

Thomas Janzen

Show, Don't Tell

Developing and Validating
a Role-Play-Based Simulation (RobS) for the
Assessment of Pre-Service EFL Teachers'
Feedback Competence on Writing

Thomas Janzen

Show, Don't Tell – Developing and Validating
a Role-Play-Based Simulation (RobS)
for the Assessment of Pre-Service EFL
Teachers' Feedback Competence on Writing

Logos Verlag Berlin



Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available on the Internet at <http://dnb.d-nb.de>.



This work is licensed under the Creative Commons Attribution 4.0 license CC BY-SA (<https://creativecommons.org/licenses/by-sa/4.0/>). Creative Commons license terms for re-use do not apply to any content (such as graphs, figures, photos, excerpts, etc.) not original to the Open Access publication and further permission may be required from the rights holder. The obligation to research and clear permission lies solely with the party re-using the material.

I acknowledge support for the publication cost by the Open Access Publication Fund of Paderborn University.

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Paderborn University.

© Copyright Logos Verlag Berlin GmbH 2025

All rights reserved.

ISBN 978-3-8325-5994-6

DOI 10.30819/5994

Logos Verlag Berlin GmbH
Georg-Knorr-Str. 4, Geb. 10,
D-12681 Berlin
Germany

Tel.: +49 (0)30 / 42 85 10 90

Fax: +49 (0)30 / 42 85 10 92

<http://www.logos-verlag.de>

Parts of this dissertation are published in:

Janzen, T., Vogelsang, C., & Rumlich, D. (2024). Feedbackkompetenz handlungsnah prüfen: Die Entwicklung einer rollenspielbasierten Simulation als Prüfungsformat für angehende Englischlehrkräfte. In D. Gerlach (Ed.), *Wissen, Können und Handeln von Fremdsprachenlehrpersonen: Professions- und Wissensforschung in den Fremdsprachendidaktiken* (pp. 221–234). Peter Lang.

This thesis is part of the project *Performanzorientierte Professionalisierung in der Lehramtsausbildung* (PERFORM-LA). The project is funded by the Federal Ministry of Education and Research in Germany as part of the *Rahmenprogramm Empirische Bildungsforschung – Nachwuchsgruppen in der Bildungsforschung*.

Funding number: NWGWIHO02.



Bundesministerium
für Bildung
und Forschung

Acknowledgements

This thesis presents the conclusion to a journey which I am thankful to have shared with so many remarkable people, for whom words can hardly express my gratitude.

First and foremost, I want to express my deepest gratitude to my supervisors, Dr. Christoph Vogelsang and Prof. Dr. Dominik Rumlich. The support the two of you have given me throughout this time is invaluable and the advice I received is priceless. Your high standards motivated me to strive for my best, while your thoughtful pragmatism allowed me to accept my limits. I am beyond thankful that you put your faith in me and gave me the necessary freedom to develop academically and personally.

I also wish to thank Prof. Dr. Peter Hohwiller and Dr. Jana Meier, not only for their contributions to this project, but also for their roles in the doctoral committee. Thank you.

The influence of Prof. Dr. Frauke Matz on my academic journey cannot be overstated, as she introduced me to the world of academia. Your support and your kindness continue to be nothing short of extraordinary. Thank you so, so much for everything.

I am beyond grateful that I had the privilege to travel this doctoral road with Philipp Wotschel. You are the best office buddy I could have hoped for. Our countless discussions sharpened my critical thinking in so many ways and improved my work beyond what I could have ever imagined. But most importantly, they always brought joy and laughter. Thank you, from the bottom of my heart.

I am so lucky that I have had the support of multiple teams who have all contributed immensely to this thesis. I am in great debt to Dr. Lea Grotegut, for your invaluable feedback and support, consistently motivating me in the writing process by kind words and delicious baked goods. Furthermore, I would like to thank the people of the *PLAZ – Professional School of Education*, particularly Dr. Annegret Hilligus, and its research department: Dr. Andreas Seifert, Ha My Truong, Dr. Franziska Schwabl, Prof. Dr. Sabine Fechner, and all participants of the *Forschungskolleg Empirische Bildungsforschung* for their comments, which have greatly improved my work. This project would not have been possible without the great work of our student colleagues: Jasmin Heinrichs, Elena Seifert, Hannah Grummel, Ella Beer, Eike Schriegel, and Carlo van Laak. Your work on- and off-screen is truly award-worthy. Thank you.

It is hard to find words to say how much I value the contributions of the amazing TEFL Team at Paderborn University. Especially, I want to thank Maike Jolmes, Julie Intveen and Benedikt Wagner, with whom I started this journey and feel particularly connected to. I also wish to thank Dr. Katharina von Elbwart, Dr. Jeannette Böttcher, Dagmar Keatinge, Leonie Schieffer, Christina Lennert,

Alexander Woitinas, Madusha Sivanandan, and Julian Schrowangen for their consistent support. I cannot overstate the appreciation I have for the work of Leigh Menschig, whose proofreading immensely improved the quality of the text. Thank you.

A heartfelt thank you goes also out to all the great people who have accompanied me on this journey in their special way. Alicia Battenfeld and Clarissa Diekmann, I am so thankful for your support, feedback and friendship over the years. I also wish to thank all of the English Language Education team at the University of Münster, especially Dr. Ricardo Römhild, Jens-Folkert Folkerts, Marius Ritter, Anika Marxl, Dr. Stephan Gabel, and Dr. Johanna Marks for all your assistance and empowering words. I would also like to thank Meral Roeben, Hoang Nguyen and Annette Methfessel for their cooperation and counsel, as well as my former colleagues at the *Berufskolleg Beckum des Kreises Warendorf* for all their flexibility. Thank you.

I also want to thank the Federal Ministry of Education and Research in Germany for the funding of this project. It is needless to say that I owe great debt to all the participants in this project – all the students who participated in the simulations, as well as the experts and stakeholders I interviewed. Thank you for your time, I wholeheartedly appreciate it.

Last, but certainly not least, I want to thank my family and friends for their understanding, the dog care and the warm meals. I wish to thank Arne, who provided me with all the books that I needed, and Markus, who supported me from day one in my decision to pursue a PhD. Pippa, my four-legged companion, I thank you for all the walks that helped me clear my mind and the joy you bring me every single day. Above all, I want to thank my wife Solveig. You have always believed in me and supported me unwaveringly. **Thank you!**

*Zuletzt möchte ich mich bei meiner Familie und Freunden für das Verständnis, die Hundebetreuung und die warmen Mahlzeiten bedanken. Mein Dank gilt Arne, der mich mit all den Büchern versorgt hat, die ich brauchte, und Markus, der mich vom ersten Tag an darin unterstützt hat, die Promotion zu verfolgen. Pippa, meine vierbeinige Begleiterin, ich danke dir für all die Spaziergänge, die mir zu einem klaren Kopf verhelfen und all die Freude, die du mir jeden Tag gibst. Der größte Dank gilt meiner Frau, Solveig. Du hast immer an mich geglaubt und deine Unterstützung war unerschütterlich. **Danke!***

Table of Contents

1	Introduction	1
Part I - (EFL) Teacher Education in Germany: From Developing Competence to Assessing Performance		
2	Structural Elements of Teacher Education in Germany.....	7
2.1	The Goal of Teacher Education	7
2.2	The Structure of (EFL) Teacher Education in Germany	10
3	Competence of EFL Teachers	12
3.1	Approaching the Terminology of Competence	13
3.2	Models of Competence in Teacher Education.....	17
3.2.1	Model of Professional Competence	18
3.2.2	Competence as Knowledge	23
3.2.3	Competence as a Continuum.....	25
3.3	Frameworks for (EFL) Teacher Education in Germany	28
3.4	Empirical Findings on Professional Knowledge of Pre-Service EFL Teachers.....	33
3.5	Competence between Knowledge and Practice	37
4	A Performance-Oriented Perspective on Constructive Alignment in the First Phase of EFL Teacher Education.....	40
4.1	Constructive Alignment	41
4.2	Learning Objectives: The Core Practice Approach in EFL Teacher Education	43
4.2.1	The Core Practice Approach	43
4.2.2	Core Practices in the EFL Context	45
4.3	Teaching and Learning Activities: Simulation-based Learning.....	47
4.3.1	Characteristics of Simulations	49
4.3.2	Typology of Simulations.....	51
4.3.3	Perception of Simulations and their Effectiveness for Learning	55
4.4	Assessment: RobS in EFL Teacher Education	57
4.4.1	Concepts of Assessment in Higher Education	57
4.4.2	Assessment in Higher Education: Students' Perceptions and Impact on Learning ..	60
4.4.3	Simulation-based Assessment in Higher and Teacher Education.....	63
4.5	Addressing the Assessment Gap in Teacher Education	65
5	Summary	68
Part II - Developing a Role-Play-Based Simulation (RobS) to Assess Feedback Competence on Writing in the EFL context		
6	The Relevance of Feedback on Writing in the EFL context.....	71
6.1	Feedback	71
6.1.1	On the History and Terminology of Feedback.....	72
6.1.2	Feedback as Literacies.....	75
6.1.3	Relevance of Feedback in (EFL) Teacher Education	76
6.2	Writing in the EFL classroom	79

6.2.1 Cognitive Process Models for (L1-) Writing.....	80
6.2.2 Approaches to Writing in the EFL classroom	84
6.2.3 Statutory Guidelines for Learning and Teaching Writing	87
6.3 Synthesis: Feedback on Writing.....	90
7 Developing a Framework for Feedback Competence on Writing in the EFL context.....	94
7.1 Methodological Reflections on Theoretical Research.....	94
7.2 Instruments for the Evaluation of Feedback Performance	95
7.3 The Framework.....	98
7.3.1 Functions of Feedback.....	99
7.3.1.1 Aim	99
7.3.1.2 Level.....	102
7.3.2 Strategies of Feedback.....	104
7.3.2.1 Communication	105
7.3.2.2 L2/FL Writing Pedagogy.....	108
7.3.3 Topics of Feedback.....	113
7.3.3.1 Diagnosis.....	113
7.3.3.2 Information	115
7.4 Discussion and Limitations of the Framework.....	117
8 A RobS to Assess Feedback Competence on Writing in the EFL context	120
8.1 Preparation	120
8.2 Simulation.....	126
8.2.1 The Role of Mia.....	126
8.2.2 Standardization of the Learner	127
8.3 Assessment.....	130
8.4 Discussion and Limitations of the RobS	140
9 Summary	142
Part III - Methodology: Investigating the Validity of the RobS	
10 Criteria for Test Evaluation	145
10.1 Overview of Test Criteria	145
10.1.1 Secondary Criteria: Practicality, Fairness and Authenticity.....	146
10.1.2 Primary Criteria: Objectivity and Reliability.....	147
10.2 From Validity to Validation.....	148
10.2.1 Traditional Approaches to Validity.....	149
10.2.2 Messick's Approach to Construct Validity	151
10.2.3 The Argument-based Approach to Validation.....	156
10.3 Implications for the Validation Process.....	160
11 IUA for the RobS.....	161
11.1 A Classification System for Validation by Schreiber & Gut (2022).....	161
11.2 Making the Claims: The IUA and Sources of Evidence for the Test	164
11.2.1 Claim 1: Model	165
11.2.2 Claim 2: Test.....	165
11.2.3 Claim 3: Performance	169
11.2.4 Claim 4: Item score	169

11.2.5 Claim 5: Test score.....	170
11.2.6 Claim 6: Latent trait.....	170
11.2.7 Claim 7: Interpreted trait.....	171
11.2.8 Claim 8: Use.....	171
12 Summary	172
Part IV - Empirical Findings: Results of the Investigations	
13 Research Design and Overview.....	173
14 (Pre-)Pilot Phase and Survey: Actor Training and Material Development .	177
14.1 Pre-Pilot Phase	177
14.1.1 Creating and Evaluating Authentic Material.....	177
14.1.2 Actor Training.....	185
14.2 Pilot Survey.....	187
14.2.1 Data Collection and Sample Characteristics	188
14.2.1.1 Data Collection.....	188
14.2.1.2 Sample Characteristics	193
14.2.2 Investigating the Verbal Triggers.....	194
14.2.2.1 Method: Stimulated Recall Interviews	195
14.2.2.2 Results of the SRIs.....	197
14.2.2.3 Discussion and Limitations	203
14.2.2.4 Conclusion	205
14.2.3 Deciding on the Test Documents	206
14.2.4 Deciding on the Research Procedure.....	209
14.3 Summary	210
15 Main Survey	211
15.1 Data Collection and Sample Characteristics.....	211
15.1.1 Data Collection.....	211
15.1.2 Sample Characteristics	213
15.2 Study 1: Pre-service EFL Teachers' Perspective on RobS as Assessment.....	215
15.2.1 Methods	217
15.2.2 Results	219
15.2.2.1 Authenticity of the RobS	219
15.2.2.2 Competences Tested by the RobS.....	222
15.2.2.3 Fairness of the RobS as Assessment	226
15.2.2.4 Use of the RobS for Summative Assessment.....	227
15.2.3 Discussion and Limitations.....	230
15.2.4 Summary	233
15.3 Study 2: Expert Interviews	234
15.3.1 Method: Expert Interviews.....	235
15.3.2 Study Design and Data Processing	237
15.3.3 Results of the Qualitative Expert Interviews.....	242
15.3.3.1 Experts' Evaluative Comments on the Performances	242
15.3.3.2 Evaluation of the Scoring Instrument.....	248
15.3.4 Discussion of the Results and Modifications to the Scoring Instrument.....	265
15.3.5 Limitations	269

15.3.6 Summary	270
15.4 Study 3: Scoring Reliability	271
15.4.1 Scoring Reliability on the Item Level	272
15.4.1.1 Rating of the Full Sample on the Item Level	272
15.4.1.1.1 Methods.....	272
15.4.1.1.2 Results	275
15.4.1.2 Inter-Rater Reliability on the Item Level	282
15.4.1.2.1 Methods.....	282
15.4.1.2.2 Results of Rating I on the Item Level	286
15.4.1.2.3 Modifications	290
15.4.1.2.4 Results of Rating II on the Item Level.....	297
15.4.1.3 Discussion of Items for the Final Scale.....	301
15.4.1.3.1 Exclusion of Items	301
15.4.1.3.2 Inclusion of Items Despite Scoring Frequency	302
15.4.1.3.3 Inclusion of Items Despite Insufficient Corrected Item-Total Correlation.....	304
15.4.1.3.4 Inclusion of Items Despite Potentially Insufficient Inter-Rater Reliability	305
15.4.1.4 Revised Reliability Values on the Item Level.....	309
15.4.2 Scoring Reliability on the Test Score Level.....	310
15.4.2.1 Descriptive Statistics and Tests on Normal Distribution for the Full Sample	310
15.4.2.2 Inter-Rater Reliability on the Test Score Level	313
15.4.2.2.1 Methods.....	314
15.4.2.2.2 Results of Rating I on the Test Score Level	314
15.4.2.2.3 Results of Rating II on the Test Score Level	319
15.4.2.3 Discussion of the Test Score-Focused Reliability	324
15.4.3 Limitations and Implications from the Reliability Investigations	326
15.4.4 Scoring Independence from Actors	328
15.4.4.1 Results	328
15.4.4.2 Discussion.....	329
15.4.5 Summary	329
15.5 Study 4: The Relationship between Test Score and Other Variables	330
15.5.1 Relevant Criteria	331
15.5.2 Hypotheses.....	334
15.5.3 Methods	335
15.5.3.1 Personal and Academic Data.....	335
15.5.3.2 Areas of the target domain	336
15.5.3.2.1 Specific PK: Declarative Feedback Knowledge	336
15.5.3.2.2 EFL-specific PCK: FALKO-E	338
15.5.3.2.3 CK / LP: The LexTale Test.....	343
15.5.3.3 Expert Ratings.....	345
15.5.4 Results	346
15.5.4.1 H1: School-related Experience	346
15.5.4.2 H2: Academic Progress and Success.....	348
15.5.4.3 H3: Facets of EFL Teacher Competence.....	351
15.5.4.3.1 Descriptive Statistics, Normal Distribution and General Inferential Statistics	351
15.5.4.3.2 Inferential Statistics	353
15.5.4.4 H4: Expert Ratings.....	356

15.5.5 Discussion	359
15.5.6 Limitations	365
15.5.7 Summary	367
16 Summary	368

Part V – Discussion: Evaluating the Validity of the RobS

17 Evaluating the Claims: The Validity Argument	373
17.1 Claim 1: Model	373
17.2 Claim 2: Test	375
17.3 Claim 3: Performance	377
17.4 Claim 4: Item Score	379
17.5 Claim 5: Test Score	381
17.6 Claim 6: Latent Trait	383
17.7 Claim 7: Interpreted Trait	384
17.8 Claim 8: Use	386
17.9 Conclusion: To use, or not to use?	387
18 Looking Back – Looking Forward	388
18.1 Summary	388
18.2 Limitations	391
18.3 Implications and Future Perspectives	395
19 References	399
List of Abbreviations	447
List of Figures	451
List of Tables	453

Appendix

A. Test Material	457
A.1. M0: Vignette	457
A.2. M1: Learner text	458
A.3. M2: The Task for the Text	459
A.4. M3: Excerpt from the Series of Lessons	460
A.5. M4: Sample Text from the Series of Lessons	461
A.6. M5: Empty Sheet for Notes	462
A.7. Role description of Mia (German)	463
A.8. Text Sample A from the Pilot Survey	465
B. Rating manual	466
C. Demographic and Academic Data Collected in the Written Survey	523
D. Interview Guidelines	525
D.1. Pre-pilot Phase: Expert Interviews	525
D.2. Pre-pilot Phase: PSETs Interviews	525
D.3. Pilot Survey	527
D.4. Main Survey	529
D.5. Expert Interview	530

E. Transcription Guidelines	540
E.1. Follow-Up Interviews and Simulations (Main Survey).....	540
E.2. Stakeholder and Expert Interviews.....	541
F. Coding Manuals.....	542
F.1. Stimulated Recall	542
F.2. Guideline Interview	548
F.3. Expert Interviews.....	607
G. Coding Samples	620
G.1. Stimulated Recall	620
G.2. Simulation	638
G.3. Guideline Interview– Main Survey	658
G.4. Expert Interview – Part 1	666
G.5. Expert Interview – Part 2	670
H. Statistical Information.....	685
H.1. Intercoding Sub-Codes PSETs Guideline Interview.....	685
H.2. Discrepancies in Inter-Rating I.....	686
H.3. Inter-Item Correlation Matrix for 27 Items.....	689
H.4. Test Score Frequencies.....	692
H.5. Normal Distribution in the Actor-specific Groups.....	693
H.6. Further Statistical Information on SPK, PCK and CK /LP	693
I. Digital Supplement.....	695

1 Introduction

Across the field, there is a growing awareness that knowledge about teaching and schools, however extensive, is not sufficient to prepare capable teachers. Nor is it enough to hope that the necessary routines and practical skills will be acquired in internships and practica, especially if the aim is to prepare teachers to enact more ambitious and equitable forms of instruction.

Urban Fraefel and Pam Grossman (2024, p. 139)

Teachers have a large impact on their learners'¹ achievement (Hattie, 2023, p. 216)². However, the potential positive impact that teachers could have on their learners' achievement is not fully realized, as a substantial share of learners fails to achieve the desired outcomes. This discrepancy is supported by findings of large-scale studies, such as PISA³ and the IQB⁴ *Trends in Student Achievement*, that show mixed results for learner achievement in Germany (OECD, 2023, p. 183; Stanat et al., 2023). This raises questions about the effectiveness of teacher education, which is supposed to prepare teachers to maximize their impact on learners' achievement. This concern is corroborated by findings from Hattie, who states that "[teachers']⁵ initial teacher training programs have little impact on how well teachers influence their students' achievement" (Hattie, 2023, p. 243), quantifying the impact with a small effect size of $d = 0.10$ (Hattie, 2023, p. 229). Thus, these findings point to potential areas for improvement in the current approach to teacher education in Germany.

A central challenge in teacher education is the gap between theory and practice. Often, a lack of action-orientation and a perceived disconnect between the acquired theoretical, academic knowledge at university and supposed actual practice in the classroom are named as reasons for this gap (Rothland, 2020, p. 133). A commonly proposed solution is to increase the share of practical phases, allowing future teachers to apply their knowledge in school internships

¹ In this thesis, the term *learner* is usually used when referring to K-12 education, i.e. primary and secondary education, while *students* refers to higher education, i.e., university. Exceptions are, of course, direct quotations.

² In this thesis, the APA guidelines are generally followed for citations. However, in-text citations will also include the page number for indirect citations, if the publication is not addressed as a whole. This is meant to increase transparency and traceability of the scientific process.

³ PISA is the abbreviation of *Programme for International Student Assessment*.

⁴ IQB is the abbreviation of Germany's *Institut zur Qualitätssicherung im Bildungswesen* (Institute for Educational Quality Improvement).

⁵ For reasons of legibility, parentheses that indicate grammatical adjustment in direct quotations always include the full word.

(Rothland, 2020, p. 133). However, there are indications that the actual application of knowledge during such authentic practical phases is quite challenging for students (Zeichner, 2010, p. 91). Possibly, the complexity of the classroom might be a reason for this finding.

One approach to address the gap between theory and practice is through an *approximation of practice* (Grossman, Compton, et al., 2009, p. 2076). In such an approach, pre-service teachers (PSTs), i.e., teacher students at university, are allowed to perform in an authentic, but reduced-complexity situation, which allows them to practice the application of knowledge without necessarily interacting with real learners (Grossman, Compton, et al., 2009, p. 2076). However, PSTs are rarely given the opportunity to approximate practice in terms of action-oriented sequences during their university-based studies (Grossman, Compton, et al., 2009, pp. 2094–2095). In medical education, such opportunities are provided, for instance, in the form of role-play-based simulations (RobS⁶), in which trained actors take on the role of standardized patients (Schnabel, 2018). In teacher education, RobS would allow future teachers to take on the role of the teacher and engage in an authentic situation with an actor, portraying, for example, a standardized learner (SL) or parent (Dotger et al., 2010).

RobS can also be used as a performance-oriented format of summative assessment. In medical education, simulations are used for the assessment of clinical skills before (future) doctors engage with real patients in high-stakes situations and potentially severe consequences (Sopka et al., 2018). Considering the relevance that assessment has for students and their learning behavior, the implementation of RobS as performance-oriented assessment formats might lead students being better prepared for actual practice in the classroom, as “assessment drives learning” (Sopka et al., 2018; see also Gulikers, Bastiaens, et al., 2008). In teacher education, however, RobS are rarely found as an assessment format in teacher education programs or addressed in research.

If RobS were used as an assessment format in teacher education, a focus should also be on subject-specific practices, i.e. feedback on writing in the context of English as a Foreign Language (EFL). While teachers need to acquire a diverse set of competences for their profession, they are still primarily responsible for teaching their subjects, such as EFL, and mastering the required practices. In the EFL classroom, for instance, providing feedback on writing can be viewed as such a relevant practice (Grossman, 2018, p. 168; Pang, 2019, p. 265; Summer & Plötz, 2024, p. 164). Feedback has proven to be one of the most impactful

⁶ RobS (Role-play-based Simulation) is used as a fixed term for the test that is at the focus of this thesis as well as for an abbreviation of any role-play-based simulation. For legibility, RobS is used as the singular as well as plural form.

variables influencing learning (Hattie, 2023, p. 320; Hattie & Timperley, 2007). However, there is also a great variance in the effect that different kinds of feedback have, indicating that it is essential for teachers to learn how to provide feedback appropriately (Hattie, 2023, p. 319; Wisniewski et al., 2020, p. 12). Feedback on writing in the EFL context has shown to be specifically relevant (Scherer et al., 2024). Conceptualizing writing as an iterative and nonlinear process, feedback is required in order for revisions to take place (Hayes & Flowers, 1980; Hyland, 2019, p. 12). As writing is also a skill that is statutorily part of Germany's EFL curricula, EFL teachers need to be able to provide appropriate feedback on writing products of their learners (Council of Europe, 2020, p. 68; Kultusministerkonferenz [KMK], 2023a, p. 16). Considering the importance of feedback in writing, this context offers a relevant focus for developing a performance-oriented assessment format.

To evaluate the quality of such a newly developed performance-oriented assessment format, the criterion of validity must also be tackled. Within an argument-based approach to validation, different claims of validity are made, investigated and discussed critically (M. T. Kane, 2013a). Following this approach, validity does not refer to the test itself but the extent to which theory and evidence support the intended interpretation of its results (AERA, APA, NCME, 2014, p. 11). From a theoretical perspective, the construct of the test must be described transparently. From an empirical point of view, data must be collected that allows to discuss the claims. All findings are then used to make an overall evaluation concerning the appropriateness of the intended score interpretation.

Based on these considerations, a gap in teacher education can be identified that is addressed in this thesis. Teacher education programs do not prepare PSTs adequately for the profession, as teachers do not fully capitalize on their potential positive impact. The gap between theory and practice is often cited as a challenge in teacher education programs. Using RobS as a way to approximate practice might be one approach to bridge this gap. As an assessment format in teacher education, RobS have the potential to influence PST's learning behavior, focusing more on action-oriented competences which are relevant for actual practice in the classroom. However, there is a lack of subject-specific, performance-oriented assessment formats. This gap is addressed specifically for pre-service EFL teachers (PSETs) in this thesis, as a RobS is developed for the assessment of PSETs' feedback competence on writing. To ensure the quality of the developed assessment format, the aspect of validity is addressed in the argument based-approach. Consequently, the following can be stated:

The aim of this thesis is to develop and validate a role-play-based simulation in the EFL context that can be used for summative assessment.

Structure of this Thesis

This thesis is structured in five main parts, each addressing a specific aspect of the research process. Each part is divided into chapters and concludes with a summary. The following section provides an overview of each part and its chapters.

In **Part I**, a more detailed look at EFL teacher education in Germany is presented and the goal of this thesis derived. In Chapter 2, the specific goals and structural elements that are relevant for this thesis are described. In Chapter 3, the term competence is focused on. Specific models of (teacher) competence are presented and the relevant statutory frameworks are summarized. Thereafter, selected empirical findings are summarized regarding professional knowledge of (pre-service) EFL teachers. Lastly, the extent to which the current approach of teacher education suffices to prepare PSETs for practice is critically discussed. In Chapter 4, a performance-oriented approach to university-based teacher education is outlined, focusing on learning objectives, teaching and learning activities, and assessment formats. Lastly, the current assessment practices in teacher education and a potential gap of performance-oriented assessment are critically discussed. As a consequence of the discussion, the aim of this thesis, i.e., developing a RobS, is described as one approach to address this gap.

In **Part II**, the design of the RobS is presented. In Chapter 6 the reasoning for the content focus of the RobS, i.e., assessing feedback competence on writing, is provided using theoretical models and empirical evidence. In Chapter 7, a framework is presented, which includes the relevant competence facets for describing and assessing feedback competence on writing holistically. In Chapter 8, the developed RobS is described as a potential assessment format.

In **Part III**, the methodological background of this thesis is outlined. In Chapter 10, different criteria for test evaluation are described. The relevance of validity is highlighted by providing a diachronic perspective on different conceptualizations of this criterion. It concludes with the argument-based approach to validation, which describes the postulation and evaluation of specific claims for an intended use of a test and is followed in this thesis (M. T. Kane, 2013a). In Chapter 11, the specific *interpretation/use argument*, which can be referred to as a list of claims for which evidence must be provided to evaluate the RobS, is presented.

In **Part IV**, the research process and the empirical findings for the test evaluation are presented. In Chapter 13, an overview of this thesis' two research cycles is given. In Chapter 14, the first cycle, i.e., the pre-pilot phase and the pilot survey, is focused on. The development and evaluation of the test material is detailed, as is the training of the actors for the RobS. In Chapter 15, all the studies that are executed in the main survey are subsumed. In Study 1, the perspective of the participants on the RobS is described. In Study 2, the qualitative expert

judgements of RobS performances and the scoring instrument is focused on. In Study 3, the process of designing the scoring instrument, based on different reliability indices on the item and test score level is detailed. In Study 4, findings of the testing of several hypotheses that are stated *a priori* to discuss the interpretation of test score are presented.

In **Part V**, the findings are discussed in terms of the test's quality. In Chapter 17, the previously made claims are evaluated using the findings generated in this thesis and a conclusion is drawn regarding the validation of the RobS. In Chapter 18, the overall thesis is summarized, limitations drawn and implication made. The appendix includes the relevant material that complements this thesis, i.e., the material for the RobS, the rating manual and transcription guideline for the RobS, the guidelines, transcription rules and coding manuals for the interviews and further statistical information.

Part I - (EFL) Teacher Education in Germany: From Developing Competence to Assessing Performance

*[The] quickest way to change student learning
is to change the assessment system.*

Lewis R.B. Elton and Diana Laurillard (1979, p. 100)

In the first part, the theoretical background for this thesis is given and the research aim derived. Firstly, the focus is put on the specific context of the study and the structure and goals of teacher education in Germany are described. Secondly, the notion of competence is focused on. The term is discussed theoretically first, before specific models of teacher competence are presented. The statutory guidelines for EFL teacher education are outlined and empirical findings presented that allow for an indication of competence development. The dissatisfaction that is associated with the status quo of teacher education is discussed as a result of these observations. Thirdly, a performance-oriented perspective is provided to address this issue. Based on the framework of constructive alignment by Biggs (1996) and current research, performance-oriented learning objectives, teaching and learning activities and assessment formats are proposed. Regarding the latter, an assessment gap in teacher education is identified, which leads to this thesis' research aim.

2 Structural Elements of Teacher Education in Germany

The aim of this first chapter is to describe the status quo of teacher education in Germany. Firstly, the aims of teacher education in Germany will be presented. Secondly, the history, structure, and organization of (EFL) teacher education in Germany will be outlined, so as to show how the development of achieving these aims is organized.

2.1 The Goal of Teacher Education

Teachers have a large impact on their learners' achievement, as stated in the introduction (Hattie, 2009, 2023). Thus, it is the job of the teacher to enable student learning in the most beneficial way. Consequently, teacher education programs should prepare PSTs effectively to do their job and impact student learning positively. The nature of the relationship can heuristically be depicted by a causal flow chart as seen in Figure 1 (Terhart, 2012b, p. 7).

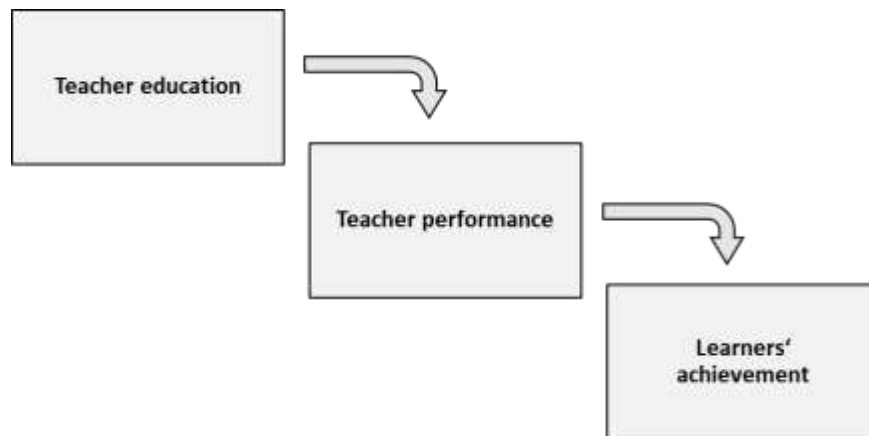


Figure 1: Causal flow chart depicting the influence of teacher education on teacher actions and the consequential influence on student learning (Terhart, 2012b, p. 7; translation TJ).

The assumption is that through teacher education, people acquire the necessary resources to fulfill the demands of the teaching profession (Vogelsang, 2014, p. 20). These resources can then be drawn upon to act adequately in concrete situations, which, in turn, have an impact on learning achievement. Consequently, learners' achievement is thus a direct result of teacher education. The exact nature of this relationship is, however, not that clear cut, as other potentially influencing variables are not included. Diez's model (2010; see Figure 2) takes mediating factors into account and adds an evaluative connotation as it includes questions to the proposed flow (Vogelsang, 2014, p. 21). In the model, the desired outcome of teacher education programs is described as "knowledge, skills, and dispositions" (Diez, 2010, p. 442), which are to be defined in standards that explicate the goals.

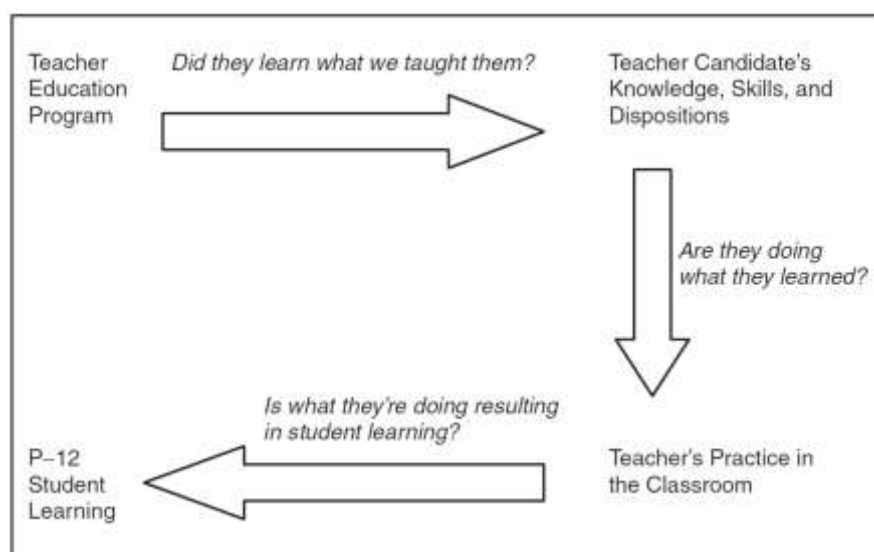


Figure 2: Flow diagram of impact on student learning (Diez, 2010, p. 442).

The first inquiry of evaluation then, should be to investigate the degree to which PSTs have achieved those standards. The second question is if teachers actually implement what they have learned in the classroom, with the subsequent final question being what the impact of their actions on P-12 student learning is, i.e., from pre-school to twelfth grade.

More recent models on the relation between teacher education and student learning have focused on teachers' professional competence. In the COACTIV⁷ study, teachers' professional competence is considered to be influential for the instructional quality, which in turn influences the students in their achievement⁸ (Krauss et al., 2020, p. 315). In her heuristic model, Voss links teacher education with a constructive understanding of teaching, creating a double offer-use-model⁹ that describes teacher education as one offer-use-situation, and the school's classroom as the other (Voss, 2019, p. 10). This model describes teachers' professional competence (see Chapter 3.2.1) as the result of teacher education, and a prerequisite for practice, similar to Diez. However, it also includes the learners and other variables like context, family or their individual learning (Helmke, 2022, p. 77; Voss, 2019, p. 11).

The cascade model by Krauss et al. (2020) is to date the most differentiating because it merges the model of competence as a continuum¹⁰ (Blömeke et al., 2015) with other relevant and salient paradigms in teacher education to describe the goal of teacher education (see Figure 3). The model includes five columns: Columns 1, 2 and 3 cover competence understood as a continuum as described by Blömeke et al. (2015), which, contrary to other models of competence, sees performance as part of competence, not a requirement for it (see Chapter 3). Columns 3 and 4, respectively, cover the offer-use-model as described, for example, by Helmke (2022, p. 77), while column 5 focuses on the outcome of the learning process on the learners' side. Evidently, teachers' competence, especially the performative aspect, i.e., the teacher's actual behavior, is central in all of these models. This part of the cascade intersects most clearly with learning – depending on how the teacher performs, learners might be more or less inclined to actually use the offers and achieve learning gains. The next chapter will describe how teacher education seeks to achieve its goal on a structural level.

⁷ COACTIV is the abbreviated form for *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers* (Baumert & Kunter, 2013).

⁸ See Chapter 3.2.1 for a detailed overview of the COACTIV model of professional competence.

⁹ See also Helmke (2022, p. 77).

¹⁰ See Chapter 3.2.3 for a detailed description of the model.

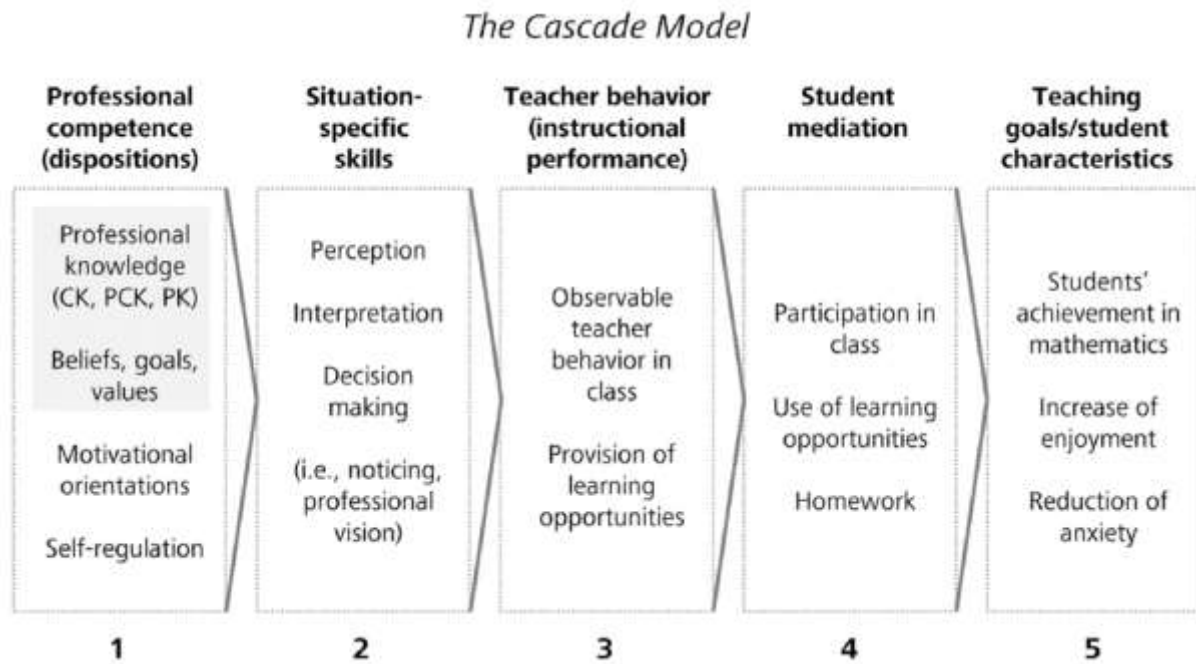


Figure 3: *The Cascade Model* (Krauss et al., 2020, p. 316).

2.2 The Structure of (EFL) Teacher Education in Germany

In Germany, the structure of teacher education, like education in general, is the responsibility of the individual federal state. In most federal states, teacher education is differentiated in three different phases:

The first phase: university-based education;

The second phase: teacher training or the induction phase, which takes place at training schools, as well as teacher education institutes (e.g., *Studienseminare*, *Zentren für schulpraktische Lehrkräfteausbildung*);

The third phase: further education as in-service teachers (Pasternack et al., 2017, p. 20; Terhart, 2007).

The main objective of the first phase is to lay the foundation for teaching the respective subjects. This includes a deeper knowledge about the subject matter, but also other pedagogical (content) dispositions (Pilypaitytė, 2013, p. 33). As a result of the Bologna process, the first phase is separated into a Bachelor's and a Master's program in most of Germany's federal states and, thus, covers at least five years of studies in total. The first phase also includes school internships of various lengths, allowing prospective teachers to gain experience in the role of the teacher. These practical phases are closely accompanied by the universities to ensure an adequate guidance and reflection (J. Meier et al., 2024, p. 312). In the state of North Rhine-Westphalia, the local context of the presented studies in Chapter 15, there are three practical phases (LABG NRW 2009, §12). In the

Aptitude and Orientation Internship (Eignungs- und Orientierungspraktikum; abbreviated as AOI) PSTs are supposed to spend five weeks at a school, mostly observing other teachers but also gaining initial experience in teaching. The AOI is supposed to provide PSTs with an opportunity to reflect on the appropriateness of the choice of the teaching profession early on in their course of study (Porsch & Gollub, 2018, p. 240). In the *Occupational Internship (Berufsfeldpraktikum; abbreviated as OI)* students are supposed to spend four weeks getting an insight into a profession that is related to the teaching profession (J. Meier et al., 2024, p. 310). Both the AOI and the OI are part of the Bachelor program (J. Meier et al., 2024, p. 312). In the Master phase, students need to finish the *School Internship Semester (Praxissemester; abbreviated as SIS)*. The SIS is the longest internship and students are supposed to gain more experience as teachers, guided by mentors at schools. During the SIS, students are encouraged to put their theoretical knowledge into practice and conduct action research (Porsch & Gollub, 2018, p. 240). The second phase of teacher education, the induction phase, takes place at dedicated institutes that focus on practical teacher training (Pilypaitytė, 2013, p. 33). Teachers in the induction phase spend most of their time teaching real classes, with and without supervision from more experienced teachers. They are guided by the institutes and are assessed predominantly based on their behavior in class (Schustereder, 2020, p. 355). After 18 months of teacher training and passing the second and final state examination successfully, teacher candidates are then certified to teach. The third phase includes further education programs that the teachers are willing to participate in on a primarily voluntary basis.

Other structural elements of teacher education refer to the specific subjects PSTs aspire to teach. For instance, PSTs for lower and upper secondary choose (at least) two different subjects for their studies, supplemented with the subject of educational sciences. Candidates who want to study for primary schools or for schools for learners with special needs usually have more subjects. One special characteristic in the education of modern foreign language teachers is that they need to include a stay abroad in their studies (for North Rhine-Westphalia, see LABG NRW 2009 §11 (10)), a peculiarity that can be dated back to the 19th century (Klippel, 2018, p. 20). The specific requirements for the stay abroad are slightly different in the various federal states. Using the example of North Rhine-Westphalia, the stay comprises twelve weeks, which can be split into smaller segments, and the target language must be spoken in the country of the stay.

In conclusion, it can be stated that the primary objective of teacher education is to develop teacher competence to prepare PSTs to perform at their possible best in the classroom, as teacher performance has a direct impact on actual learner achievement. In Germany, teacher education is structured in three

phases. The first phase is situated at the university, focuses on the acquisition and learning of knowledge, and the first exploration into the field by the means of internships. The second phase allows teacher trainees to actually teach in class as they work at schools, while being guided by teacher education institutes. In the third phase, teachers can further educate themselves. The following chapter will describe, in detail, the notion of EFL teacher competence as the focus of EFL teacher education.

3 Competence of EFL Teachers

The aim of this chapter is to provide the necessary theoretical foundation to discuss teachers' professional competence¹¹. The results of *PISA 2000* not only led to a shift in the German education system, but it also led to a shift towards standardization in German teacher education, and, thus, also a turn towards teachers' competences¹² (Blömeke, 2006, p. 26). While there seems to be a consensus that teachers should acquire competences, a closer look at the individual definitions and approaches reveals that there are differences in the understanding of what competence means (Bresges et al., 2014, p. 9). Apparently, consensus exists only therein that the term is used without a definitory consensus (Frey & Jung, 2011, p. 54; K. Schneider, 2019, p. 1939; Weinert, 2001, p. 45). Consequently, as "a generally valid definition of competence that would satisfy scientific criteria is not a realistic goal" (K. Schneider, 2019, p. 1940), this chapter will not seek to do so. Rather, this chapter will start with an exploration of the term, before providing an overview of the most prominent conceptualizations of competence in teacher education: the model of professional competence (Baumert & Kunter, 2006, 2013) and competence as a continuum (Blömeke et al., 2015). These models are chosen, as they have been quite influential in conceptualizing competence in teacher education research and allow for a discussion of the extent to which teacher education fulfills its goal. EFL teacher competence will be explored with regard to the legal guidelines so as to provide the specific context in which the RobS and this thesis operate. Lastly, an overview into current research of EFL teacher competence, and the extent to which teacher education is successful, will be provided. This chapter will conclude with a discussion on

¹¹ For reasons of cohesion, this thesis will only address the competence-oriented approach of teacher education. For an overview of different approaches to teacher education, see Cramer et al. (2020, pp. 131–214).

¹² In this thesis, competences, i.e., the plural of competence, is used, if specific areas in which a person can acquire competence are addressed, e.g., competences in teaching English and Chemistry. Competence in the singular form may refer to a single area of competence, or to the general broader notion. Competency, is not used synonymously to competence, but rather describes a smaller dimension of competence, with competencies referring to several different subdimensions.

competence and whether the current approach taken in teacher education suffices in achieving its goal of preparing teachers for practice.

3.1 Approaching the Terminology of Competence

The concept of competence can be located within different traditions in linguistic (Chomsky, 1965), cognitive, social and educational sciences (Klieme & Hartig, 2007; Klieme et al., 2008; K. Schneider, 2019, pp. 1938–1939; Zierer & Arnold, 2017). Chomsky introduces competence merely as a technical term¹³ to describe the linguistic system that is shared by all humans (Klieme et al., 2008, p. 4). It is dissimilar to performance, which is what can be observed by humans producing language: “We thus make a fundamental distinction between competence (the speaker-hearer’s knowledge of his language) and performance (the actual use of language in concrete situations)” (Chomsky, 1965, p. 4). Thus, performance is the potential “reflection” (Chomsky, 1965, p. 4) of competence. However, this logic applies only to an idealized version of performance. Considering, for example, that the various cognitive processes in the production of oral language might result in irregularities in the language performance, the speaker’s actual competence might be hidden (see also Levelt, 1995). This rather dichotomic understanding, then, sees competence as unbound from context (Klieme & Hartig, 2007, p. 16).

While competence is also understood to be a prerequisite for action or performance, in conceptualizations that stem from educational or psychological contexts it is more strongly related to specific situational demands (Jung, 2010, p. 9; Klieme & Hartig, 2007, p. 16). The probably most-frequently used point of departure in research of competence in education, at least for Germany, is Weinert’s definition of competence, which found prominence after being used for the *Bildungsstandards (BISTAs)* by *The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK)*^{14,15} (Krauss, Lindl, Schilcher, Fricke, et al., 2017, p. 35; Marks, 2023, p. 40). According to Weinert, competences are

¹³ The technicality that has since been associated with competence is the target of much criticism (K. Schneider, 2019, p. 1940).

¹⁴ The conference is on the following abbreviated as KMK based on the German name *Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland* and its common use. “Länder” refers to the 16 federal states in Germany.

¹⁵ According to the KMK, it “is the oldest conference of ministers in Germany and plays a significant role as an instrument for the coordination and development of education in the country. It is a consortium of ministers responsible for education and schooling, institutes of higher education and research and

die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen. (Weinert, 2014, pp. 27–28)¹⁶

(the cognitive skills that are available to, or learnable by, individuals in order to solve specific problems, as well as the related motivational, volitional and social dispositions and skills to use the solutions to the problems successfully and responsibly in variable situations.)¹⁷

Competence is thus a necessary requirement – from a cognitive, motivational, volitional and social perspective – for finding solutions to a specific problem, but also to use solutions variably and responsibly (Krauss, Lindl, Schilcher, Fricke, et al., 2017, p. 17). Following the logic of Weinert, certain characteristics have been attributed to the idea of competence. They serve as a suitable approach to further explore the idea of competence, specifically against the backdrop of teacher education. The following compilation is influenced by the work of Marks (2023, pp. 39–51) and Vogelsang (2014, pp. 29–30), but specified and focused to the specific context of this thesis, i.e., performance-oriented assessment of PSTs.

Competences can be learned

Any individual can acquire competences. Although this statement might be classified as a truism, there has been a longstanding assumption that the quality of a teacher is defined by their personality, rather than their ability (Mayr, 2014, p. 189). While personal dispositions might be a contributing factor, i.e., as a motivational disposition (see Chapter 3.2.1), competences can still be learned by all those who are willing to do so, and research is focused on identifying appropriate scenarios for competence development to occur (Blömeke et al., 2015, p. 5). Interestingly, in Mayr's reading, Weinert states explicitly that only the cognitive aspects of competence can be learned (2014, p. 203), while this is not overtly stated by most other researchers who refer to the broader idea of the learnability of competences (Blömeke et al., 2015, p. 5; Klieme & Hartig, 2007, p. 17). Consequently, competences are also teachable in the sense that, by the arrangement of targeted and appropriate learning scenarios, a teacher can provide opportunities for a student or learner to acquire and develop competences (Voss, 2019,

cultural affairs, and in this capacity formulates the joint interests and objectives of all 16 federal states." (Kultusministerkonferenz [KMK], 2023b)

¹⁶ Although generally following APA in its 7th edition for citation, occasionally bloc citations are used for quotes of less than 40 words, if it is in German and requires an additional translation.

¹⁷ For German block citations that are provided in German and English, the translation by the author will be provided in parentheses subsequently and not explicitly indicated.

p. 11). Due to the fact that competences are bound to specific contexts and situations, they “cannot be directly taught” (Weinert, 2001, p. 63), but they need to be acquired “*in situ*” (Marks, 2023, p. 51; emphasis in the original). For teacher education, that means that future teachers must be given the opportunity to develop these competences in action, e.g., in school or in school-approximating scenarios like internships, lab schools, or other activities (see also Chapter 4.3).

Furthermore, as learning is a continuous process, competence acquisition cannot be described as a dichotomy; rather, there are levels of competence (Blömeke et al., 2015, p. 7; Frey & Jung, 2011, p. 7; Kunter et al., 2013, p. 64). These levels are mostly seen as distinct and can be developed subsequently. Although individuals might have achieved the same certificate and work in identical positions, e.g. as teachers, there might still be “interindividual differences” (Kunter et al., 2013, p. 64) between teachers in their respective competence levels. Reasons for this are usually found either in their individual aptitude, e.g., “stable personality characteristics” (Kunter et al., 2013, p. 64) or in their particular qualification process (Kunter et al., 2013, p. 65). Weinert argues that both hypotheses contribute to the evident differences: “[Inter-individual] competence differences are always a confounded result of inherited talents, developmental regularities, general environmental influences, and the quantity and quality of specific learning opportunities” (2001, p. 61). Hence, competences can be learned in principal, but not every individual may acquire the specific competences, even though they are given the same opportunities as others who have successfully acquired the competence. The principle of learnability has also drawn criticism as this would “[trivialize]” (K. Schneider, 2019, p. 1948) the concept of competence because “[...] every person is competent who does not dispose of a competence, but can in principle learn it” (Roeger, 2016, p. 157 as cited in K. Schneider, 2019, p. 1948). However, the notion of being competent usually applies to people that indeed have acquired competences, the principal of potentially being competent is not equal to being competent.

Competence includes cognitive, as well as affective-motivational aspects

In his definition, Weinert distinguishes at least two different aspects of competence, those that are cognitive in nature, and those that are motivational, volitional and social (Weinert, 2014, pp. 27–28). The last three aspects are also often subsumed as “non-cognitive” (Vogelsang, 2014, p. 29) or affective-motivational (Blömeke et al., 2015, p. 7; Marks, 2023, p. 41) dispositions. These dispositions are “latent traits (i.e., human constructions of unobserved processes); they cannot be directly observed but have to be inferred from observable behavior”

(Blömeke et al., 2015, p. 3)¹⁸. Cognitive aspects usually refer to different domains of knowledge. In the area of teacher education, Shuman's taxonomy of content knowledge, pedagogical knowledge and pedagogical content knowledge is regularly referred to in this context (Shulman, 1986). Affective-motivational aspects include, for example, beliefs, professional motives, self-efficacy and self-regulation (Blömeke, 2011, p. 15).

The differentiation of competence in these distinct facets allows for an analytical perspective of competence, though the aspects remain connected and interdependent (Marks, 2023, p. 41). The actual relation of cognitive and affective-motivational dispositions is, however, still discussed as, in some approaches, cognitive dispositions seem to dominate and affective-motivational dispositions are seen more as one influencing factor (Meinhardt, 2018, p. 163). Zlatkin-Troitschanskaia and Seidel argue that, despite the still rather underexplored relationship, competent action would require at least a minimal amount of both cognitive and affective-motivational dispositions, as excellent cognitive dispositions would not be able to compensate for a lack of motivation (2011, p. 229).

Competences can be measured

Competences can be measured, otherwise competence development could not be traced and assessed. This is in alignment with the principle of learnability and the analytical perspective. However, it has been described as difficult to actually measure competences (S. Meier, 2015, p. 82). Dividing competence into smaller facets of cognitive and affective-motivational dispositions, i.e., competencies, facilitates measurement. Smaller fragments of competence can be focused on initially and later integrated to provide a holistic description of an individual's competence.

Competences are manifested in performance

As cognitive and affective-motivational dispositions are "latent traits" (Blömeke et al., 2015, p. 3), they are not directly observable but rather a prerequisite for behavior. Consequently, competence can only be observed in performance – reflecting Chomsky's distinction from the beginning. Thus, competence could be viewed as a "disposition whose existence shows itself in actual action, the performance" (Vogelsang, 2014, p. 28; translated by TJ).

¹⁸ On a nomenclatural note, Marks points out that the term *disposition* is sometimes used to refer solely to affective-motivational aspects, while *cognition* is used to refer to cognitive dispositions. In this thesis, the nomenclature provided by Blömeke et al. (2015) will be used, who refer to cognitive or affect-motivational dispositions.

Competence is specific for a domain

Competence always relates to a specific domain. The phrase ‘to be competent’, when not used in a rather mundane context, is usually supplemented by specifying the area which one claims to be competent *in* (Marks, 2023, p. 46; Müller-Ruckwitt, 2008, p. 181) – signaling that it is situation-, domain- and content-specific (Vogelsang, 2014, p. 29). Weinert further establishes the concept of action competence, which has been used in the context of occupational research to conceptualize the competences needed for professions. It can be defined as follows:

The theoretical construct of action competence comprehensively combines those intellectual abilities, content specific knowledge, cognitive skills, domain-specific strategies, routines and subroutines, motivational tendencies, volitional control systems, personal value orientations, and social behaviors into a complex system. Together, this system specifies the *prerequisites required* to fulfill the demands of a particular professional position. (Weinert, 2001, p. 51; emphasis added)

Thus, action competence is specific to a certain professional domain and its innate requirements. It is essential that people who want to work in a specific professions, e.g., teachers, acquire these specific action competences, as these are needed to perform the tasks of the profession, such as creating learning opportunities or providing the necessary feedback in the learning process (Marks, 2023, p. 46).

In summary, competence

- can be learned or acquired by engaging in specifically-designed learning scenarios,
- includes cognitive as well affective-motivational dispositions,
- can be measured,
- is manifested in performance, and
- is specific for a domain and is called action competence when being used to describe prerequisites for a profession.

The following chapter will further explore how that action competence has been conceptualized for teacher education.

3.2 Models of Competence in Teacher Education

Against this backdrop and understanding of competence, a selected overview of models that specifically seek to conceptualize competence in teacher education will be provided. Specifically the models of Baumert and Kunter (2006, 2013), Neuweg (2014) and Blömeke et al. (2015) are focused on.

3.2.1 Model of Professional Competence

As a consequence of *PISA 2000*, the *DFG*¹⁹ funded program COACTIV was embedded in the 2003/2004 PISA cycle (Baumert et al., 2013, p. 2). The aim, among others, was to examine the professional knowledge of mathematics teachers. As part of that project, the prominent model of professional competence²⁰ of teachers was developed by Baumert and Kunter (2006, 2013)²¹:

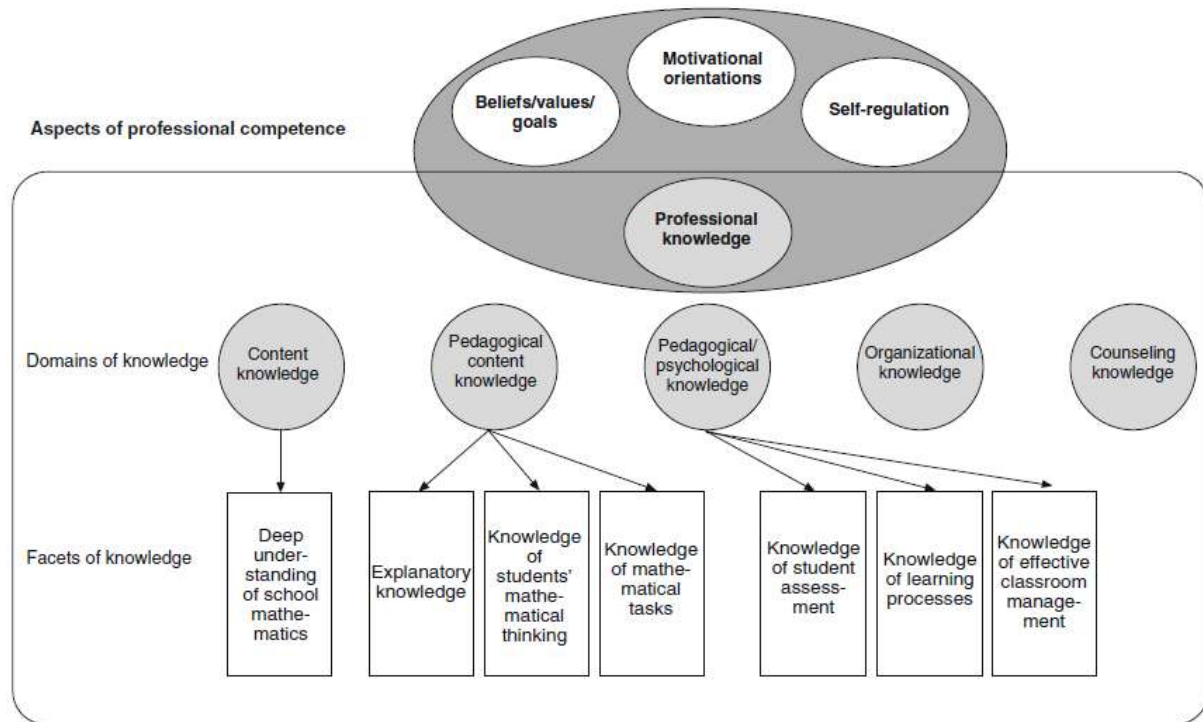


Figure 4: Model of professional competence used in the COACTIV study (Baumert & Kunter, 2013, p. 29).

The model is created as a generic model for teacher competence that combines previous research on competence and can be used and tested in their empirical investigations (Baumert & Kunter, 2013, p. 25). The model succeeds in that it connects “the demands of a multi-dimensional competence-oriented profession [...] with the analysis of demands of professional practice” (S. Meier, 2015, p. 61; translated by TJ). The model of professional competence consists of four aspects (see Figure 4): “Specific declarative and procedural knowledge (competence in

¹⁹ *DFG* is the abbreviation of *Deutsche Forschungsgemeinschaft*, the German Research Foundation.

²⁰ Interestingly, Baumert and Kunter have dropped the apposition ‘action’ for the model description in the English-language publication (2013) while referring to “professioneller Handlungskompetenz” in the older German-language publication (2006, p. 482). For consistency, the term *professional competence* will be used when referring to either publication of the model.

²¹ The model serves as an example for other similar models that conceptualize action competence similarly and build upon this model (see for example Lehmann-Grube and Nickolaus (2009)).

the narrow sense: knowledge and skills) / Professional values, beliefs, and goals / Motivational orientations / Professional self-regulation skills" (Baumert & Kunter, 2013, p. 28). Using these factors, a clear connection to Weinert's model of competence is evident in that Baumert and Kunter include cognitive and affect-motivational aspects their model. Regarding the cognitive aspect, the authors structure this disposition into different knowledge facets that are included in Shulman's taxonomy (1986).

Shulman distinguishes between three facets: *general pedagogical knowledge* (PK), *content knowledge*, differentiated in the two sub-facets of *subject-matter content knowledge* (CK) and *pedagogical content knowledge* (PCK), and *curricular knowledge* (Shulman, 1986, p. 9). Later, Shulman slightly modifies the categories by putting the previously mentioned four facets on an equal footing as categories of knowledge and adding the categories *knowledge of learners and their characteristics*; *knowledge of educational contexts* and *knowledge of educational ends, purposes, and values, and their philosophical and historical grounds* (Shulman, 1987, p. 8). Shulman's taxonomy has been quite influential in recent years, having been adapted by German researchers (see Bromme, 1997) and modified to include additional knowledge categories that, for example, include knowledge of technical aspects (see Rodríguez Moreno et al., 2019). This development is to be seen positively, as Shulman's proposal should be regarded as a "heuristic model" (Vogelsang, 2014, p. 39; translated by TJ) not a de facto structure of knowledge. It can and should thus be updated, expanded and modified to include knowledge aspects that are relevant and, in the best case, can be empirically proven to be relevant for teachers' professional competence. For the COACTIV study, Baumert and Kunter adapt three of Shulman's knowledge types: CK, PCK and PK, modifying the latter to also include aspects of psychological knowledge (2013, p. 30). They adapt only these three, arguing from a rather pragmatical point of view that these dimensions have prevailed in previous studies, but do not give any substantial reasons why the other aspects are excluded (Baumert & Kunter, 2006, p. 482). The categories of *organizational knowledge* and *counseling knowledge* are added to the generic model, but neither is included in COACTIV "for reasons of research economy" (Baumert & Kunter, 2013, p. 36).

The three knowledge types PK, CK and PCK represent the dominant aspects of teachers' cognitive dispositions (Krauss et al., 2020, p. 311) and are a presumably essential prerequisite for performance. In COACTIV, PK includes the knowledge of student assessment, learning process and effective classroom management (Baumert & Kunter, 2013, p. 29). While Vogelsang states that there is no generally accepted concept of the internal structure of PK (2014, p. 41), Baumert and Kunter claim that there is a "broad consensus" (2013, p. 34) of the

facets listed in Figure 5, which additionally include historical aspects of education and knowledge about empirical social research, systemizing previous research.

Table 2.1 Facets of general PK

1. Conceptual knowledge of the <i>foundations</i> of education
– Educational philosophy, educational theory, and the historical foundations of schooling and instruction
– Theory of institutions
– The psychology of human development, learning, and motivation
2. General pedagogical knowledge of instructional planning
– Metatheoretical models of lesson planning
– Domain-general principles of lesson planning
– Instructional methods in the broad sense
3. Knowledge of classroom management and orchestration of learning opportunities
– Patterns of instructional practice
– Variation of social forms and methods of learning
– Rules and routines of effective classroom management
– Creating a constructive and supportive learning environment
4. Knowledge of domain-general principles of diagnostic testing and assessment
– Learning and achievement: basic diagnostic skills
– Assessment and evaluation of learning processes
– Feedback
– Summative testing and assessment
5. Basic knowledge of the methods of empirical social research

Figure 5: Facets of PK (Baumert & Kunter, 2013, p. 35).

CK includes deeper knowledge about the content that the teacher needs to convey to their learners. This is naturally dependent on the subject – an EFL teacher, for example, is supposed to have knowledge about the English language, anglophone culture and literature. Within each subject, a consensus should be reached with regard to the content learners must acquire in that subject during their time in school. These are usually documented in governmental guidelines like the supranational *Common European Framework of References* (CEFR) (Council of Europe, 2001, 2020) or the national *BISTAs* (KMK, 2004, 2023a). Teachers' knowledge should exceed the content described in these documents as they need to have a deeper understanding of the subject-matter (Baumert & Kunter, 2013, p. 29). This knowledge is also prescribed in statutory guidelines (see Chapter 3.3).

PCK is described by Shulman as "the blending of content and pedagogy into an understanding of how particular topics, problems, or issues are organized, represented, and adapted to the diverse interests and abilities of learners, and

presented for instruction" (1987, p. 8). While the idea of content specific knowledge is generally applicable to all subjects, it is, at the latest, with this facet that Baumert and Kunter leave the generality of their model and turn towards the subject-specificity of COACTIV, focusing on mathematics teachers (2013, pp. 31–34). The specific facets of PCK for mathematics teachers, which they describe as explanatory knowledge, knowledge of students' mathematical thinking, knowledge of mathematical tasks, cannot directly be adapted to other subjects, which quite likely require different kinds of PCK. In FALKO-E²², a successor of COACTIV, PCK for English teachers includes the first two of the previously described facets, but modifies the third facet and frames it as "knowledge about the learning and teaching potential of tasks and texts" (Kirchhoff, 2017, p. 120; translated by TJ). This modification is also made in the other language-oriented *FALKO* tests, as these subjects have a larger focus on the use of texts and their potential for the acquisition of cultural, literary or language-related competences (Kirchhoff, 2017, p. 121).

While the model of professional competence puts an emphasis on cognitive aspects, it also addresses several affective-motivational dispositions like "[beliefs]/values/goals"; "[motivational] orientations" and "[self-regulation]" (Baumert & Kunter, 2013, p. 29). The first of these aspects contains epistemological beliefs, subjective theories about teaching and learning, value commitments and goal systems (Baumert & Kunter, 2006, p. 497, 2013, pp. 37–38). Following Fenstermacher's widely applied distinction (Baumert & Kunter, 2006, p. 496), knowledge is the "justified true belief" (Fenstermacher, 1994, p. 8) and is categorically different from belief²³. Beliefs do not have to commit to the same standards as knowledge as they, for example, do not necessarily have to be consistent or justified (Baumert & Kunter, 2013, p. 37). Teacher beliefs or teacher belief systems might still influence classroom action, potentially more than knowledge (Pajares, 1992, p. 311). In COACTIV, beliefs are again distinguished (Baumert & Kunter, 2013, p. 37). *Epistemological beliefs* can be seen as teachers' meta-cognition about knowledge and how it is acquired, represented and structured (Kitchener, 2002, pp. 92–93). *Subjective theories about either teaching or learning* include assumptions that teachers have made, based on their own

²² For further information on FALKO-E, see Chapters 3.4 and 15.5.3.2.2. in which the test will be further described. FALKO is the abbreviation of *Fachspezifische Lehrerkompetenzen*, and the prefix *E* refers to the subject of English (Krauss, Lindl, Schilcher, Fricke, et al., 2017).

²³ There is not necessarily consensus on the statement that knowledge and belief are categorically different, as is noted by Vogelsang (2014, p. 42) or Baumert and Kunter (2006, p. 496); (2013, p. 37) and it is difficult to make that strict differentiation.

previous experience, and are hence also subject-specific²⁴. For EFL teachers, for example, beliefs about language learning and second language acquisition are of particular importance and have been investigated (Mattheoudakis, 2007). For the sake of completion, the notions of *values* and *goals*, two aspects that are not considered of particular interest in competence research, will be presented briefly (S. Meier, 2015, p. 76). *Value commitments* include teachers' ideas of professional moral and ethos, and how dedicated they are in terms of executing certain values like "justice, caring [and] veracity" (Harder, 2014, p. 172; translated by TJ; see also Oser, 1996, p. 236). *Goal systems* describe teachers' assumptions with regard to what aims school as an institution should achieve (Vogelsang, 2014, p. 43).

Motivational orientation as a disposition is responsible for teachers' "psychological functioning" and includes "teachers' self-related cognitions — especially control beliefs and self-efficacy beliefs [...] — and intrinsic motivation" (Baumert & Kunter, 2013, p. 38). While the concept of motivation has a long research history (e.g., Wigfield & Eccles, 2020; Woolfolk, 2010, p. 378) and spans a multitude of influential theories (Ryan & Deci, 2000; Wigfield & Eccles, 2000), the model of professional competence by Baumert & Kunter focuses only on teachers' intrinsic motivation (Baumert & Kunter, 2013, p. 39). Generally, intrinsic motivation describes activities that are based on the inherent satisfaction that they entail, e.g., experiencing joy while playing football, while extrinsic motivation focuses more on the external reward that the completion of the activity will have, e.g., financial compensations (Ryan & Deci, 2000). Within intrinsic motivation, Baumert and Kunter primarily focus on teacher enthusiasm (2013, p. 39), which can be "conceptualized as the enjoyment, excitement, and pleasure that teachers experience during teaching" (Lazarides et al., 2018, p. 2). The enthusiasm can be directed towards the activity, e.g., teaching, or the topic or subject, e.g., English (Kunter, 2013, p. 277). While the concept of control beliefs includes a person's beliefs on the outcome that a certain behavior triggers (Bandura, 1997; Holzberger et al., 2013, p. 774), it has largely been replaced by Bandura's concept of self-efficacy (Baumert & Kunter, 2006, p. 502). Self-efficacy²⁵ can be defined as a person's confidence or trust in their capability of dealing with

²⁴ Vogelsang argues that the use of subjective theories by Baumert and Kunter as an umbrella term for beliefs is misleading, as both aspects cannot be equated (2014, p. 43); others do not make that strict distinction (S. Meier, 2015, p. 77).

²⁵ The use of the terms self-efficacy, self-efficacy beliefs and self-efficacy expectations as well as their German counterparts *Selbstwirksamkeit*, *Selbstwirksamkeitsüberzeugung* and *Selbstwirksamkeitserwartungen* is inconsistent in research. As this concept is not of primary concern for this thesis, trying to systematically define these terms and the terms of the respective source used will be refrained from.

complex tasks (Holzberger et al., 2013, p. 774) and is rooted in Bandura's social-cognitive theory (Bandura, 1997; S. Meier, 2015, p. 74).

Lastly, self-regulation in the COACTIV model is defined as "teachers' ability to budget personal resources in the professional context" (Klusmann, 2013, p. 291). This facet of professional competence aims more at teachers' health as it covers how teachers can cope with the predominantly mental and emotional challenges, demands and stress that the job of a teacher includes. The underlying assumption is that teachers with increased self-regulatory skills are better equipped to deal with more demanding situations, leading to a generally higher well-being, which in turn, increases a) teacher persistence, as well as b) their instructional quality (S. Meier, 2015, pp. 75–76).

Generally, while also focusing on non-cognitive aspects, the model of professional competence can best be described as knowledge-oriented as it constitutes knowledge as "[the] core of professionalism" (Baumert & Kunter, 2013, p. 28), assuming that teachers' knowledge is a prerequisite for action in the classroom. However, this assumed relationship has sparked further discussion concerning the extent to which knowledge can be used in practice, which will be addressed in the next chapters.

3.2.2 Competence as Knowledge

Knowledge that PSTs are meant to acquire during their education is, as Neuweg points out, not an end in itself, as its purpose is to be enacted in the action of the job (2022a, p. 11). The relationship between the two concepts is still debated contentiously and requires a brief definitory introduction before it can be discussed further. Neuweg's model (2014) has often been used to differentiate between different concepts of teacher knowledge (see Figure 6).

As *knowledge 1*, Neuweg conceptualizes knowledge that is canonically and systematically considered essential to be acquired in teacher education, also calling it professional knowledge (D. Gerlach, 2022, p. 36; Neuweg, 2014, p. 584). *Knowledge 1* can be distinguished from the other concepts in that it can mostly be considered as objectively true, being subject to scientific scrutiny, which is why it is also referred to as "research knowledge" (Wieser, 2016, p. 595). An exploration of what constitutes *knowledge 1* for foreign language teachers shows that, for example, foreign language teachers should know about competence orientation and language acquisition (Caspari & Grünwald, 2022, p. 110). *Knowledge 2*, then, is classified as what the teacher actually knows, describing their actual cognitive representations of knowledge. It can also be described as "personal knowledge" (Wieser, 2016, p. 595). In comparison to *knowledge 1*, this kind of knowledge does not have to be verifiably true, and is considered to be subjective. It includes both explicit knowledge, e.g., what teachers are able to

retrieve when asked for reasoning action, and implicit knowledge, e.g., what teachers cannot necessarily verbalize, with both being potentially influential for action (D. Gerlach, 2022, p. 36). *Knowledge 3* is the external reconstruction of action by an observer and is thus attributed to a person, e.g., by a researcher (Neuweg, 2014, pp. 584–585). *Knowledge 3* can also be described as “practical knowledge” (Wieser, 2016, p. 595) – or “performance” (Vogelsang, 2014, p. 111).

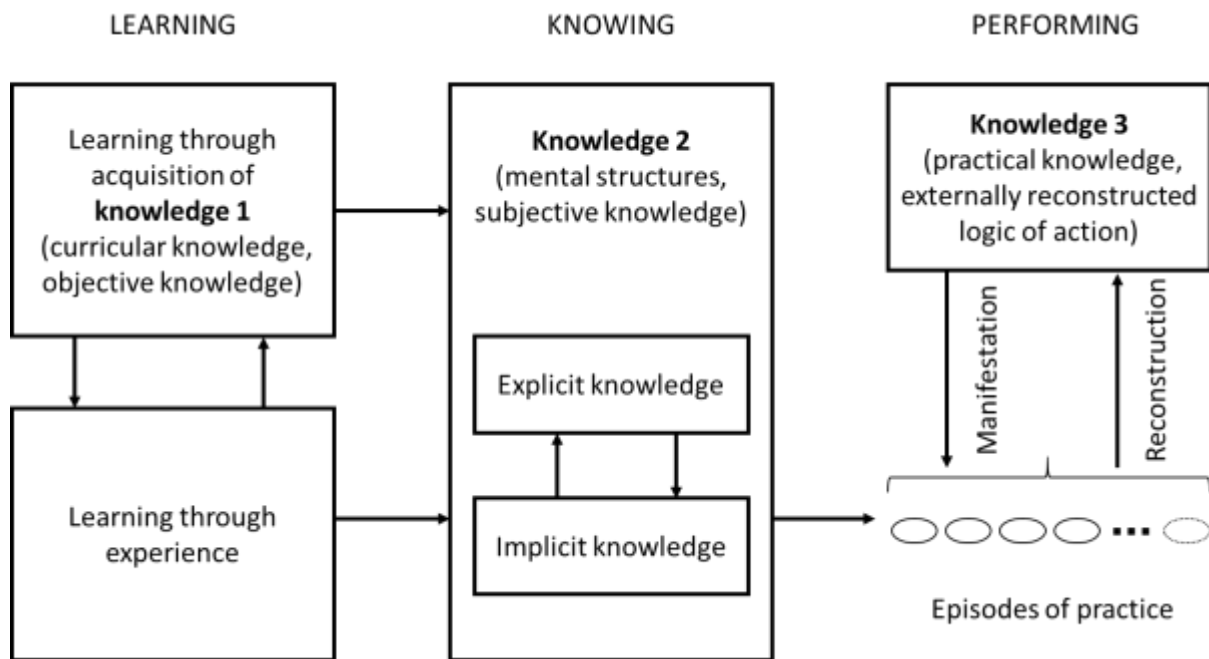


Figure 6: Concepts of “teacher knowledge” (Neuweg, 2014, p. 585; translated by TJ in concurrence with Wieser, 2016, p. 595).

Putting Neuweg’s thoughts on knowledge and the model of professional competence into relation, some tensions can be ascertained. Many aspects that Baumert and Kunter consider to be part of professional competence, e.g., cognitive dispositions in general, but also beliefs, could be subsumed as being part of *Knowledge 2* (2013). Neuweg makes a clear distinction between teachers’ cognitive structures (*knowledge 2*) and their action (*knowledge 3*). According to Neuweg (2022a, p. 23), this clear distinction between knowledge and action, and knowledge and competence is lacking in the model of Baumert and Kunter, who argue that “knowledge and ability” are both aspects of professional competence (Baumert & Kunter, 2006, p. 481) and that practical knowledge is manifested as ability (Baumert & Kunter, 2006, p. 483). While it is clear that knowledge (*knowledge 2* in Neuweg’s terms) is the core aspect in the model of professional competence, its relation to ability and performance (*knowledge 3*) is not made clear, arguably to a degree that no real difference can be made (Neuweg, 2022b, p. 1300). The root issue seems to be a historically reasoned dichotomic

distinction between competence and performance. Since Chomsky, competence and knowledge are both considered to be internal latent traits of a person that are, to some extent, made visible in their performance. Thus, both knowledge and competence are made out to be dispositions for performance and the two terms, though not necessarily used synonymously, are often closely linked together and knowledge is considered to be a main aspect of teachers' professional competence. One approach that tries to consolidate this discussion is the idea from Blömeke, Gustafsson & Shavelson of "[competence viewed] as a [continuum]" (2015).

3.2.3 Competence as a Continuum

The review article "Beyond Dichotomies" (2015) by Sigrid Blömeke, Jan-Eric Gustafsson and Richard J. Shavelson has had a considerable impact on the conceptualizations of teacher competence. The authors' main contribution is a "'grand' structural model" (Blömeke et al., 2015, p. 3) that to their mind, overcomes "conceptual controversies" (Blömeke et al., 2015, p. 3) of higher education competence research by modeling competence as a continuum (see Figure 7). This model will be presented first and then compared to the model of professional competence.

The point of departure for Blömeke et al. lies in the difficulty of "complex ability constructs that are context-specific, trainable, and closely related to real life" (Koeppen et al., 2008, p. 61). They describe the titular dichotomy as two concepts of competence. On the one hand, there is a more traditional, analytical interpretation focusing on complex abilities, which is rooted in educational science and its objective in fostering competence. On the other hand, there is a more holistically-oriented interpretation that focuses more on behavior in real-life professional situations, which has its origins in selecting adequate candidates for a job (Blömeke et al., 2015, p. 5). The authors, however, find similarities in the two conceptualizations, in that, 1) competence is being constituted by multiple smaller dimensions (competencies) of various kinds (e.g., cognitive, affective, motivational), which can be acquired, and 2) competence refers to the mastering of authentic real-life situations as a model or norms (Blömeke et al., 2015, pp. 5–6). Using these similarities as a foundation, the authors further aim to overcome the remaining dichotomous view of " (a) competence *is* performance in real-world situations, more specifically, whether behavior is the focus of competence, or (b) behavior is the *criterion* against which cognition and affect-motivation are validated as measures of competence" (Blömeke et al., 2015, p. 6; emphases in the original). Whereas the first interpretation argues that the previously addressed competencies influence performance dynamically *in situ*, the second approach sees these competencies as summands that make up competence, which

rather directly lead to performance in a cumulative sense (Blömeke et al., 2015, p. 6). The authors then point out that this rather polarized research perspective neglects research concerning the nature of the bond between disposition and performance (Blömeke et al., 2015, p. 7). As a conclusion, they propose that “instead of insisting on an unproductive dichotomy view of competence, in particular knowledge or performance, competence should be regarded as a process, a continuum with many steps in between” (Blömeke et al., 2015, p. 7; emphasis in the original).

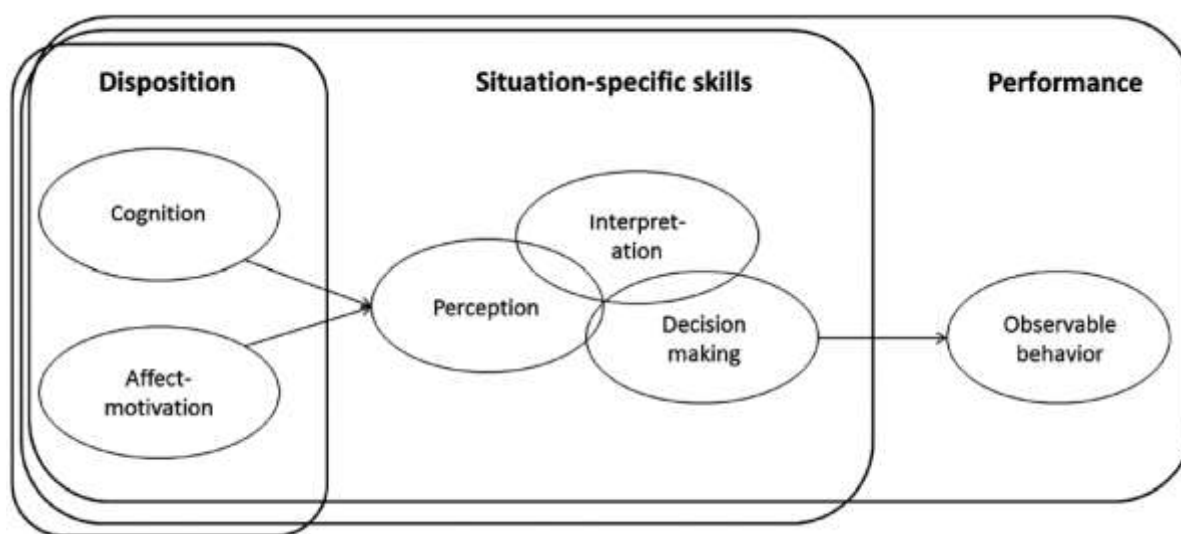


Figure 7: Model of competence viewed as a continuum (Blömeke et al., 2015, p. 7).

The proposed model differentiates three different “steps” (Blömeke et al., 2015, p. 7) of competence: disposition, situation-specific skills and performance. The dispositional aspect most clearly relates to the model of professional competence in that it encompasses cognitive and affect-motivational facets, similar to the model by Baumert and Kunter (2006, 2013) and Neuweg’s idea of *knowledge 2*. This aspect addresses the rather analytical perspective of competence. Performance, on the other end of the continuum, addresses the requirement to actually act in situations and display one’s competence, resembling *knowledge 3* (Neuweg, 2014, p. 585). In between those two ends of the continuum are situation-specific skills that “may mediate” (Blömeke et al., 2015, p. 7) between both ends, e.g., professional vision (Weger, 2019, p. 15)²⁶. A

²⁶ The question of how professional knowledge and performance interact has also been raised by Bromme who, similarly to the presented model, discusses teachers’ routines in classroom situations and their diagnostic competence as situated between teachers’ knowledge and their classroom activities (1997, pp. 198–200). A secondary reception of Bromme’s model by Frey and Jung can also be seen as a precursor for the model by Blömeke et al. (2015) (2011, p. 16).

hypothetical feedback conversation between a teacher and a learner in the English language classroom may serve as an example of this model. In this scenario, the learner asks the teacher for some feedback on a text, but time is limited. The teacher scans the text and diagnoses several important aspects for the learner to improve upon, but decides to focus on one specific aspect they²⁷ consider to be the most important, such as the text missing specific genre aspects. The teacher is well-capable of using their knowledge of the English language, and how to provide adequate feedback (dispositions) but given the limited time in the situation and the volume of feedback opportunities, they decide to prioritize a certain aspect (situation-specific skills) and provide feedback on only one aspect (performance)²⁸. In the end, it is the teacher's performance in the situation that is crucial for the success of this learning opportunity for the learner.

By dissolving the dichotomy, and stating that competence is both knowledge and performance, Chomsky's initial definition is not necessarily refuted, but rather clarified, as he regards competence as knowledge (dispositions), with performance as the observable use of that knowledge (Chomsky, 1965, p. 4). The model by Blömeke et al. (2015) expands on that by adding situation-specific skills as a mediator that could explain why knowledge does not necessarily translate into performance. Weinert makes a similar point, criticizing that many competence definitions neglect the context of performance, potentially leading to invalid assumptions of competence (2001, pp. 58–59). Critically, it should be noted that not unlike the model of professional competence, the model only takes the teacher's perspective into account, not considering whether the performance has an impact on the learner (Krauss et al., 2020, p. 314). Nonetheless, Blömeke, Gustafsson and Shavelson succeed in that they are able to connect different strands of competence research, allowing to relate the different facets to each other under one united framework.

The respective positioning on competence has considerable implications for teacher education. Following the assumption laid out by Baumert and Kunter (2013), professional knowledge and other dispositions will eventually lead to performance in the classroom. Within this paradigm, teacher education then has as its goal the development of these dispositions. Understanding performance as a constituting aspect of competence, as is done in the model by Blömeke et al. (2015) implies that the performative aspect must be considered in teacher education. Consequently, as all three phases of teacher education regard the

²⁷ Singular *they* is used in this thesis to avoid assigning genders and to comply with data protection when it relates to specific participants in the empirical part of this thesis.

²⁸ See also Marks (2023, p. 46).

development of competence as their goal, they should all include the performative aspect of competence to some extent.

In order to discuss the understanding of competence in German teacher education, the following two chapters will first describe the relevant frameworks for EFL teacher education in Germany and, second, present selected, relevant empirical findings that describe the extent to which the goals that are set, are obtained.

3.3 Frameworks for (EFL) Teacher Education in Germany

While the general goal of teacher education has been outlined on a rather meta-level in Chapter 2.1, the specific objectives that are required for EFL teacher education in Germany are presented in the following. Primarily, this serves as a description of the different requirements of each phase. Additionally, any test, like the RobS, that is intended for use in teacher education, should be anchored in the statutory guidelines, which makes a brief description necessary²⁹. Before that, the context of these frameworks will be provided by way of a brief historical outline.

Towards Standardization in German (EFL) Teacher Education

According to Terhart, “[the] history of teacher education in Germany can be described as a continuous process of classical modernization” (2019, 4), which is to say that the only real constant, throughout its history, is change³⁰. The “disastrous” (Marks, 2023, p. 18; translated by TJ) results of the *PISA 2000* (Baumert, 2001), sent shock waves through Germany’s society on all levels as the so-called “PISA Shock” (Terhart, 2021, p. 57; translated by TJ) and made a systematic change in (teacher) education unavoidable (Marks, 2023, p. 18). A shift towards what learners should be able to do, often phrased as ‘output’ orientation, and away from what should be learned (i.e., input-orientation) commenced (Klippel, 2018, p. 25; Marks, 2023, p. 18)³¹. Similarly, an input-oriented governance in teacher education was replaced by a focus on output (Blömeke, 2006, p. 26; Marks, 2023, pp. 18–19). The premise is that better teacher education leads to an improved in-class teacher performance and finally to an increase in learner achievement (König, 2014, pp. 18–19; Terhart, 2012a). The adoption of this idea in Germany followed an international movement towards more standardization

²⁹ This notion refers to curricular validity, see Chapter 10.2.1.

³⁰ For a more comprehensive overview on the history of language teacher education in Germany specifically, see Klippel (2018, 2019, 2022) or Doff (2019).

³¹ For a more detailed discussion of the implementation of competences in education post-PISA, see Marks (2023, 18 ff.).

and output orientation in teacher education that started at the end of the 1980s (Klieme et al., 2008, p. 3). In the German-speaking regions, this movement was spearheaded by the Swiss researchers Oser and Oelkers (2001) (see also Blömeke, 2006, p. 26; Marks, 2023, p. 24; Terhart, 2014, pp. 303–304). Both developments, PISA and the movement towards standardization, induced the KMK to take action. The purpose of the KMK is the cross-federal³² governance of all education-related aspects, thus the education of teachers is one of their concerns. In this role, they introduced a first document outlining *Standards für die Lehrerbildung: Bildungswissenschaften*, which translates to *Standards for teacher education: Educational sciences* (translated by TJ)³³, in the following referred to as *KMK04(/19)*³⁴. The publication was preceded by a general discussion of the scope of the standards, with the KMK first seeking Terhart's expertise (2002). The KMK followed Terhart's expertise in many aspects, but ignored it in one of its key propositions: Developing standards for "all areas and levels of teacher education" (Terhart, 2014, 303; emphasis in the original; translated by TJ). Although the development of standards for the subjects and subject-pedagogies (including those for new foreign languages like English or French) were planned to follow the publication of *KMK04* (Terhart, 2014, p. 305), members of these disciplines still did not feel the subjects to be adequately represented (Marks, 2023, p. 26). In the year 2008, the *Ländergemeinsamen inhaltlichen Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung*, in the following referred to as *KMK08(/19)*³⁵, were introduced as the standards for the specific subjects, like English (as part of the new foreign languages) (KMK, 2019a). Consequently, efforts have been made for teacher education in Germany to become more standardized and output-oriented, principally enabling an evaluation of the acquisition of teachers' professional competence. In the following, a selected overview of *KMK04/19*, with a focus on those standards relevant to this thesis,

³² In Germany, (teacher) education is mostly under the jurisdiction of the individual federal state, thus the KMK serves as a body that allows the federal states to try to achieve unified decisions for Germany.

³³ The discourse around the German terms *Bildung* and *Erziehung*, both often subsumed under and translated with *education* (Biesta, 2016, p. 832) is an inherently German phenomenon, dating back to Humboldt and Kant (see, for example, Koller, 2014). In the context of this dissertation, the term *educational sciences* is used as a translation of *Bildungswissenschaften*, concurring with how Terhart summarizes the KMK's pragmatic definition as all the aspects of teacher education, that are not part of the individual subject, e.g., Chemistry or English or their individual subject-pedagogy (German: *Fachdidaktik*) (2012, 25).

³⁴ Abbreviation in concurrence with Marks (2023). The updated version (2019) will be abbreviated as *KMK04/19*.

³⁵ Abbreviation in concurrence with Marks (2023). The updated version (2019) will be abbreviated as *KMK08/19*.

will be provided. As *KMK08/19* describes necessary competences for foreign language teachers, those will be presented in more depth.

KMK04/19

In *KMK04/19* four different competence areas are differentiated:

- Teaching/instructing (*Unterrichten*)
- Educating (*Erziehen*)
- Assessing/evaluating (*Beurteilen*)
- Innovating (*Innovieren*) (KMK, 2019b; translated by TJ).

Each competence area includes individual competences, eleven in total. Each competence is subdivided into descriptors that focus on the theoretical and practical part of teacher education, 88 descriptors in total (KMK, 2019b; Marks, 2023, p. 317). As an example, the competence area assessing/evaluating (competences 7 and 8) is described³⁶. Competence 7 states the following: “teachers diagnose learners’ learning predispositions and learning processes; they support learners in a target-oriented manner and counsel learners and parents” (KMK, 2019b, p. 11; translated by TJ). With regard to the theoretical descriptors, teachers are supposed to be knowledgeable about options of diagnosis and be able to diagnose and accommodate heterogeneous learners with diverse needs. They are supposed to possess knowledge of the factors influencing learning processes, be aware of the forms of especially gifted learners, the basics of learning process diagnostics, as well as principles of counseling students and parents. Finally, teachers are able to recognize the relevance of collaboration with various stakeholders in education. The practical descriptors include teachers’ ability to identify learning levels, potentials, obstacles and progress, as well as individuals’ initial learning situations, providing targeted support, including for highly gifted learners. Teachers are supposed to be able to adequately use forms of counseling, differentiating between situations of counseling and assessment, and be able to cooperate with pedagogical personnel and use digital tools for diagnostic purposes (KMK, 2019b, p. 11).

Competence 8 states that “teachers diagnose learners’ development of achievement and assess learning processes based on transparent assessment criteria” (KMK, 2019b, p. 12; translated by TJ). For the theoretical aspects of teacher education, a specification is made in that teachers are supposed to know different forms of assessment and feedback, including the impact, their function, as well as the (dis-)advantages. Moreover, they are supposed to be familiar with

³⁶ Considering the fact that the test in focus in this thesis is aiming at assessing prospective teachers’ competence in providing feedback, the competence area *assessing/evaluating* is of particular interest. A comprehensive description of all competence facets is, however, not considered necessary.

different frames of reference of assessment, and the tension that exists between formative feedback and “the societal functions of assessment” (KMK, 2019b, p. 12; translated by TJ). Lastly, teachers are also supposed to be aware of the potential and limitations of digital tasks and assessment formats. For the practical portion of teacher education, the standards include the conceptualization of tasks and the adequate application of assessment models and criteria. Teachers are supposed to agree on assessment criteria with their colleagues, be able to give reasons for their evaluation appropriate to the respective group, and also provide the next steps for future learning. Teachers are supposed to use digital tools for diagnostic purposes in a reflective way and also use the results of assessment as a form of feedback on their own actions as a teacher (KMK, 2019b, p. 12). Overall, Marks concludes that *KMK04* includes ten assessment-related descriptors (2023, p. 317).

KMK08/19 – New Foreign Languages

In *KMK08/19*, the subject-specific requirements for teacher education are added by focusing on the subject content (CK) and its pedagogy (PCK). The chapter of each subject or subject group includes two distinct sections. The first section lists the subject-competences, and the second one lists the content that the respective sub-area should cover in the first phase of teacher education.

The competence descriptors can be aligned to the different areas of study and the content they must cover. For new foreign languages, these areas are:

- Practical Language Use (*Sprachpraxis*),
- Linguistics,
- Literary Studies,
- Cultural Studies,
- Foreign Language Pedagogy (*Fachdidaktik der Fremdsprachen*) (KMK, 2019a, pp. 45–46; translated by TJ).

In the context of practical language use, students are expected to possess language skills that closely resemble native proficiency, which they need to keep updated. In the areas of language, literature, and cultural studies, students are expected to acquire an in-depth knowledge, including current topics and methodologies of research. A significant portion of the requirements relates to subject-specific pedagogy. Students are supposed to be able to utilize insights from the subject disciplines and specific pedagogical approaches to analyze, evaluate, and prepare texts for instruction from linguistic and cultural-literary perspectives. Additionally, they are supposed to be capable of presenting discipline-specific inquiries and scientific findings appropriately, and explaining the significance of the field. Students should also possess an understanding of foreign language learning processes, particularly those regarding multilingualism,

heterogeneity, and inclusion in instruction. Students are supposed to comprehend the communicative, intercultural, and text-related foreign language competence of learners and have initial experiences in planning, conducting, and reflecting on instructional units to support them. In the realm of inclusion, graduates are expected to collaborate with other pedagogical personnel and be sensitive to the need for accessible learning materials. Regarding digitalization, students are supposed to be able to track relevant developments in their field, reflect on them, and effectively utilize them for instruction, including individualized support. The aspect of diagnosis and support is mainly articulated within subject-specific didactic content (KMK, 2019a, p. 44).

Pertaining to the content of studies, only the content that is required in the area of PCK is referred to. The *KMK08/19* includes a list of eleven elements that the students must acquire knowledge about:

- language acquisition theories and individual requirements of language acquisition, also against the backdrop of migration-relation multilingualism and intercultural contexts
- subject-pedagogical specific approaches to diagnosis, assessment, and support that is based on the results
- theories and methods of communicative foreign language learning
- target-, learner- and subject -appropriate planning, execution and reflection of competence-oriented foreign language instruction, considering the individual needs in heterogeneous learning groups
- theories, targets and processes of linguistic and intercultural learning, including their implementation in class
- literature-, text-, culture- and media-pedagogical theories, aims and processes
- subject-pedagogical specific characteristics for the respective foreign language classroom
- demands of bilingual³⁷ learning and instruction
- concepts, media and methods of an inclusive foreign language classroom
- types of cooperation in the classroom with special needs teachers and other pedagogical personnel concerning planning, execution and reflection of inclusive instruction
- subject-related text competence as a basis for the selection and adaptation of authentic texts for the foreign language

³⁷ In the context of the German educational system, instruction of subject matter with a vehicular language other than German is referred to as bilingual, although the terminology of CLIL (Content and Language Integrated Learning) is more common internationally (Rumlich, 2015, p. 41).

classroom and the development of text-based tasks. (KMK, 2019a, p. 46; translated by TJ)

The document combines the requirements for all new foreign languages in one set, thus it does not consider differences between the languages. The following chapter will provide insights into empirical findings that focus on measuring competence facets that are outlined in these frameworks. A focus is placed on knowledge, as it plays a prominent role on the models of teacher competence, as well as the guidelines by the KMK (Blömeke, 2006). This allows for an exploration of the extent to which the goals of EFL teacher education, as set out by these guidelines, are achieved.

3.4 Empirical Findings on Professional Knowledge of Pre-Service EFL Teachers

In this chapter findings are presented with regard to the professional knowledge of PSETs³⁸. As previously discussed, the basic assumption in teacher education is that knowledge acquired at university leads to performance in instructive situations. The findings may provide an indication of the extent to which teacher education programs contribute to the competence development of EFL teachers in Germany. What is more, this concentration on cognitive dispositions in this chapter is reasoned as knowledge is often considered the primary aim of university-based teacher education, as indicated by the frameworks of the KMK. Moreover, the models of teacher competence, especially the model of professional competence by Baumert and Kunter (2006, 2013), also include aspects of teacher knowledge. However, the professional knowledge that EFL teachers need is debated, on the one hand, conceptually with regard to its definition and scope and, on the other hand, concerning the appropriate approaches for research (see, for example, the *ZFF*³⁹ special issue on foreign language-specific PCK by Grünewald et al., 2022⁴⁰). Research perspectives on foreign language-specific professional knowledge differ, as some strands follow a qualitative-reconstructive approach (for an overview, see Bonnett, 2024), concentrating on understanding reasons for teacher performance in the sense of Neuweg's *knowledge 3* (2014, p. 585). In the recent past, however, there has also been an effort to develop test instruments to empirically measure professional

³⁸ For reasons of cohesion, i.e., focusing in PSETs, other empirical findings in the EFL context that might be of interest and which do not specifically focus on knowledge are not included here. This relates especially to results of the *DESI* study (*Deutsch Englisch Schülerleistungen International*), in which video data on the actual performance of in-service EFL teachers is collected and analyzed (see Klieme, 2008).

³⁹ *Zeitschrift für Fremdsprachenforschung*, a journal focusing on foreign language education.

⁴⁰ Especially the contributions by Caspari and Grünewald (2022), Legutke et al. (2022) and D. Gerlach (2022).

knowledge quantitatively and evaluate the progress of professionalization by the means of knowledge development. There have been three prominent projects that have sought to develop and validate tests and use them to investigate the professional knowledge of (future) EFL teachers: TEDS-LT⁴¹ (Blömeke, Bremerich-Vos, et al., 2013; Blömeke et al., 2011), PKE⁴² (König et al., 2016; Roters et al., 2015) and FALKO-E (Kirchhoff, 2017)⁴³. The results of these projects will be discussed in this chapter, focusing mostly on CK and PCK, as they allow for a broader empirical perspective on the status quo regarding professional knowledge of (pre-service) EFL teachers and knowledge development throughout the different phases⁴⁴.

Several similarities can be found within these projects regarding their conceptualization. First, the differentiation of professional knowledge is alike in that Shulman's taxonomy was used to distinguish between the areas of PK, CK and PCK (Shulman, 1986, p. 9). Second, the descriptors of *KMK08* serve as a general theoretical foundation for constructing the items of CK and PCK (Marks, 2023, p. 93). In TEDS-LT, the individual university-specific curricula are also used for orientation and experts consulted (Roters et al., 2011, p. 80). In PKE, "[several] pilot studies and expert reviews were conducted to develop the final set of 33 test items" (König et al., 2016, p. 327). In FALKO-E, further items are included normatively for PCK, as *KMK08* only relates to the first phase of teacher education, and validated in expert interviews⁴⁵ with researchers and teachers (Kirchhoff, 2016, p. 85, 2017, p. 122). Third, professional knowledge is measured in written pen-and-paper tests, which are predominantly designed with closed or semi-closed items in TEDS-LT and PKE, whereas the FALKO-E test uses mostly open items (Marks, 2023, p. 95).

⁴¹ TEDS-LT is the abbreviation of *Teacher Education and Development Study – Learning to Teach* (Blömeke, 2011).

⁴² PKE is the acronym for *Professionelle Kompetenz von Englischlehrkräften: Fachdidaktisches Wissen angehender Englischlehrkräfte – Konzeption, Messung, Validierung* which translates to *EFL teachers' professional competence: Future EFL teachers' PCK – Conception, measurement, validation* König et al. (2016).

⁴³ The individual design and construction of the PCK-specific test parts is focused on in Chapter 15.5.3.2.2 in the context of finding appropriate instruments to measure PCK as an external variable. An in depth summary is also provided in Marks (2023).

⁴⁴ Empirical findings on EFL teachers that are consequently not included here are those focusing on qualitative data exploring the reflection of professional knowledge (for an overview, see Gerlach, 2022, pp. 32–33), affect-motivational dispositions (see Doll & Schwippert, 2011; Schwippert et al., 2013), situation-specific skills (for an overview on professional vision, see Hüttner, 2019; Weger, 2019) or other specific aspects of EFL teacher competence, e.g., diagnostic competence (Fleckenstein et al., 2018; Jansen et al., 2020; Sebastiani, 2023).

⁴⁵ Expert interviews were conducted as think-aloud interviews, see Kirchhoff (2016) for more information on the validation process.

In the TEDS-LT study, professional knowledge, as conceptualized by Baumert and Kunter (2006, 2013), is tested at two different times in the first phase to investigate the process of professionalization of PSETs (Jansing et al., 2013, p. 78). In the test, CK is measured in two dimensions, "British and American literary studies" and "linguistics" (Roters et al., 2011, p. 87; translated by TJ), while PCK is separated into three different subdimensions "(1) Methods/concepts/analyses, (2) processes/strategies/language acquisition, and (3) literature pedagogy / interculturality)" (Jansing et al., 2013, p. 80; translated by TJ). However, the differentiation of PCK occurred merely for the second measurement, as the psychometric statistics for the PCK part were considered insufficient for the first measurement (Jansing et al., 2013, p. 79). The results of T_1 ($n_1 = 462$) show that the dimensions of CK and PCK correlate positively ($r = .395$ to $r = .604$), indicating that the facets measure aspects of the same construct (Roters et al., 2011, pp. 94–95). As only 7 % of participants participated in both measurements ($n_2 = 420$; $N_G = 862$), the data is rather quasi-longitudinal, as primarily earlier and later semester groups are compared (Jansing et al., 2013, p. 81). What is more, the reliability of the scales is relatively low, complicating the interpretation of the results (Jansing et al., 2013, p. 80). Still, the comparison of both data points shows that a significant increase of knowledge is only evident for the first two PCK subdimensions, while the remaining PCK and CK dimensions are relatively stagnant (Jansing et al., 2013, p. 95). Interestingly, there are no significant correlations between PCK or CK in relation to PK for the entire cohort of PSETs, while there are for PSTs of German and Mathematics (König et al., 2011, p. 150)⁴⁶. Moreover, PK seems to be generally lower for students of English and German than for those of Mathematics (König et al., 2013, p. 160). Lastly, the final graduation grade⁴⁷ seems to be significantly predictive for the average professional knowledge of pre-service EFL teachers ($\beta = -.44$) (Blömeke, Buchholtz, & Bremerich-Vos, 2013, p. 179).

PKE can be classified as the sequel to TEDS-LT as it builds directly on that work. For instance, it includes the same instruments to measure PK⁴⁸ and CK (König et al., 2016, pp. 326–327). It differs in that it uses the Cambridge Placement Test to measure language proficiency (LP) (König et al., 2016, p. 327),

⁴⁶ There is only a positive correlation between PCK and PK for the subset of pre-service EFL teachers studying for lower secondary (König et al., 2011, p. 151).

⁴⁷ Occasionally, this grade is compared to the *Grade Point Average* (GPA). However, in this thesis the last grade before entering teacher education is collected, which can also relate to other grades, e.g., grade of apprenticeships.

⁴⁸ The authors use the abbreviation GPK for general pedagogical knowledge. For consistency, it will be referred to as PK in this thesis.

whereas in TEDS-LT participants are only asked to self-evaluate their proficiency based on the CEFR scale (Roters et al., 2011, p. 84). Also, PKE focuses on further developing the items to measure PCK, as the results of TEDS-LT are not considered to be sufficiently reliable. PCK is conceptualized along the “three content dimensions: knowledge of curriculum, teaching strategies, and students” (König et al., 2016, p. 327), and sufficient reliability scores are achieved this time (König et al., 2016, p. 328). Findings indicate that PCK highly correlates with CK and PK ($r = .69$), while there are moderate correlations between CK and PK ($r = .40$), as well as LP and CK ($r = .52$), and LP and PCK ($r = .43$), while correlations are rather low for LP and PK ($r = .24$), which is not surprising considering that these constructs are the furthest apart content-wise (König et al., 2016, pp. 328–329). As part of PKE, knowledge of PSETs is measured in two different phases quasi-longitudinally, i.e., Master students in the first phase ($n = 216$), and teacher-trainees in the second phase (induction) ($n = 228$) (König et al., 2016, pp. 326–328). Findings indicate that PK and PCK is significantly higher for the group in the induction phase, while CK remains stagnant (König et al., 2016, p. 329; Tachtsoglou & König, 2018, p. 24). Results seemingly align with findings of TEDS-LT as during the process of professionalization, PCK seems to increase, while CK does not. This finding is corroborated by König et al., who find even more pronounced differences for PK and PCK between students in their Bachelor and Master phase (2017, p. 29). Taking the graduation grade into account, it is significantly predictive for CK ($\beta = -.27$), PCK ($\beta = -.32$) and PK ($\beta = -.18$)⁴⁹.

FALKO-E is the EFL-specific part of the larger FALKO project in which multiple disciplines collaborated to develop pen-and-paper tests to evaluate professional knowledge of teachers using the design of COACTIV as a connective framework (Krauss, Lindl, Schilcher, & Tepner, 2017, p. 10). In FALKO-E, subject-specific professional knowledge is similarly conceptualized as CK and PCK, while PK is apparently not included in this study (Kirchhoff, 2017, p. 118). In the CK test, deeper knowledge on relevant aspects that are taught in school is focused on in nine items, including aspects of linguistic, literary and cultural knowledge (Kirchhoff, 2017, pp. 120–122). The PCK test is presented in detail in Chapter 15.5.3.2.2 and thus not further elaborated on here. The FALKO-E test consists of mostly open items and is validated in a study with a sample of $N = 186$ participants. The sample is made up of two subsets, PSETs ($n = 106$) and in-service EFL teachers ($n = 80$). A significant positive correlation between PCK and CK is found ($r = .43$), further supporting the notions that both facets are part of the same knowledge domain. Comparing the two subsets, the findings indicate that in-service

⁴⁹ Negative coefficients are a result of the German grading system using lower numbers for better grades, see Chapter 15.5.4.2.

teachers achieve significantly more points in CK than pre-service teachers, while there is no significant difference with regard to PCK (Kirchhoff, 2017, p. 136). Moreover, the final graduation grade seems to be significantly predictive for both PCK ($\beta = -.25$) and CK ($\beta = -.30$). Interestingly, the professional experience of teachers does not seem to correlate significantly with either CK ($r = -.21$), or PCK ($r = -.05$), and also seems to be rather negative (Kirchhoff, 2017, p. 141).

While the results of these projects are insightful, they should be discussed with their limitations in mind. Firstly, the primary concern of these studies is that of validation. The test results are thus of an explorative nature and part of the process of test construction. This is especially evident in TEDS-LT, whose PCK items are completely reconsidered for PKE due to their insufficient reliability scores. Secondly, none of these studies are of a strictly longitudinal nature, which prohibits statements regarding actual professionalization processes, i.e., true growth in professional knowledge. Thirdly, results of these studies are partially contradictory. While there are significant differences of PCK in TEDS-LT when comparing Bachelor and Master students and in PKE when comparing Master students and teacher trainees, this is not corroborated by findings in FALKO-E. Conversely, the FALKO-E study only finds significant differences for CK, while this is not corroborated by either TEDS-LT or PKE. The following chapter will seek to discuss to what extent knowledge, which arguably develops to a certain extent as indicated by these findings, relates to PSTs' ability to perform in instructional settings.

3.5 Competence between Knowledge and Practice

Based on these theoretical and empirical findings, this chapter seeks to discuss the extent to which EFL teacher education in Germany actually achieves its goal of preparing PSTs for the tasks of the profession, focusing on the contributions of the first phase. The goal of teacher education is discussed based on several flow chart models, each of which puts teachers' performance at the center as it has the most direct influence on learner achievement (Diez, 2010; Krauss et al., 2020; Terhart, 2012b; Voss, 2019). Looking at the standards formulated in the statutory framework of the KMK, however, they are often criticized for focusing too much on the cognitive dispositions. Blömeke, for instance, claims that the descriptors in *KMK04* mostly focus on aspects of "knowing" (2006, p. 29; translated by TJ), e.g., knowing about different procedures of assessment, but do not include teachers' ability to that knowledge, consequently sidelining performative

aspects and locating them only in the second phase of teacher education⁵⁰ (Marks, 2023, pp. 316–317). Critically, Blömeke states the following:

In einer zweiphasigen Ausbildung kann man sicherlich nicht so weit gehen zu sagen, dass die Studierenden am Ende der ersten Phase unterrichten können. Dennoch sollen sie etwas können, was kognitiv anspruchsvoller ist als lediglich zu "kennen". (Blömeke, 2006, p. 30)

(In a two-phasic structure of teacher education, one can surely not go so far as to say that students should be able to teach at the end of the first phase. However, they should be able to do something more cognitively demanding than just "know".)

Universities, as the institution mainly responsible for the first phase, should thus not only focus on theory and declarative knowledge but go beyond and include instances of performance, thus laying the foundations adequately for the second phase.

This focus on theoretical knowledge is also reflected in the different models of competences, especially the model of professional competence by Baumert and Kunter, which conceptualizes knowledge as the "core of professionalism" (2013, p. 28). While the model suggests that knowledge can be a resource for performance, it does not clearly differentiate between competence and performance (Neuweg, 2022a, p. 23; Vogelsang, 2014, p. 45). Generally, there is apparently no clear distinction between knowledge and competence in the German discourse (Neuweg, 2022b, p. 1300). With the model proposed by Blömeke et al., performance is conceptualized as part of competence, going beyond existing dichotomies (2015, p. 7). If the development of competence is thus the goal for each phase in teacher education, the first phase should consequently not disregard its performative facet, but include it. Conceptualizing competence as more than knowledge and inclusive of performance also further emphasizes Blömeke's critique of the KMK guidelines as too narrowly focused on knowledge (2006, p. 30). This leads to the broader discussion of theory and practice in teacher education.

Described as an "unsolved problem" (Rothland, 2020, p. 133; translated by TJ), the discourse on the relation between theory and practice is dominant in teacher education, dating back to the beginning of the 20th century (J. Dewey, 1904; Pilypaitytė, 2013, p. 36). While there is a similar discussion in the legal or medical profession, it seems to be particularly intense in teacher education (Neuweg, 2022a, p. 12). A plethora of related discussions occur under this umbrella phrase, though generally *theory* refers to the research and science-related

⁵⁰ Baumert and Kunter make the same point, but refer to Terhart's expertise and not the *KMK04* standards (2006).

aspects of teacher education, i.e., acquisition of theoretical knowledge, whereas *practice* refers to the professional action, i.e., doing the actual job (Rothland, 2020, p. 133). The root cause for the discussion possibly stems from the approach of “theory-to-practice” (Carlson, 1999, p. 203): The university lays a foundation that consists of mostly theoretical or formal knowledge (Fenstermacher, 1994, p. 12), which can then be put into practice at school (Wideen et al., 1998, p. 167). While this approach has prevailed for decades and is reflected in the two-phasic structure of teacher education in Germany and models of competence, PSTs encounter challenges in actually applying this knowledge in practice, resulting in a perceived “gap between theory and practice” (F. A. Korthagen, 2010, p. 409).

While the acquisition of knowledge can be described as somewhat successful, as described in Chapter 3.4, several studies indicate that (pre-service) teachers encounter challenges in applying that knowledge (e.g., Zeichner, 2010). Specifically at the start of their professional life and in insecure situations, teachers turn to teaching how they have experienced it as a learner rather than what they were taught at university (Müller-Fohrbrodt et al., 1978, p. 39). There are multiple possible reasons for these findings, although at the core is the challenge of the application of knowledge in practice (Widlak, 1983, p. 213). Several reasons naturally factor into individual challenges teachers face when it comes to applying their knowledge (F. A. Korthagen, 2010). Rather recently, practical phases have been included into university-based teacher education in Germany with the distinct aim to narrow this perceived gap and facilitate the process of applying knowledge in practice (Ulrich et al., 2020, p. 6). As described by Zeichner, however, so-called “field-experiences” do not necessarily allow for the application of knowledge, either:

A perennial problem in traditional college- and university-sponsored teacher education programs has been the lack of connection between campus-based, university-based teacher education courses and field experiences. Although most university-based teacher education programs now include multiple field experiences over the length of the program and often situate field experiences in some type of school-university partnership (e.g., professional development schools, partner schools), the disconnect between what students are taught in campus courses and their opportunities for learning to enact these practices in their school placements is often very great even within professional development and partner schools. (2010, p. 91)

While students generally self-report that their competences increase after spending time at school, there is little empirical evidence that there actually is a measurable increase (for an overview, see Ulrich et al., 2020, p. 49). Within these practical phases, the focus has shifted more towards reflection, especially

“reflection-on-action” (J. Meier, 2022, p. 27). The basic assumption (although condensing vast research on this topic⁵¹) is that PSTs should use theory to think about their experience and come to a reasoned conclusion on its efficiency, hypothesizing about alternative actions that could have been taken instead, thus offering one way to connect theory to practice (F. Korthagen & Nuijten, 2022, p. 24). These reflective skills are certainly relevant, however, the jury is still out on whether these skills have an influence on (pre-service) teachers’ subsequent actions (Neuweg, 2022a, p. 22). While the issue of theory-practice-connection is being discussed and approaches are being taken to consolidate the gap, there are still challenges that students face when it comes to actually using theory in practice. Circling back to the introductory quote, neither the acquisition of knowledge, nor the implementation of internships seems to be having the desired impact on teacher education (Fraefel & Grossman, 2024, p. 139). Thus, new approaches seem necessary to address this issue.

In consequence, a focus on theory and knowledge can be observed for the first phase of teacher education with regard to the statutory frameworks and the conceptualizations of prominent competence models. This rather narrow perspective of competence may not necessarily suffice in preparing PSETs for the “core business of teaching” (Baumert & Kunter, 2013, p. 26), i.e., “instructional performance” (Krauss et al., 2020, p. 316). While there seems to be a development in knowledge, its application in practice is challenging (Zeichner, 2010, p. 91). Considering performance as an aspect of competence as proposed by Blömeke et al. (2015) and including performance in the first phase of teacher education might be beneficial in better preparing PSETs for practice. The next chapters will lay out how, using constructive alignment, such an approach can be taken.

4 A Performance-Oriented Perspective on Constructive Alignment in the First Phase of EFL Teacher Education

This chapter will discuss how performance can be included in the first phase of teacher education as one approach to address the challenges PSTs face in applying knowledge in the classroom. First, constructive alignment is proposed as a general framework for higher education (Biggs, 1996; Biggs & Tang, 2011). Second, a performance perspective is taken and possible approaches detailed for each of the three dimensions of constructive alignment: learning objectives, teaching and learning activities, and assessment. A special emphasis is put on

⁵¹ For a more nuanced and detailed discussion of reflection, see, for example, Klempin (2019); J. Meier (2022).

assessment of performance, as assessment is particularly influential for learning behavior.

4.1 Constructive Alignment

A framework that can be used to discuss how performance may be included in the first phase of teacher education is that of constructive alignment (CA) (Biggs, 1996; Biggs & Tang, 2011). Biggs' framework builds on the idea of "instructional alignment" (S. A. Cohen, 1987), which describes the congruence between curriculum and assessment (Biggs, 1996, p. 350). CA adds the notion of constructivism by learning activities: "The 'constructive' aspect refers to the idea that students construct meaning through relevant learning activities. That is, meaning is not something imparted or transmitted from teacher to learner, but is something learners have to create for themselves" (Biggs, 2003, p. 1). According to CA, when designing university courses, three facets should align to create an internally harmonized system, i.e., "a stable equilibrium" (Biggs, 1996, p. 350): The objectives that are to be reached, the teaching and learning activities (TLAs) that are planned and executed, and the assessment that seeks to evaluate whether the objectives have been achieved (see Figure 8).

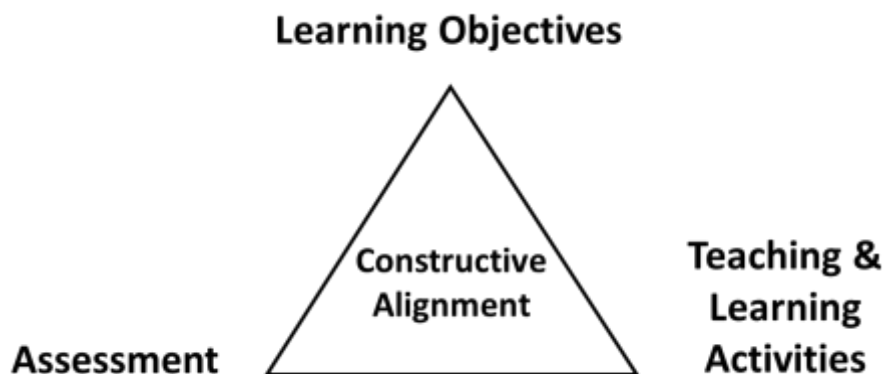


Figure 8: Constructive alignment in higher education (Biggs, 1996; figure adapted from Deibl et al., 2018, 295; Gallagher, 2017, p. 3013).

First, courses need to set the specific learning objectives. Biggs argues that in order to achieve high levels of mastery, high-level goals, e.g., the "performative aspect of understanding" (Biggs, 1996, p. 351), should be set. Biggs mainly distinguishes between *declarative* and *functioning knowledge*. Whereas *declarative knowledge* refers to "knowing about things", *functioning knowledge*, "is what professionals are concerned with; they use theory to inform their decisions on what to do in their professional context, be it solving problems, designing buildings, planning teaching or performing surgery" (Biggs & Tang, 2011, pp. 81–82). Objectives that are set for higher levels need to be more expansive than declarative knowledge because "[good] students will turn declarative into functioning

knowledge in time, but most will not if they are not required to" (Biggs, 2003, p. 3). Thus, objectives relating to functioning knowledge must be set for all students to increase the likelihood that they are achieved. However, those high-level, performative goals are rarely found at university (Biggs, 1996, p. 351). This aligns with Blömeke's criticism that the implemented descriptors by the KMK do not reach far enough, relying only on (declarative) knowledge (2006, p. 29). As has been elaborated in the previous chapter, these frameworks are hence not sufficient because their focus on theory and knowledge does not allow for an adequate preparation of PSETs at university for practice. Using the model by Blömeke et al. (2015), and referring to the arguments by Biggs, learning outcomes should also include performative aspects when developing competence holistically. Consequently, the evaluation of students' learning should (at least in part) also be operationalized in criterion-oriented objectives that consider the performative aspect (Biggs, 1996, p. 353).

Based on these criterion-oriented objectives, TLAs that demand students to perform as required should be used (Biggs, 1996, p. 353; Biggs & Tang, 2011, p. 58). Educators need to design TLAs that follow a constructive approach, i.e., "learners arrive at meaning by actively selecting, and cumulatively constructing, their own knowledge, through both individual and social activity" (Biggs, 1996, p. 348). Biggs is especially critical of lectures as they do not necessarily allow for the attainment of these higher levels due to their lack of activity (Biggs, 1996, p. 353). Certainly, teaching activities at university are less unidirectional in current teacher education, as, for example, seminars are usually included as courses that are more interactive (Hilger et al., 2015, p. 59). However, if performative aspects are included as objectives for these courses, learning activities need to enable performance. In teacher education, for instance, this includes internships as a learning activity outside of the campus as well courses that allow for planned practice of performance.

As a last point, assessment seeks to evaluate whether students have achieved the objectives through the executed TLAs. Consequently, assessment should be designed so that it reflects the objectives appropriately and actually measures what students are supposed to learn in their courses (Biggs & Tang, 2011, p. 99). The purpose of assessment, however, is also dependent on the respective perspective: While for the teacher, "it checks the quality of learning, [...] for students, it defines what is to be learned" (Biggs, 2003, p. 3). Assessment tasks should thus also address the higher levels of understanding, i.e., performance, because "if the set assessment tasks address lower cognitive level activities than those nominated in the curriculum objectives, equilibrium will be achieved at a lower level" (Biggs, 1996, p. 350). Performance assessment could thus be defined as "a type of assessment that requires students to actually perform, demonstrate,

construct, develop a product or a solution under defined conditions and standards" (Khattari & Sweet, 2013, p. 3). For the learner, the way they are assessed is relevant to their learning behavior (Biggs, 2003, p. 3). If students' performance is the focus of assessment, they are incentivized to focus on their performance in the preparation for the assessment, so the idea. This sentiment can best be summarized with the truism that "assessment drives learning" (Juan Fischer et al., 2024, p. 233; Sopka et al., 2018, p. 97).

In the following three chapters, each of the described aspects of CA will be explored in more depth regarding the performative facet of competence, describing concepts and empirical findings for the first phase of teacher education.

4.2 Learning Objectives: The Core Practice Approach in EFL Teacher Education

If performance is to be focused on in the first phase, adequate learning objectives need to be formulated, which can then be targeted. As the implemented standards for the first phase focus more on knowledge, a different approach may be considered to describe learning objectives from a performance perspective, such as the core practice approach (Grossman, 2018; Grossman & Fraefel, 2024).

4.2.1 The Core Practice Approach

In response to the criticism that initial teacher education, especially the university-based phase, is too reliant on knowledge and not adequately preparing teachers for the profession, the core practice approach has emerged, providing a different perspective on the focus of teacher education (Grossman & Fraefel, 2024, p. 1; Grossman et al., 2018, p. 3; McDonald et al., 2013, p. 378). Core practices can be defined as practices "that are fundamental to teaching and grounded in disciplinary and pedagogical goals that teachers enact to support student learning" (Grossman & Fraefel, 2024, p. 2). They can be characterized as

- practices that occur with high frequency in teaching;
- practices that teachers can enact in classrooms across different curricula or instructional approaches;
- practices that allow teachers to learn more about students and about teaching;
- practices that preserve the integrity and complexity of teaching; and
- practices that are research-based and have the potential to improve student achievement. (Grossman, Hammerness, & McDonald, 2009, p. 277; Grossman et al., 2018, pp. 4–5)

The individual core practice may vary in grain size, as some conceptualizations focus on larger actions, like lesson planning, and other focus on smaller

techniques like providing oral corrective feedback (Chan, 2023; Grossman et al., 2018, p. 4; Pang, 2019, p. 269). Defining and establishing discipline-related core practices also allows for the establishment of a common language for teacher educators to discuss and describe the demands for teachers (Levine et al., 2023, p. 80; McDonald et al., 2013, p. 381). What is more, the core practice approach “provides teacher educators with the opportunity to address teaching as a complex task, while also enabling them to focus on key components with novice teachers” (Grossman, Hammerness, & McDonald, 2009, p. 277). At the University of Michigan, the group of TeachingWorks has developed a set of 19 “high-leverage practices” (Grossman et al., 2018, p. 5) or “high-level teaching practices (HLTPs)” (Glisan & Donato, 2017, p. 1)⁵². They include, for instance, “[diagnosing] particular common patterns of student thinking and development in a subject-matter domain”, “interpreting the results of student work, including routine assignments, quizzes, tests, projects, and standardized assessments” or “[providing] oral and written feedback to students” (Grossman, 2018, pp. 164–168; TeachingWorks, n.d.).

Similar to competences, core practices are to be acquired in a tiered process that gradually increases the performative component. In fact, core practices research is less focused on conceptual-theoretical discussions, but rather pragmatic, more strongly emphasizing the process behind mastering these practices (Grossman, 2018; McDonald et al., 2013, p. 379). In a cross-professional analysis, Grossman et al. identify three distinct approaches that are taken in different professions to prepare for practice: “representations, decompositions, and approximations of practice” (2009, p. 2055). In *representations of practice*, novices are presented with examples of the profession, either by observations, videos or recounts (Grossman, Compton, et al., 2009, pp. 2064–2068). In *decompositions of practice*, the specific components are identified and singled out individually “for the purposes of teaching and learning” (Grossman, Compton, et al., 2009, p. 2069). This allows, for example, focusing on specific elements of the profession, like designing a task. These components may also be described as core practices if they fulfill the previously described characteristics. *Approximations of practice* give novices the opportunity to perform the practice in a complexity-reduced setting. For instance, “[students] may be asked to simulate certain aspects of practice through activities such as role-plays” (Grossman, Compton, et al., 2009, p. 2076). In these less complex situations, students may test out different ways to deal with challenges, receive feedback and thus modify and refine their practice. Thus, approximations of practice allow for errors and deliberate

⁵² These are also abbreviated as HLPs. HLPs and HLTPs are used interchangeably with core practices. For consistency, the term core practice is used in this thesis.

practice to master practices (Grossman, Compton, et al., 2009, p. 2077). McDonald et al. (2013) have proposed a learning circle for teacher education in which these elements can be located. In a first step, pre-service teachers might be introduced to representations of practice through videos or transcripts, before analyzing them in more depth in a decomposition of practice (McDonald et al., 2013, p. 382). In a second step, pre-service teachers might collaborate on planning an activity and try out their planned practice in a microteaching scenario as an approximation of practice before actually engaging in practice with real students (McDonald et al., 2013, p. 383). This could also function as an opportunity for reflection in an *investigation of practice* (Grossman, Hammerness, & McDonald, 2009, p. 274; McDonald et al., 2013, p. 383). This practice-based approach does, however, not imply that theory should be excluded in teacher-education programs (Forzani, 2014, p. 367). Rather, it allows for a productive connection of knowledge to professional performative demands, with knowledge functioning as a potential resource (Fraefel, 2022, p. 17). Also, practice should not be imitated and copied as proposed in the craft model (Glisan & Donato, 2017, p. 7; Müller-Hartmann & Schocker-von Ditfurth, 2014, p. 12). Rather, own practice should be developed by using the theory in performance-oriented practice (Summer & Plötz, 2024, p. 161). Therefore, this approach also has to be distinguished from the implementation of practice-elements like internships (Forzani, 2014; Summer & Plötz, 2024, p. 160). Criticism of the core practice approach relates to its negligence of equity and justice (Philip et al., 2019, p. 252), or the deliberateness with which core practices are defined and used in teacher education (Chan, 2023).

Arguably, mastering core practices could, thus, be understood as the habitualized realization of competence, i.e., performance in the sense of Blömeke et al. (2015). Core practices describe the relevant actions that teachers need to be able to execute, providing a performative set of learning objectives. The underlying approach describes a methodology through which, with a gradual increase of complexity regarding the performance of practice, students can accomplish mastery of these core practices.

4.2.2 Core Practices in the EFL Context

Although core practices originated in a more general approach to teacher education, they can and have been conceptualized for specific contexts, like the (foreign) language classroom (Chan, 2023; Grossman, 2018, pp. 174–183; Martel, 2021, p. 2; Summer & Plötz, 2024, p. 161). The research into EFL specific core practices is, while growing continuously, still rather scarce, especially in Germany (Pang, 2019, p. 259; Summer & Plötz, 2024, p. 166). Glisan and Donato (2017,

2021) propose a set of ten core practices for the foreign language classroom, as presented in Figure 9:

- 1) Facilitating target language comprehensibility
- 2) Building a classroom discourse community
- 3) Guiding learners to interpret and discuss authentic texts
- 4) Focusing on form in a dialogic context through PACE (Presentation, Attention, Co-construction, Extension)
- 5) Focusing on cultural products, practices, and perspectives in a dialogic context
- 6) Providing oral corrective feedback to improve learner performance
- 7) Establishing meaningful and purposeful context for language instruction
- 8) Planning for instruction using an iterative process for backward design
- 9) Engaging learners in purposeful written communication
- 10) Developing contextualized performance assessments

Figure 9: *Core Practices in the Foreign Language Classroom (Glisan & Donato, 2017, 2021 as cited in Summer & Plötz, 2024, p. 164.*

Each of these core practices is decomposed and defined more specifically in rubrics so that they can be more specifically learned and taught (Summer & Plötz, 2024, p. 164). For example, the practice of “Providing oral corrective feedback to improve learner performance” includes rubrics such as “Use of Implicit vs. Explicit [Corrective Feedback]” (Glisan & Donato, 2017, p. 160). The authors argue that their comprehensive list of core practice is evidence-oriented (Glisan & Donato, 2017, p. 10), while Barahona, comments that “the selection process of how and who identified the set of recommended practices is not defined” (2020, p. 3). In an analysis of foreign language classroom videos, Zhai applied the rubrics of the described core practices 1) and 2) and found that these occurred frequently in these videotaped situations (2019). Pang proposes a set of five subject practices, for the EFL classroom as seen in Figure 10 (2019, p. 265):

1. offering comprehensible input as exposure to models of language use;
2. explicit teaching of key language elements in context to develop knowledge about language;
3. supporting interactive task engagement in target communication areas with opportunities for re-hearsing language forms, specific skills, and communicative routines;
4. stretching learner output with a focus on form–function relationships in mastering the basic genres of communication; and
5. giving informing feedback on use of language with reference to the target language and learner interlanguage features.

Figure 10: *Subject practices for the EFL classroom according to Pang (2019, p. 265).*

While similarities between Pang as well as Glisan and Donati are evident, for example in the first of each of their core practices, there are also differences. For instance, Pang situates these practices explicitly within the framework of task-

based language learning (2019, p. 266) and “claims the primacy of explicit teaching practices for teaching grammar” (Barahona, 2020, p. 4). What is more, the described sets consider core practices of different grain sizes, complicating comparisons between, but also within, proposals. Also, these core practices are developed in specific national contexts and are thus not reflective of the requirements of Germany’s EFL classrooms, missing, for example, aspects of the 2023 *BISTAs* for the foreign languages⁵³ such as “foreign language-specific digital competence” (KMK, 2023a, p. 9; Summer & Plötz, 2024, p. 166). Such a culturally-specific approach has, for example, been taken by Barahona, who explored the perceived core practices for EFL teachers in the Chilean context in a Delphi study with 46 EFL teacher educators (2020). In that study, the core practices of “*Facilitating target language comprehensibility*, *Building a discourse community* and *Focusing on cultural products (connecting with other content areas)*” (2020, p. 10; emphases in the original), which are proposed by Glisan and Donato (2017), are agreed upon for the Chilean context by an expert panel.

While not binding for teacher education in Germany, these sets of core practices provide an indication of the activities that are considered essential for the EFL context. They explicitly claim to delineate the requirements for EFL teachers based on practice and described using practice. Although the KMK descriptors are oriented towards the necessary requirements for teaching as well, the description using core practices allows for a performance perspective for teacher education programs. In the following chapter, (role-play-based) simulations, as one specific option for TLAs that aim at the development of core practices, are focused on.

4.3 Teaching and Learning Activities: Simulation-based Learning

One promising approach that could be implemented to acquire the mastery of core practices, focuses on performance at the university level, and is rooted in constructivist learning theory is that of simulation-based learning (SBL) (Levin, Frei-Landau, & Goldberg, 2023, p. 2). In general, “SBL offers learning with approximation of practice, allows limitations of learning in real-life situations to be overcome, and can be an effective approach to develop complex skills” (Chernikova, Heitzmann, Stadler, et al., 2020, p. 502). Simulations can thus be viewed as a method within the core practice approach, specifically approximations of practice (Grossman, Compton, et al., 2009, p. 2077). SBL can be related to experiential or authentic learning (Levin, Frei-Landau, Flavian, & Miller, 2023, p. 2). These approaches allow learners to use their knowledge in a performative way,

⁵³ For legibility, the abbreviation *BISTAs* will also be used when referring to the standards specifically for foreign languages.

resembling their future professional actions (Herrington et al., 2014, p. 401; Levin, Frei-Landau, Flavian, & Miller, 2023, p. 2). For decades, simulations have been a cornerstone in the education and training of many professions, such as aviation (Myers et al., 2018, p. 2) or medicine (St. Pierre, 2018, p. 2), to name a few. Mostly, simulations have been used in contexts “where real-world skills practice is logistically challenging, dangerous, or costly” (Kaufman & Ireland, 2019, p. 3). In educational or training contexts, learners engage physically with simulations “for the purpose of teaching or assessment” (Cook et al., 2013, p. 874; see also Kim et al., 2016, p. 1). In simulations, participants are able to practice performance “without risking harm” (Kaufman & Ireland, 2019, p. 3), repeat the situation, and make mistakes from which they can learn by receiving feedback, before starting surgery on actual patients or flying real planes (Jossberger et al., 2022, p. 1).

In the context of (teacher) education, simulations have not really been much discussed until recently. Arguably, the potential risk of PSTs negatively impacting real learners when practicing their performance is not considered sufficiently high enough to warrant much use of simulations. Possibly, the effect of teacher performance on the learner is often not as immediately visible as in other professions, probably contributing to the lack of implementation of simulations in which PSTs can practice. However, teachers’ wrong diagnostic decisions, as one example, may have severe negative impacts on learners, especially if these decisions influence the decision on academic tracks in the German school system. Nevertheless, as findings from other professions have indicated that simulations can have a positive impact on the learning process, research in that field has also grown steadily in teacher education (Chernikova, Heitzmann, Fink, et al., 2020, p. 1; Kadel et al., 2023, p. 20; Kaufman & Ireland, 2019, p. 2; Levin, Frei-Landau, Flavian, & Miller, 2023, p. 2). In Germany, the Standing Scientific Commission on Education Policy (abbreviated as *SWK* for the German name *Ständige Wissenschaftliche Kommission*), an advisory board for the KMK, has recently also emphasized the value of simulations for the acquisition of competence in teacher education (Ständige Wissenschaftliche Kommission [SWK], 2023, p. 74). While Kaufman and Ireland argue that “[simulations] are becoming more widely available for pre-service teacher education” (2019, p. 4), others state that simulations are still in its “infancy with few large scale, high quality studies and a lack of conceptual clarity” (Ade-Ojo et al., 2022, p. 872). This chapter then seeks to provide a brief overview on the concept and types of simulation, especially in teacher education, and outline current findings that explore the effects and perceptions of SBL.

4.3.1 Characteristics of Simulations

Simulations are generally defined as “a simplified but accurate, valid, and dynamic model of reality” (Kaufman & Ireland, 2019, p. 1). For teacher education specifically, they might “offer dynamic, rule-based and often simplified imitations of classroom events” (Theelen et al., 2019, p. 16), but may also depict situations that are related to the profession, such as parent-teacher talks (Dotger et al., 2010). Thus, simulations are both flexible and structured in their conceptualization (Levin, Frei-Landau, Flavian, & Miller, 2023, p. 11). The notions of simplification, authenticity, flexibility, and structure can be regarded as critical aspects of simulations and will be elaborated on further (Dieker et al., 2014, p. 22; Grotegut, 2023, p. 17).

Simplifications can best be described using the notion of “approximation of practice” (Grossman, Compton, et al., 2009, p. 2076), as described in Chapter 4.2.1. The simulation should focus on specific aspects of the practice and not try to include a plethora of different components. Simplification does not, however, necessarily lead to situations that are easier to manage, as especially in practice contexts, simulations might emphasize a specific practice more strongly, while it might be more muted in real-life (Levin, Frei-Landau, Flavian, & Miller, 2023, p. 14). Similarly, such opportunities for learning may also be more elaborate than in real-life situations, allowing for an in-depth exploration of a scenario (Grossman, Compton, et al., 2009, p. 2077). For instance, a feedback conversation in school might not last more than three minutes, while a simulation may be extended to provide a more extensive opportunity to practice and to show one’s competence (see also Grotegut, 2023, p. 17).

Authenticity, likewise, is concerned with the replication of reality. The authenticity of a simulation might be described as the degree of resemblance between the simulation and the actual demands of the task (Chernikova et al., 2024, p. 16; Gulikers et al., 2004, p. 69; Gulikers, Kester, et al., 2008, p. 173). Importantly, as the learners are active participants in these simulations and construct meaning within the activity, it is *their perception* that affects the potential learning achievement (Stürmer et al., 2024, p. 3280). Thus, authenticity can be distinguished from fidelity, which

in the context of simulation-based learning is considered as a close as is possible reproduction of an object reality whereas authenticity may be considered as a subjective interpretation/ response to a constructed situation in which the student interacts with context, other students, facilitators and technology with varying degrees of fidelity. (Bland et al., 2014, p. 1113)

While the design of simulations has long been evaluated with regard to fidelity, Hamstra et al. argue for a change:

The field of simulation should shift emphasis away from structural properties of the simulator (i.e., physical resemblance) to functional properties of the entire simulation context that align with learning objectives (i.e., functional task alignment). There is now plenty of evidence that physical resemblance can be reduced with minimal or no loss of educational effectiveness, provided there is appropriate correspondence between functional aspects of the simulator and the applied context. (2014, p. 390)

Thus, fidelity is not a mean, in and of itself, and can be reduced, if the context allows for it. Consequently, in order to create an authentic simulation, only aspects of fidelity that are relevant for the alignment between learning objectives and simulation have to be considered for development. Authenticity can then be constructed by the individual through different aspects of fidelity: Physical elements, i.e., the setting or objects in the room, conceptual elements, i.e., the task or scenario, or experiential elements, i.e., experiencing the situation of the simulation (Levin, Frei-Landau, Flavian, & Miller, 2023, p. 9). One aspect that is consistently argued to contribute to an authentic perception of simulation is the use of actors taking on the role of communicative partners (Levin, Frei-Landau, Flavian, & Miller, 2023, p. 3). For instance, a simulation might take place in a room that includes several aspects of a regular classroom, but has adults acting as ninth graders. Such a simulation contains aspects of physical fidelity that contribute to authenticity, but also others that potentially impede an authentic perception for the participant.

Simulations are flexible and dynamic, i.e., they are adaptive to the participants' actions, but also structured, e.g., the beginning of the simulation is usually identical and the adaptive reactions are designed systematically (Levin, Frei-Landau, Flavian, & Miller, 2023, p. 12; Sauv   et al., 2007, p. 251; St  rmer et al., 2024, p. 3280). The participants have a certain extent of control in the simulation, arguably more so than in real-life, where other factors might influence the situation (Sauv   et al., 2007, p. 251). Based on participants' behavior, the simulation may change in a certain direction, according to simple *if-then* formulations. For instance, if participants of a classroom simulation ask certain questions, then learners in that simulation will provide some sort of reaction, e.g., in the form of a verbal answer, which might steer the subsequent course of the simulation. The progress of a simulation is not, however, accidental. Usually, simulations are initiated in an identical manner. Then, possible kinds of behavior from the participant are anticipated and reactions planned, so that the simulations might proceed in a systematic and consistent manner. In complex simulations, these *if-then* formulations may branch out to complex systems, resembling decision trees (Gr  newald & Levin, 2022, p. 6; Kaufman & Ireland, 2019, p. 5; Siker, 2023, p. 340). In role-play based simulations in the educational contest, one

structuring element in the dynamic scenario could be “verbal triggers” (VTs) (Dotger et al., 2010, p. 133). VTs can be described as “specific questions, information, and/or comments” (Dotger et al., 2010, p. 133) that the counterpart, i.e., the actor, utters to elicit certain reactions. They are used flexibly in the simulation, but also allow for standardization without the complexities of decision trees.

4.3.2 Typology of Simulations

There are different types of simulations with which one could engage. On a basic level, simulations can be described on a spectrum of (A) *Technology*, i.e., digital or analogue, (B) *Human Interaction*, i.e., the extent to which the participant is required to interact with a human communication partner, and (C) *Fidelity*, i.e., the extent to which the simulation reflects a real life scenario (Ade-Ojo et al., 2022, p. 862; Chernikova, Heitzmann, Stadler, et al., 2020, p. 503; Nickl, 2023, p. 7; Sauvé et al., 2007; Stürmer et al., 2024, p. 3281)⁵⁴. In the following, these spectra will be discussed and exemplified using selected simulations, preferably from the realm of teacher education. Of course, the selected simulations could be placed on and used to discuss each spectrum, the discussion of individual simulations is thus only exemplary and not exhaustive.

Technology

The spectrum of technology reaches from digital to analogue simulations. Digital simulations are usually differentiated from games as they do not have a competitive aspect, while in games a specific aim has to be achieved, following predetermined rules (Grotegut, 2023, p. 17; Kaufman & Ireland, 2019, p. 3; Sauvé et al., 2007, p. 253; Wouters et al., 2013, p. 250). Other terms that are often synonymously used with digital simulations are, e.g., computer-based simulations (Nickl, 2023, p. 1; Theelen et al., 2019, p. 16) or virtual simulation (Ade-Ojo et al., 2022, p. 863; Angelini, 2023, p. 45; Dieker et al., 2014, p. 21; Stevens & Kincaid, 2015, p. 41). Digital simulations can take various forms. In video-based simulations, for instance, the participant watches previously filmed, usually scripted, video clips and reacts to them (Codreanu et al., 2022, p. 40; Grotegut, 2023; Kulgemeyer et al., 2021). One example of such a simulation was developed in the project *Visit-Math* as part of the larger DFG-funded project COSIMA⁵⁵ (Codreanu et al., 2022). In that simulation, participants’ diagnostic skills are to

⁵⁴ For a discussion specifically on simulated diagnostic simulations, see Machts et al. (2023).

⁵⁵ The self-described goal of the COSIMA group is “to analyse how simulation-based learning environments in the university context can be designed and implemented in order to facilitate the acquisition of diagnostic competences, in particular in early and intermediate stages of competence acquisition” (COSIMA Research Unit, n.d.).

be fostered as they take on the role of an intern who has to “assess the students’ mathematical argumentation skills so that [the teacher] can choose tasks for individual learning support in a subsequent class based on the participants’ observations” (Codreanu et al., 2022, p. 40). The simulation consists of four steps: First participants familiarize themselves with the context, i.e., the *diagnostic situation*, by receiving additional information, e.g., lesson plans (Codreanu et al., 2022, p. 40). Second, the participants are presented with the *diagnostic task* as described above (Codreanu et al., 2022, pp. 40–41). In the *diagnostic process*, participants watch videos of simulated learner-teacher interactions in several cycles (i.e., videos of the same learner at different points) and take notes (Codreanu et al., 2022, pp. 42–43). These videos are scripted situations, which last about one minute each and include a trained teacher and actual eighth-grade learners (Codreanu et al., 2022, pp. 37–39). In the last step, participants have to formulate a diagnosis and assess various aspects of the learner, e.g., mathematical content knowledge (Codreanu et al., 2022, p. 43). Similarly, in the digital GEPoS⁵⁶ simulation, participants’ diagnostic competence is also focused upon as they take on the role of a sixth-grade teacher (Grotegut & Klingsieck, 2023, p. 59). Conversely, GEPoS is avatar-based as these simulated situations are designed as 3-D animated video sequences (Grotegut, 2023, p. 26). Participants are asked to collect, systematize and evaluate data, which is presented through various artefacts, e.g., test results, notes, documents, as well as observations of the animated simulated virtual conversations with parents and learners and, ultimately, a simulated lesson in which learners show specific features with regard to learning challenges or unwanted behavior (Grotegut, 2023, pp. 25–26; Grotegut & Klingsieck, 2023, p. 59).

Analogue simulations, i.e., those that do not include digital elements or technology in a primary manner, (Chernikova, Heitzmann, Stadler, et al., 2020, p. 503; Kadel et al., 2023, p. 29) have the participant interact directly with physical objects or human beings. They are sometimes also referred to as physical simulations (Ade-Ojo et al., 2022, p. 862) or live simulations (Kron et al., 2021, p. 7). One common approach to analogue simulations is the role-play. While role-play-based simulations (RobS⁵⁷) might also be digitally mediated (Theelen et al., 2019, p. 17), e.g., through a video conference system (Fecke & Müller, p. 224; Stürmer et al., 2024, p. 3288), they are usually conducted in an analogue manner

⁵⁶ GEPoS is the abbreviation of *Game- und e-learningbasierte, problemorientierte und selbst-gesteuerte Lernumgebung* (Game- and e-learningbased problemoriented and self-regulated learning environment; translated by TJ).

⁵⁷ The term *Role-play-based Simulations* is always abbreviated as RobS in this thesis for consistency, not only when referring to the test described in Chapter 8.

(Ade-Ojo et al., 2022, p. 869) and could be described as “in vivo” (Braun, Schwabe, & Klein, 2018, p. 6) simulations. In RobS in teacher education, the participant usually takes on the role of the teacher, engaging with one or multiple learners (e.g., Braun, Schwabe, & Klein, 2018; Kulgemeyer & Tomczyszyn, 2015; Marczyński et al., 2022; Telgmann & Müller, 2023), parents (e.g., Dotger et al., 2010; Gartmeier et al., 2015; Gerich & Schmitz, 2016) or, potentially, other stakeholders like colleagues (Braun, Schwabe, & Klein, 2018, p. 6). In teacher education, such analogue role-plays are sometimes referred to as microteaching, which can be defined as a “technique in which a short lesson is taught to an audience, video-recorded, reviewed, and analyzed” (Cavanaugh, 2022, p. 23). Microteaching is often part of courses in which fellow students take on the role of the class without being specifically instructed, a difference from to simulations in general and the RobS listed above, which usually has people trained for a specific role to ensure standardization (Cavanaugh, 2022, p. 32).

Human interaction

The spectrum of human interaction ranges from actual face-to-face conversations with humans to no communication at all. Analogue RobS can be found on the one end of the spectrum as they demand actual human interaction. There are, however, also examples of mixed-reality simulations which are designed as digitally mediated interactions (Ade-Ojo et al., 2022, p. 863). In the *TLE TeachLive™* simulation, for instance, the participant interacts with a group of teacher, parent, or learner avatars “that were created using 3-D modeling and computer animation techniques. The resulting avatars are controlled by artificial intelligence and a human operator who embodies the avatars” (Dieker et al., 2015, p. 12) and which are claimed to “look, talk, and interact like typical humans” (Dieker et al., 2015, p. 12). This simulation can also be classified as a “digital puppetry-based virtual [environment]” (Dieker et al., 2014, p. 28) as a human is responsible for the reaction of the avatars. Avatar behavior is based on a “decision tree” (Siker, 2023, p. 340) which prescribes the branches of behavior that depend on the participants’ actions. This simulation is not necessarily focused on a specific skill and can be used for a variety of scenarios (Dieker et al., 2015, p. 13). The simulation developed by Fecke and Müller focusses on effective communication in role-plays (p. 216). In their digital simulation, which is based on the analogue role-play scenarios by Braun, Schwabe, and Klein (2018), both the participant and the actor use avatars and meet online in a virtual room to act out a situation, which is relevant for future teachers (Fecke & Müller, p. 222). Within such a virtual learning environment, AI-based avatars may also serve as counterparts, allowing for interaction which is not human-based (Afzal et al., 2024; Fecke et al., 2023; Fink et al., 2024).

Simulations that are less-interactive are those that do not have a direct communication partner. The *Cook School District Simulation* is web- and algorithm-based, and allows PSTs “to explore connections between their actions as teachers and the learning and engagement of simulated students” (Girod & Girod, 2006, p. 481). Participants can navigate the websites, make instructive decisions for a simulated class of learners and see potential outcomes (Girod & Girod, 2006, p. 486). Similarly, the simulation *SCR^{Bio}* provides participants with simulated teacher-learner interactions in the biology classroom, targeting the diagnosis and evaluation of learner answers that are provided to the participants in written form (Julian Fischer et al., 2022, p. 1770; Julian Fischer et al., 2021, p. 221). The simulation *DiaSim* focuses on pre-service foreign language teachers, specifically French and Spanish (Grünewald & Levin, 2022, p. 4). It was piloted as a paper & pen simulation and subsequently turned into a digital simulation (Grünewald & Levin, 2022, p. 5). The simulation includes aspects of games, in that there are several levels and a gamemaster (Grünewald & Levin, 2022, p. 4). In the simulation, participants are provided with an initial video of a lesson sequence and have to take on the role of the teacher afterwards, using further text and audio material to decide on subsequent actions (Grünewald & Levin, 2022, p. 4). Based on a decision tree, the gamemaster unlocks a different path for the participants and, in the end, participants have to decide on options to support learners optimally (Grünewald & Levin, 2022, p. 4).

Fidelity

The spectrum of fidelity is concerned with the extent to which the simulation replicates actual reality. Naturally, RobS might have a relatively high degree of fidelity just due to the notion of human interaction. In teacher education, one aspect of fidelity might refer to the number of interactors. Classroom-simulations usually have multiple simulated learners, either digitally, as described by Dieker et al. (2015), or in an analogue format (Telgmann & Müller, 2023, p. 5; Riegger & Rotthoff, 2022). The *TeachEye-ClasS* simulation, for example, while still being reduced in complexity and thus approximating practice, has a duration of 20 minutes and includes seven trained actors as learners who display, for instance, disruptive behavior (Telgmann & Müller, 2023, p. 6). Specific actions by the actors are planned for different time frames, allowing for coordination and standardization of the simulation (Telgmann & Müller, 2023, p. 6). Other RobS might further reduce complexity by focusing on single learners, though simultaneously decreasing fidelity. For instance, the simulation developed by Kulgemeyer and Tomczyszyn focuses on the explaining competence of physics teachers using one-to-one situations with a participant and an actor (2015, p. 117). High-fidelity simulations might facilitate participants’ “suspension of disbelief”,

i.e., their feeling of immersion in the situation, through attention to detail, while low-fidelity simulations make concessions in their approximation of practice (Kaufman & Ireland, 2019, p. 11; Kim et al., 2016, p. 2; Levin, Frei-Landau, Flavian, & Miller, 2023, p. 14). For instance, in microteaching simulations used by Holstein et al. (2022), age-appropriate actors are used, while in other simulations, other university-students take on the role of younger learners (Kulgemeyer & Tomczyszyn, 2015), displaying the potential variability of fidelity in that regard. Other aspects of fidelity refer to the immersiveness, especially from a spatial perspective. Simulations that are designed by the means of augmented reality (AR) or virtual reality (VR) allow for an immersive experience in the digital space, potentially increasing fidelity (Ade-Ojo et al., 2022, p. 863). After exploring the different types of simulations in the context of teacher education, the following chapter will present selected empirical research in the area of simulations and SBL.

4.3.3 Perception of Simulations and their Effectiveness for Learning

SBL is situated within the learning paradigm of constructivism (Levin, Frei-Landau, & Goldberg, 2023, p. 2). While in medical education the effectiveness of SBL has been explored to a larger extent, research in teacher education is rather scarce (Chernikova, Heitzmann, Stadler, et al., 2020, p. 512). Nonetheless, the existing research that has focused on SBL in teacher education has explored a variety of adjacent topics, especially concerning digital simulations⁵⁸ and simulations focusing on diagnostic capabilities (Chernikova, Heitzmann, Fink, et al., 2020). In the following, selected findings will be presented that consider SBL mostly in the context of teacher education, focusing on its effectiveness and perception.

In more general findings, the meta-analysis of Chernikova, Heitzmann, Stadler, et al. ($N = 145$ studies, $n = 7$ for teacher education) concludes that SBL has an overall positive effect on fostering complex skills with $g = .85$, $SE = .08$ (2020, p. 513). Literature reviews by Ade-Ojo et al. (2022) and Theelen et al. (2019) corroborate this notion. While overall results are quite heterogenous, the skills of technical performance, diagnosis, and problem solving seem to benefit most from SBL (Chernikova, Heitzmann, Stadler, et al., 2020, p. 513). Role-plays have an estimated effect of $g = 0.63$, $SE = 1.12$, based on $N = 26$ studies (Chernikova, Heitzmann, Stadler, et al., 2020, p. 518). Interestingly, results for RobS are lower than those for other types of simulations, although those other studies

⁵⁸ See the following meta-analyses and literature reviews for an overview: Ade-Ojo et al. (2022); Chernikova et al. (2024); Chernikova, Heitzmann, Fink, et al. (2020); Chernikova, Heitzmann, Stadler, et al. (2020); Grotegut (2023, p. 20); Theelen et al. (2019).

were mostly from medical education and included (physical) models (Chernikova, Heitzmann, Stadler, et al., 2020, p. 514). Notably, simulations that are classified as highly authentic show better results than those that are considered less authentic (Chernikova, Heitzmann, Stadler, et al., 2020, p. 518). In a subsequent meta-study, Chernikova et al. found that while simulations with high authenticity have a significantly larger effect size than those with a low degree of authenticity, “[the] level of physical resemblance of the learning scenario did not have a significant impact on the effects of the simulation” (2024, p. 21). Moreover, while authenticity plays a notable role in the discussion of simulations (e.g., Levin, Frei-Landau, Flavian, & Miller, 2023), it “can account for only about 2 % of the variance in the effectiveness of simulations” (Chernikova et al., 2024, p. 22), calling that emphasis on authenticity in research into question.

The notion of authenticity is further complicated as its perception cannot necessarily be predicted - it “is in the eye of the beholder” (Gulikers, Bastiaens, et al., 2008, p. 401). For instance, it could be assumed that video-based simulations are perceived as less authentic than live role-play simulations, as the latter resemble reality more closely. In a study by Stürmer et al., however, the contrary is found, i.e., the video-based simulation is perceived as more authentic by the participants (2024, p. 3290). Potentially, these findings are the results of another variable, specifically the representation of the learner, as the video-based simulation includes images and audio of an actual sixth-grade learner, while the role-play was conducted with adults as actors (Stürmer et al., 2024, p. 3286). This argument is also supported by Levin, Frei-Landau, Flavian, and Miller, who emphasize the relevance of the actor for the authenticity of simulations (2023, p. 13). One potential reason stated by the authors is that “authentic design features on the physical level are more overt to novices, and therefore primarily used for the assessment of task authenticity” (Stürmer et al., 2024, p. 3293). In the same study, the familiarity with the simulation and other learning prerequisites also has a positive impact on authenticity (Stürmer et al., 2024, p. 3291). These findings emphasize that the perception of authenticity is dependent on the personal and individual experience of the participant and is constructed through several different aspects by the individual.

While the benefits for SBL are evident, there also limitations that need to be addressed. For instance, learners might feel overwhelmed due to the complexity of the task and, potentially, the digital aspect of the simulation, making a graduated introduction to the simulation favorable (Angelini et al., 2024, p. 516; Stürmer et al., 2024, p. 3293). Also, especially with high-fidelity simulations and RobS that use live actors, high financial and timely resources are necessary, which is why, especially for RobS, digital substitutes are often implemented (Kaufman & Ireland, 2019, p. 10). Lastly, Kaufman and Ireland argue “that no

rigorous research demonstrates skills transfer from the simulated classroom environment to authentic classrooms. Without this evidence base, it is understandable that teacher education administrators are unwilling to invest in this new technology” (2019, p. 11). Consequently, further research is required that seeks to explore this transfer.

4.4 Assessment: RobS in EFL Teacher Education

Assessment is the third aspect of CA and, from the perspective of the learner, arguably the most important one as learners adjust their learning process towards the specifics of assessment (Biggs, 2003, p. 3), i.e., “assessment drives learning” (Juan Fischer et al., 2024, p. 233; Sopka et al., 2018, p. 97). In alignment with core practices and SBL, this chapter details the exploration of options for the assessment of performance. First, relevant initial terminology that facilitates the discussion around assessment is laid out. Second, research on students’ perception of assessment and the impact of assessment on learning is presented. Third, simulation-based assessment formats, primarily in the form of RobS, which exist in the context of higher education are presented as a way to assess the performative facet of competence.

4.4.1 Concepts of Assessment in Higher Education

In order to adequately discuss performance-based assessment formats, some terminology needs to be introduced and contextualized. The terms of assessment and feedback are essential to this thesis, and the latter is discussed in depth in Chapters 6 and 7. A first distinction can however be made, in that “feedback refers only to the form in which the results of formative or summative assessment are communicated” (De Florio-Hansen, 2018, p. 276), while assessment is

[any] systematic method of obtaining information, used to draw inferences about characteristics of people, objects, or programs; a systematic process to measure or evaluate the characteristics or performance of individuals, programs, or other entities, for purposes of drawing inferences; sometimes used synonymously with *test*. (AERA, APA, NCME, 2014, p. 216; emphasis in the original)

Assessment is thus a process in which information is collected that is used to provide feedback. Although often used synonymously with assessment, tests are standardized instruments that are used in assessment (AERA, APA, NCME, 2014, p. 224). Assessment can, however, also occur by systematically evaluating behavior of learners in class, i.e., without a test instrument.

A frequently employed distinction in the assessment context is made with regard to its purpose. *Summative assessment* is usually conducted at the end of

a specified learning unit or program in order to obtain information of the learning achievement and is also described as assessment *of* learning (AERA, APA, NCME, 2014, p. 224; De Florio-Hansen, 2023, p. 22; W. Stadler, 2018, p. 172). A common example of summative assessment in teacher education is the written exam (Grotegut et al., in preparation). Results of summative assessment are also often the basis for providing specific credit points or certificates in higher education. *Formative assessment*, conversely, is executed during the instruction process, with the purpose of obtaining information that can be used to improve learning and teaching procedures, and is thus also called assessment *for* learning (AERA, APA, NCME, 2014, p. 219; De Florio-Hansen, 2023, p. 22; W. Stadler, 2018, p. 172). A common example in higher education are portfolios, which provide PSTs with information on how to improve in specific aspects during the learning process (Cañadas, 2023, p. 522). Similarly, feedback may be described as summative if information with regard to the achievement is provided, and formative if it includes information on how to improve (De Florio-Hansen, 2023, pp. 26–27). As conceptualizations of feedback have increasingly focused on the formative aspect to an extent that it may be viewed as the dominant perspective in literature, feedback is also understood to be formative in the context of this thesis (De Florio-Hansen, 2023, p. 27).

Assessment can be viewed from various different perspectives. One perspective that allows for a classification regarding how formats of assessment reflect the requirements of the profession is the “framework for clinical assessment” (Miller, 1990, p. 63), which was originally developed for the context of medical education but is adaptable to teacher education (Kulgemeyer et al., 2023, p. 606) (see Figure 11).

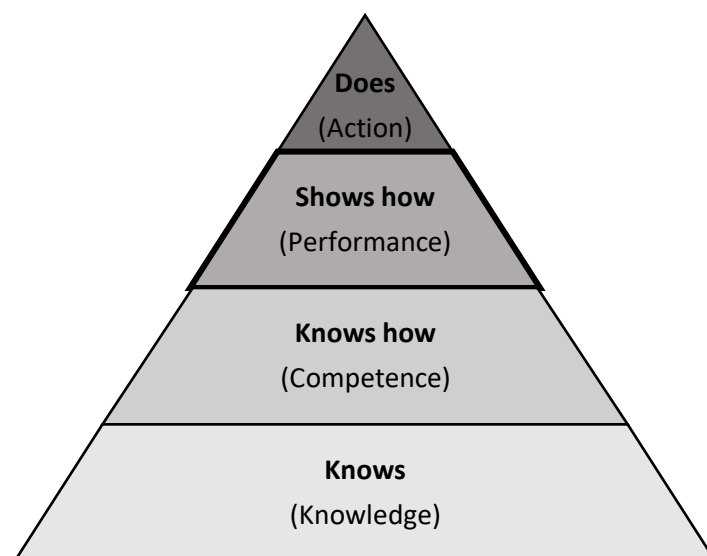


Figure 11: Framework for clinical assessment (Miller, 1990, p. 63; adapted and emphasis added by TJ).

The base level of assessment is that of *knows*, i.e., knowledge. Assessment at that level targets the declarative knowledge of students, i.e., usually in a written test (Kulgemeyer et al., 2023, p. 607). It most clearly relates to cognitive dispositions (Blömeke et al., 2015, p. 7) or professional knowledge (Baumert & Kunter, 2013, p. 29) described in the models of competence. The next level is that of *knows how*, i.e., competence. The notion of competence in Miller's framework is not to be equated with the concept of competence as discussed in Chapter 3. Rather, this might relate to situation-specific skills as described in the model by Blömeke et al. (2015) (Kulgemeyer et al., 2023, p. 607). This level primarily tests the application of knowledge, i.e., how knowledge can be used in specific situations (Wass et al., 2001, p. 946) and could also be tested in written exams. For instance, a PSET could be asked to decide which of two versions of a text is more suitable for mediation and provide reasons (Kirchhoff, 2017, p. 124). The next two levels can also be classified as "performance assessment" (Wass et al., 2001, p. 946) as they may relate to the level of performance in the model by Blömeke et al. (2015, p. 7). The level *shows how*, i.e., performance, assesses the behavior of students in simulated situations (Kulgemeyer et al., 2023, p. 607). They are thus "in vitro" (Wass et al., 2001, p. 946) as they are not conducted under the full complexity of a real-life situation. Such an assessment could be conducted as a RobS, in which core practices of the teaching profession are focused on. Lastly, *does*, i.e., action, refers to assessment formats in which the participants have to fulfill the demands of the profession in real situations, i.e., "in vivo" (Wass et al., 2001, p. 946), and are assessed based on their behavior under the full complexity of the practice (Miller, 1990, p. 63). In teacher education, graded observations of classroom situations fall into that category, as trainee teachers in the second phase have to prepare for and teach lessons with actual learners under the full complexity of the situation (Schustereder, 2020, p. 355). Of these four levels of assessment, *shows how* seems to most closely align with the outlined learning objectives described as core practices, and the TLAs which focus on mastering the core practice by SBL, an equally complexity-reduced situation in which an approximation of practice occurs in the first phase of teacher education.

Performance assessment, as assessment on that level is described, "refers to a type of assessment that requires students to actually perform, demonstrate, construct, develop a product or a solution under defined conditions and standards" (Khattri & Sweet, 2013, p. 3). If this assessment is "intended to highlight the real world nature of tasks and contexts" (Khattri & Sweet, 2013, p. 2) and includes a certain extent of fidelity, it could also be described as authentic assessment (Gulikers et al., 2004, p. 69). If assessment is to be aligned with core practices and SBL, it is necessary that the assessment format is also perceived as authentic, in that it approximates practice adequately (Gulikers et al., 2004,

p. 68). While the general notion of authenticity has been discussed in the context of simulations, Gulikers et al. developed a framework for authentic assessment which offers an additional perspective (2004). They propose five dimensions, whose relation to a criterion which stems from the real world contributes to an authentic perception of the assessment: "(a) the assessment task, (b) the physical context, (c) the social context, (d) the assessment result or form, and (e) the assessment criteria" (Gulikers et al., 2004, p. 70). The *assessment task* should be one that the students perceive to be relevant for the profession and that resembles the actual learning tasks (Gulikers et al., 2004, p. 71). This could, for instance, refer to a core practice as previously outlined. The *physical context* can be approximated by the fidelity of the situation, similar to how Levin, Frei-Landau, Flavian, and Miller describe the fidelity of physical elements in a simulation (2023, p. 9; Gulikers et al., 2004, p. 74). The *social context* refers to the extent to which the assessment resembles the interactive nature that the similar task has in real life (Gulikers et al., 2004, p. 74). For instance, if the task is to plan a lesson, this could be done alone as this is usually not done collaboratively. If the task is to provide feedback to a learner, assessment would require a person to interact with. The dimension of *assessment result or form* refers to the output or format of assessment and the extent to which it resembles real life, allows inferences about the underlying competence in a holistic manner, and provides students the opportunity to present and reason their performance (Gulikers et al., 2004, p. 75). Lastly, the dimension of *assessment criteria* refers to the extent to which the criteria are made explicit and transparent for the learning process, they relate to the actual format of the assessment, and are relevant for mastering such a situation in real-life (Gulikers et al., 2004, p. 75).

4.4.2 Assessment in Higher Education: Students' Perceptions and Impact on Learning

In this chapter, findings that describe the relationship between assessment and learning will be presented. The "effect that tests have on learning and teaching" (Hughes, 2003, p. 53) can also be described as washback⁵⁹ and is similar to the notion of "assessment drives learning" (Sopka et al., 2018, p. 97). Negative washback is often associated with a "*teaching-to-the-test*" (W. Stadler & Kremmel, 2018, p. 51; emphasis in the original) approach, in which instruction only serves to prepare for assessment and assessment is designed to test what is easy (Hughes, 2003, p. 53; W. Stadler & Kremmel, 2018, p. 51). However, if there is a constructive alignment and assessment is reflective of requirements of the job,

⁵⁹ The effect is occasionally also referred to as "*backwash*" (Hughes, 2003, p. 53; emphasis in the original).

it may be more beneficial. In the following, some selected empirical findings are presented that allow insights into that theoretical assumption.

In a literature review, Struyven et al. argue that “students’ perceptions about assessment and their approaches to learning are strongly related” (2005, p. 336). On the one hand, this relates to how learners steer their learning approaches towards assessment, on the other hand, the approaches students take to study also influences how they perceive assessment (Struyven et al., 2005, p. 325). For instance, if students perceive assessment to only focus on factual knowledge, they rather take a surface approach to learning, albeit arguably preferring such an assessment, e.g., a multiple-choice format, might be perceived as less difficult to prepare for and is thus preferred by students (Struyven et al., 2005, p. 336). Generally, the most used assessment formats in higher education are tests or written exams, as indicated by findings from Sweden (Pereira, Niklasson, & Flores, 2017, p. 160), Portugal (Pereira, Assunção Flores, & Barros, 2017, p. 450; Pereira, Niklasson, & Flores, 2017, p. 160) for a variety of study programs, and Germany for teacher education specifically (Grotegut et al., in preparation). The aspects of fairness and authenticity seem to be of particular importance in students’ perceptions (Sambell et al., 1997; Struyven et al., 2005).

For students, the perception of fairness is of “crucial importance” (Sambell et al., 1997, p. 362). Fairness relates not only to the possibility of cheating, but also to whether the students perceive the invested effort in the learning process to be appropriately reflected in the assessment format, allowing them to display their actual competence (Sambell et al., 1997, p. 363). In an interview study with $N = 13$ cases, Sambell et al. found ten conditions for assessment to be fair and have a positive effect on learning (1997, pp. 365–366). Accordingly, assessment is supposed to, for example, “[relate] to authentic tasks”, “[encourage] students to apply knowledge to realistic contexts” or “[accurately measure] skills and qualities, as opposed to an over-reliance on memory or regurgitation of facts” (Sambell et al., 1997, pp. 365–366). Comparing learner-centered methods, e.g., portfolios, and traditional methods of assessment, e.g., written exams, the former is considered fairer and more effective by students (Flores et al., 2015, p. 1531; Pereira, Assunção Flores, & Barros, 2017, p. 448; Pereira, Cadime, Brown, & Flores, 2022, p. 11). In the Portuguese context, students of educational sciences perceive assessment in general to be fairer than other groups, potentially as they report to receive more learner-centered assessment than other programs of studies, as well (Pereira, Cadime, Flores, et al., 2022, p. 81).

Students also report that learner-centered forms of assessment have a positive impact on the quality of their learning and that “assessment is most effective when it encourages students to apply knowledge in real contexts/situations” (Pereira, Assunção Flores, & Barros, 2017, p. 448), reflecting the notion of

authentic assessment (Gulikers et al., 2004). Generally, if assessment is perceived as more authentic by students, it is also more likely to lead to deeper learning processes (Gulikers, Kester, et al., 2008, p. 182). As has been outlined previously, the notion of authenticity is itself difficult, as it is a subjective notion (Gulikers, Bastiaens, et al., 2008, p. 401; see Chapter 4.3.3). Considering that authenticity relates to how assessment is perceived to reflect a professional situation, the extent to which the individual has experience in that profession might also have an impact on their perception; results are however contradictory concerning the influence of experience on different aspects of authenticity (Gulikers, Bastiaens, et al., 2008, p. 409; Gulikers, Kester, et al., 2008, p. 182). Results of a literature review suggest that authentic assessment, i.e., formats that require students to apply their knowledge in situations that are related to real-life situations (i.e., portfolios, simulations), might have a positive impact on the learning experience due to “engagement in learning and students’ satisfaction and increased efforts to achieve educational goals” (Sokhanvar et al., 2021, p. 8).

As assessment and how its perception by PSTs and other students does matter because it might influence their learning behavior, the question arises as to how students perceive the implemented assessment formats. Generally, it can be derived from the previous results that as written tests and exams dominate, the extent of authenticity is arguably not high. This is reflected in a study by Wagener et al. who interviewed $N = 170$ PSTs in Germany and found that students predominantly did not perceive assessment formats to be relating to requirements of the profession (2019, p. 217). Arguably, this perceived lack of coherence between assessment and job requirements indicates that new assessment formats might be necessary. One promising approach that would allow for authentic, performance-based and learner-centered assessment are (role-play-based) simulations (e.g., Matsumoto-Royo & Ramírez-Montoya, 2021, p. 9; Struyven et al., 2005, p. 326). However, such assessment formats are not common in university-based teacher education, as indicated by findings of a study with $N = 620$ PSTs at a German university (J. Meier et al., 2023, p. 91). If asked, those students predominantly consider role-play-based assessment formats to be relevant for their profession and authentically depict requirements of the profession (J. Meier et al., 2023, p. 92). However, PSTs are more skeptical regarding the fairness of such an assessment format, questioning the objectivity and reliability (J. Meier et al., 2023, p. 92). As most of the participants did not have experience with role-plays for assessment purposes, these conceptions of RobS need to be considered with caution. The following chapter will then provide an overview of simulation-based formats that aim at assessing performance, i.e., the level of *shows how*, in an authentic manner.

4.4.3 Simulation-based Assessment in Higher and Teacher Education

Several performance-based assessment formats that aim at the level of *shows how* have been designed for different professions. A general overview for teacher education can be found in the scoping review by Albu and Lindmeier (2023). In the following, the focus will mostly be on performance assessment that is based on simulations, so as to continue the outlined thread of CA.

Similar to SBL, simulations have been used for assessment in medical education for decades (Khan et al., 2013, 4138). Usually, simulations are part of an *Objective Structured Clinical Examination* (OSCE) (Wass et al., 2001, p. 946). OSCEs consist of several stations which seek to assess clinical skills of future doctors (Sopka et al., 2018, p. 108). These stations focus on different skills with different complexities. For instance, participants might have to interpret an X-Ray-image in one station and take the history of a simulated or standardized patient (SP) in another. Especially the use of SPs allows for a performance-oriented assessment. SPs are amateur or professional actors that have received training and instructions to portray a role in the OSCE (Obliers et al., 2018, p. 1869). In their training, SPs usually receive a script, information on the pathology and symptoms, and a detailed patient biography to depict their roles (Obliers et al., 2018, p. 1881). These roles might be less complex, e.g., simply requiring the student to measure blood pressure, or more complex, if, for instance, specific symptoms for a disease have to be replicated (Obliers et al., 2018, p. 1884). This last aspect could also include make-up or prosthetics to enhance realism, i.e., fidelity, which is essential for an accurate depiction and diagnosis (Gormley et al., 2012, p. 384). Other simulations with SPs focus more on the communicative aspect of the profession. These might include situations of medical history taking, i.e., asking patients for relevant information to come a correct diagnosis (Bajis et al., 2019), or breaking bad news to patients, i.e., informing them about a terminal disease of a relative (Polivka et al., 2024). Participants are assessed in each station of the OSCE based on checklists that often include a combination of dichotomous behavioral items and global rating scales (Khan et al., 2013, 1441; Setyonugroho et al., 2015, p. 1483; Sopka et al., 2018, p. 108). In the case of history taking, for instance, the scoring rubrics might consist of a list of potential pieces of information which have to be asked for “satisfactorily” (Khan et al., 2013, 1441), e.g., drug and family history. As a similar concept, an *Objective Structured Teaching Examination* (OSTE) can be used to assess the capabilities of medical educators (Sturpe et al., 2013).

Taking the template of medical education, role-plays are developed in the *KomPrü* project⁶⁰ focusing on the assessment of communication skills (Braun, 2021; Braun et al., 2016; Braun, Schwabe, & Klein, 2018). The authors conceived ten different role-play scenarios based on Habermas' different types of communication, a literature review and expert interviews (Walz et al., 2024, p. 5). Participants have five minutes to prepare and the interactions have a duration of about ten minutes (Braun, 2021, 7). These role-plays focus on relevant job-related situations such as "[convincing] my co-worker to arrive on time" or "[explaining] to someone the reason for a poor performance" (Braun, 2021, p. 5). For each role-play, a distinct observation sheet is created for scoring purposes (Braun, 2021, 4). They include several items, which are to be rated on a four-point rating scale (Braun, Athanassiou, et al., 2018, p. 45). In studies by Walz (2023)(2023) and Walz et al. (2024), these role-plays have also been used for assessment purposes in the context of teacher education.

Generally, however, "assessment with simulations appears less common" (Kaufman & Ireland, 2019, p. 4) in teacher education, especially in a context with simulated or standardized communication partners. Nonetheless, there are some projects which have sought to use simulations as a means of evaluation. One simulation-based approach to assessment includes parent-teacher conversations (Gartmeier et al., 2015; Gartmeier et al., 2011; Gerich & Schmitz, 2016; Walker & Dotger, 2012). Walker and Dotger developed an assessment instrument for simulated parent-teacher conferences that were initially developed by Dotger et al. (2010; 2008) (2012, p. 65). Before the simulation, participants receive documents describing the case to prepare for the conversation while in the simulations with simulated parents, standardization is achieved by the means of VTs, i.e., "specific statements, questions, or concerns" (Walker & Dotger, 2012, p. 65). The scoring instrument for the simulation consists of seven categories, which are to be rated on a five-point scale (Walker & Dotger, 2012, p. 73). The agreement of coders concerning the identification of the seven categories, not the actual scoring, is reported as 87 % (Walker & Dotger, 2012, p. 66). In the *ProfKom* project, a RobS is created using the SP model with actors portraying standardized parents (Gartmeier et al., 2011, p. 420). For scoring, a five-point scale is used which describes different competence levels (Gartmeier et al., 2015, p. 452; Gartmeier et al., 2011, p. 421; Wiesbeck, 2015). In the main study, an ICC-value of .54 is achieved (Gartmeier et al., 2015, p. 452). Gerich and Schmitz also developed a test to assess counseling talk simulations with a standardized parent that focuses on the respective child's learning process (2016,

⁶⁰ *KomPrü* is part of the larger research project *KoKoHs* (Pant et al., 2016).

p. 288). Similar to the previous simulations, participants receive material to prepare for the simulation and subsequently perform in the simulation with a standardized parent (Gerich & Schmitz, 2016, p. 290). Their rating system has a score range from zero to two points for each of the behavior-based variables, reaching acceptable inter-rater reliability scores for the variables ($ICC = .76$ to 1.00 , two raters) (Gerich & Schmitz, 2016, p. 292).

For a subject-specific context, most performance-assessment formats focus on STEM⁶¹ subjects (Albu & Lindmeier, 2023, pp. 762–763). In the context of physics teacher education, Kulgemeyer and Tomczyszyn have developed a test scenario for the measurement of explaining competence (2015, p. 111; see also Kulgemeyer & Riese, 2018). In this simulated situation, the participants have to explain a physical phenomenon to one SL who has received training to act in an identical manner (Kulgemeyer & Tomczyszyn, 2015, pp. 118–119). In order to create comparability, specific questions have been designed which the actors are supposed to use to vary the degree of abstractedness of the explanations (Kulgemeyer & Tomczyszyn, 2015, p. 118). Participants have ten minutes to prepare for the simulation and have access to material, i.e., figures and graphics that they can use in the simulation (Kulgemeyer & Tomczyszyn, 2015, p. 118). For scoring, dichotomous items are used (Kulgemeyer & Tomczyszyn, 2015, p. 121) and an inter-rater agreement of 73 % - 97 % is reached for individual items (Kulgemeyer & Riese, 2018, p. 1404). Also in physics teacher education, Schröder et al. developed a standardized performance-based instrument to measure lesson planning skills (2020). While not using SLs in their test, the instrument can be situated in the realm of *shows how* as creating lesson plans for a fictional class is an *in vitro*, performative action that is part of the teaching profession (Schröder et al., 2020, p. 107). Scoring is also based on dichotomous items as the inter-rater agreement is not sufficient with more differentiated items (Schröder et al., 2020, p. 111). Agreement for the items in the final scale is reported as 87.6 % or Cohen's $\kappa = .75$ (Schröder et al., 2020, p. 114). For the foreign language or EFL context, no research on RobS or other simulations that have an evaluative component are found to date. This gap will be further addressed in the next chapter.

4.5 Addressing the Assessment Gap in Teacher Education

Based on the framework of constructive alignment, the previous chapters have provided a possible approach as to how the performative aspect of competence

⁶¹ STEM is the abbreviation of Science, Technology, Engineering and Mathematics (*MINT* in the German context).

could be included in teacher education. As shown in the last chapter, assessment seems to be of particular importance for the students' learning process. However, here also seems to be the greatest gap, concerning alignment and research in (EFL) teacher education.

Assessment in the first phase of teacher education, i.e., at the university level is rather focused on the knowledge-related levels (*knows*, *knows how*), while in the second phase, assessment is rather immediately conducted under the full complexity of the situation (*does*), reflecting the statutory requirements by the KMK (see Chapter 3). Hence, there is a gap for the assessment of the performative level, i.e., *shows how*, as it is not included in the different assessment formats at university. Relating the levels of assessment stated by Miller (1990) to the model of competence as a continuum by Blömeke et al. (2015), it is evident that the performative level of competence, i.e., observable behavior, is not included in the assessment formats at the university level, but only by graded classroom observations in the second phase of teacher education (see Figure 12). As previously shown, RobS have been successfully implemented as an assessment format for the level of *shows how* in medical education for decades, e.g., as part of OSCEs (Sopka et al., 2018, p. 98). While RobS do exist in the research context of teacher education (see Chapter 4.3.2), they are not comprehensively implemented at the university level, especially as assessment formats (J. Meier et al., 2023, p. 91). The gap in assessment is thus reflective of the alleged gap between theory and practice, as perceived by PSTs in general, and indicated by the challenges PSTs have in applying their knowledge to practice (Zeichner, 2010, p. 91; see Chapter 3.5). Performance assessment in the sense of Miller (1990), i.e., *shows how*, could thus contribute to closing that perceived gap. Such a focus would also allow for the assessment of the performative facet of competence at university, as described by Blömeke et al. (2015), extending the options and the repertoire of competence assessment. Thus, RobS can still be described as measuring competence, only focusing on the performative aspect of the continuum. As summative assessment has an influence on learning, such a performance-oriented and authentic format could lead to PSTs being better prepared for practice as they adopt a deeper approach to learning (see Chapter 4.4.2). As in medical education, role-play-based simulations are a possible assessment format for that level. There is precedent for simulation-based learning activities in teacher education, also in the form of microteaching (see Chapter 4.3.2). The simulations, either in the context of SBL or assessment, that the already developed overwhelmingly stem from the subjects that are not related to the foreign language classroom, except for *DiaSim* (Grünwald & Levin, 2022). While the concept of RobS as found in medical education may be adaptable to situations in the individual subjects, these simulations need to be developed by the

individual subject-pedagogies to ensure the subject-specific core, while still building upon research from other areas and professions (Marks, 2023, p. 34). For RobS to be used for assessment purposes, however, specific criteria need to be considered (Kaufman & Ireland, 2019, p. 4).

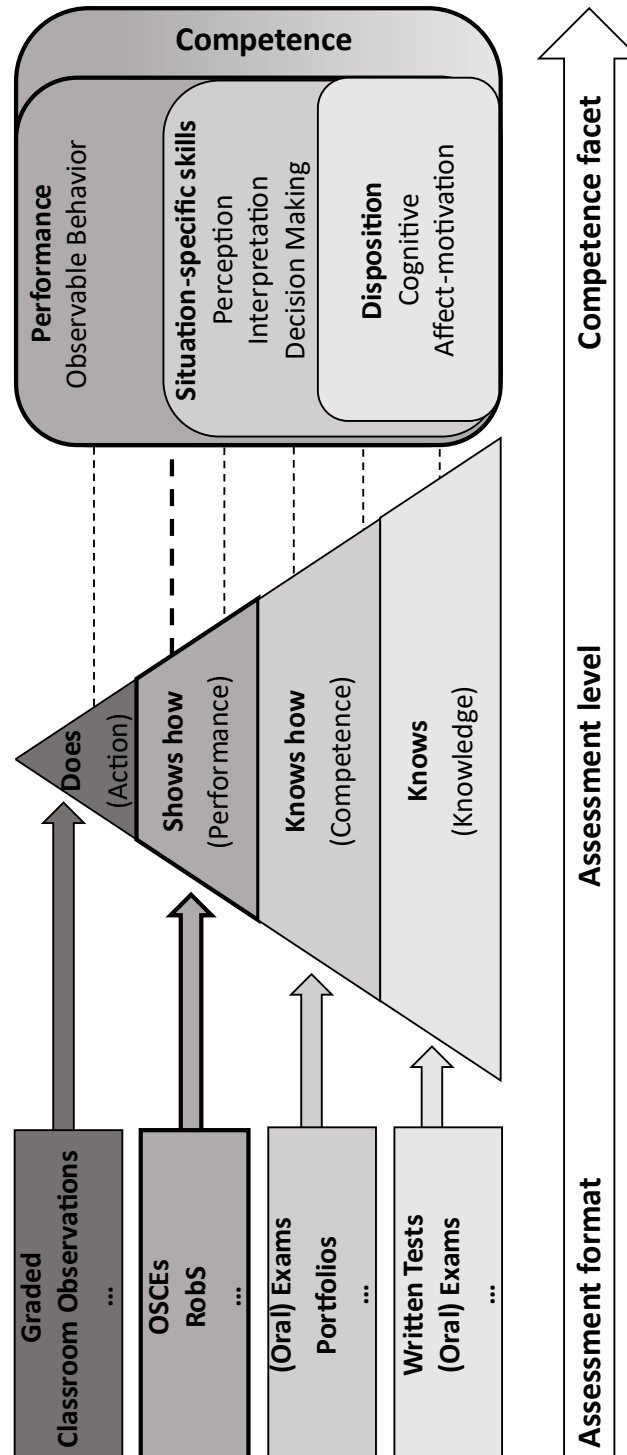


Figure 12: Relation of assessment formats in teacher education, assessment levels (Miller, 1990) and competence facets (Blömeke et al., 2015).

Based on the review of previous research, both conceptually and empirically, some guidelines can be derived for the development of RobS for the purpose of summative assessment. Considering the focus of the RobS, core practices are a suitable concept as they can be used to describe specific practices that teachers need to be able to master. They can be related to the performance-oriented facet of competence as described in the model by Blömeke et al. (2015). Core practices extend the goal of what PSTs need to be capable of beyond a mere focus of knowledge as found in the descriptors by the KMK, especially in the first phase of teacher education. Nonetheless, in order for the RobS to be implemented sensibly at German universities, the focus of the RobS should also tie into the statutory guidelines. When designing simulations, the characteristics of simplification, authenticity, flexibility, and structure must be considered (see Chapter 4.3.1). The simulation should approximate the core practice that is chosen to be the focus of the RobS. This approximation should be sufficiently authentic, i.e., in terms of the model by Gulikers et al. (2004), especially from the perspective of the PSTs. Their perception is decisive for how they steer their learning process. Fidelity, as discussed, may contribute to authenticity, but is not a necessary criterion for an authentic perception (see Chapters 4.3.3 and 4.4.2). The simulation should be flexible, in that it is adaptive to the participant. Especially for assessment purposes, however, a high degree of standardization is also required in the RobS, as it must be ensured that the assessment is perceived as fair and each PST has the same opportunity to show their capabilities, i.e., what they have prepared for. For the use of simulations as summative assessment specifically, Kaufman and Ireland also point out that “the simulation needs rigorous validity” (2019, p. 4). Discussing the extent to which a newly developed test can be considered valid is essential to understand its merits, and is a prerequisite to its use and implementation.

Consequently, the aim of this thesis is to develop and validate a role-play-based simulation in the EFL context that can be used for summative assessment.

5 Summary

The primary goal of teacher education can be described as developing teachers' competence so that learners' achievement is impacted most positively. The presented flow charts all put the instructional behavior of teachers at the nexus point, as their actions and behavior most clearly influence learning. In order to develop the competence of teachers, there is a three-phasic structure in Germany. The first phase is university-based and includes the learning of subject knowledge and first practical phases to gain experience in the profession. The

second phase is more practice-oriented and takes place at schools and specific institutions, while the third phase includes the further education of in-service teachers.

In order to prepare teachers for practice, the aim of teacher education is to develop the required competences. Looking at the meaning behind the term competence, it is evident that there is no real consensus between or within disciplines. Based on the frequently used definition by Weinert (2014), some characteristics can still be identified that can be attributed to competence in general, e.g., their learnability, distinguishability, measurability, performativity, and specificity. For teacher education, several models of competence are proposed. The model of professional competence by Baumert and Kunter (2006, 2013) describes several facets, but describes knowledge as the core of competence and a prerequisite for performance. Neuweg (2014) conceptualizes different kinds of knowledge, differentiating between knowledge that teachers are supposed to learn (*knowledge 1*), have learned (*knowledge 2*), and their action (*knowledge 3*). The model by Blömeke et al. (2015) sees competence as a continuum which ranges from cognitive and affect-motivational dispositions, over situation-specific skills to performance as the observable behavior of competence, consolidating the differences and proposing a unifying approach to competence. The frameworks that prescribe the content and competencies that are to be included in the first and second phase show that knowledge is the primary focus at university (Blömeke, 2006; KMK, 2019a, 2019b). Empirical findings into professional knowledge of (pre-service) EFL teachers provide mixed results regarding the development of different facets (Jansing et al., 2013; Kirchhoff, 2017; König et al., 2016). Overall, there seems to be an increase of professional knowledge throughout the different phases of teacher education. Nevertheless, PSTs in general do not feel adequately prepared to apply that knowledge in action (Zeichner, 2010). Emphasizing the performative aspect of competence in the first phase of teacher education could be one approach to address this issue.

The framework of constructive alignment describes the congruence between the learning objectives, teaching and learning activities, and assessment for higher education course design (Biggs, 1996; Biggs & Tang, 2011). Using that framework, a performance-oriented approach to the first phase of teacher education can be outlined. Learning objectives can be described as core practices, which at least might expand the knowledge-oriented descriptors for competence (Grossman, 2018). As TLAs, a simulation-based approach to learning can be employed (Chernikova, Heitzmann, Stadler, et al., 2020). Simulations have been used in other domains for decades, and a precedence can be found in microteaching scenarios (Cavanaugh, 2022). Simulations allow for performance-oriented training in a situation that is an “approximation of practice” (Grossman,

Compton, et al., 2009, p. 2076), and thus simplified, but still authentic, flexible and standardized (Dieker et al., 2014). SBL has shown to be effective for acquiring more complex skills (Chernikova, Heitzmann, Stadler, et al., 2020). Simulations that are considered to be more authentic tend to be more effective than those that are less authentic to a certain extent, but it is difficult to design simulations that are perceived to be authentic by the specific target group (Chernikova, Heitzmann, Stadler, et al., 2020; Chernikova et al., 2024). One common approach is that of role-play-based simulations, i.e., interactions with trained actors, as the use of actors may contribute to an authentic situation (Levin, Freilanda, Flavian, & Miller, 2023). Assessment can have an impact on the way students learn (Struyven et al., 2005), especially their perception of authenticity may be influential (Gulikers, Bastiaens, et al., 2008). However, traditional assessment formats, i.e., written tests, are generally not considered authentic, while learner-centered approaches, i.e., portfolios, are (Pereira, Assunção Flores, & Barros, 2017). RobS might be used as authentic, performance-oriented and learner-oriented assessment formats, as is done in medical education (Sopka et al., 2018). In teacher education, simulations have been designed for the fields of general education and STEM for evaluative purposes, but mostly in research contexts (Albu & Lindmeier, 2023).

Other than in medical education, simulations are not comprehensively implemented for summative assessment purposes at universities in the first phase of teacher education. PSTs also do not consider traditional assessment to reflect the demands of the profession (Wagener et al., 2019). Comparing the traditional assessment formats of teacher education against the frameworks of Miller (1990) and Blömeke et al. (2015), it is evident that there is a gap of assessment formats on the performance level in the first phase. This is reflective of the general perception of a theory - practice gap in teacher education. RobS, similar to medical education, could fill that assessment gap in teacher education, if they can fulfill the outlined requirements. Thus, developing and validating a RobS that can be used for summative assessment in teacher education is the aim of this thesis. This will be done in the context of EFL teacher education, as there is a general scarcity of simulations in this field. The next part of the thesis will focus on the test that is developed to address the gap, elaborating on its distinct topic of assessment and construction.

Part II -

Developing a Role-Play-Based Simulation (RobS) to Assess Feedback Competence on Writing in the EFL context

*Some countries have made declarations about the year of feedback
(I wonder what they do the following year, drop it?)*

John Hattie (2023, p. 320)

The focus of this part of the thesis is on the first part of the aim: The construction and development of the RobS for summative assessment. First, the topic of the RobS is reasoned and, thus, the relevance of feedback for the learning and writing process is discussed and the necessity for the assessment teachers' competence in that regard is emphasized. Second, the framework will be presented in detail. The framework is preceded by a reflection on the methodological approach taken for its construction and a review of existing evaluative tools, and concludes with a reflection of its limitations. Third, the RobS is presented in detail, structured along its three distinct phases, preparation, simulation, and assessment, concluding with limitations that relate to the test design.

6 The Relevance of Feedback on Writing in the EFL context

While the focus of this thesis is on developing and validating a RobS, it is not less important to discuss the relevance and reasons for the choice of the competence that is focused on: feedback on writing in the EFL classroom. This chapter will explore the different components and make the case for why feedback on writing is a suitable focus for the RobS.

6.1 Feedback

It is hard to find a topic in educational research that has been more prominent in the 21st century than that of feedback. The introductory quote by Hattie emphasizes this, while simultaneously criticizing that feedback is not consistently made a priority (2023, p. 320). A very brief history and clarification on the definition of the term will serve as a gateway into this topic, before the relevance of feedback is illustrated by showing empirical evidence, theoretical approaches, and relevant guidelines for (EFL) teacher education. Regarding feedback, the amount of available research on theoretical and empirical findings has continued to increase in the last decade(s), especially since the meta-studies by Kluger and DeNisi (1996) and Hattie and Timperley (2007). It is thus not the purpose of this chapter to account for all possible approaches, empirical results and

limitations in the field, but rather to point to key findings that illustrate the significance of feedback⁶².

6.1.1 On the History and Terminology of Feedback

Historically, systems theory can be used as a starting point to discuss feedback (Boud & Molloy, 2013, p. 701; Narciss, 2006, pp. 15–16). A dynamic system⁶³ can be described as striving to have a control variable reach a certain goal or keep it at a specific level. An outdoor swimming pool, for example, should have a minimal water temperature (control variable) of 25 °C. Within that system, there is a regulation device which seeks keep to this goal, e.g., the water temperature at 25 °C. Part of that system is a measuring device, which controls for the actual state of the control variable, e.g., a thermometer. This device provides the information for the system about the actual state of the water temperature – it provides *feedback*. Based on that feedback, different options can be pursued to adjust the control variable, e.g., add more hot water if the temperature is too low. Based on the feedback, there needs to be an adjustment because “[if] the information on the gap is merely stored without being utilized to alter the gap, it is not feedback” (Ramaprasad, 1983, p. 5). The purpose of feedback from a systems-theoretical perspective is thus of a corrective nature, initiated and required by the diagnosis of a gap between the actual and target state (Narciss, 2006, pp. 15–16).

The perspective from systems theory offers a descriptive and neutral model, against which conceptualizations of feedback in education can be discussed diachronically (Boud & Molloy, 2013, p. 701; Kopp & Mandl, 2014b; Narciss, 2006, pp. 15–16). While at its core, feedback has always focused on the transmission of information, its conceptualization has also been a reflection of current streams in science in educational and learning research (Schluer, 2022, p. 16). Researchers with a background in behaviorism saw feedback as a means to influence behavior towards a desired goal they have set. By using reinforcement as a type of feedback, certain behavior is positively rewarded in the hope of increasing its frequency, examples for this can be found in the works of Thorndike or Skinner (Fürstenau, 2019, pp. 15–16). While behavioristic approaches have been considered inappropriate for describing learning processes in humans because “input [...] does not equal intake by learners” (Schluer, 2022, p. 16), the idea of feedback as a means to adjust behavior has remained. Later

⁶² For a comprehensive review and overview on models of feedback throughout history, see, i.e., Lipnevich and Panadero (2021).

⁶³ The provided description here is rather superficial and focuses on the relevant aspects of systems theory. See Narciss (2006, pp. 14–19) for a detailed description.

definitions of feedback in the educational context emphasize the information that is provided to a person and take a cognitivist perspective (Schluer, 2022, p. 16). Here, feedback is seen as unidirectional, monologic and focused on the transmission of information from the teacher to the learner (Ajjawi & Boud, 2017, p. 253). The purpose is to provide information on a potential gap between the current and desired state. Mostly, the provided information focuses on whether the desired state is achieved, i.e., knowledge of result (KR), or it includes information on what a correct solution for the desired state could be, i.e., knowledge of correct result (KCR) (e.g., Lotz, 2016, p. 133).

A common definition, which fits this view for feedback, is found in the paper by Hattie and Timperley⁶⁴ (2007). They state that "feedback is conceptualized as information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007, p. 81). In this definition, feedback is also considered the consequence of performance, similar to previous approaches, and transmission of information is emphasized (Hattie, 2009, p. 174; Rütli-Joy, p. 55). Hattie and Timperley (2007) also offer a prominent model that has frequently been used to describe feedback (see Figure 13).

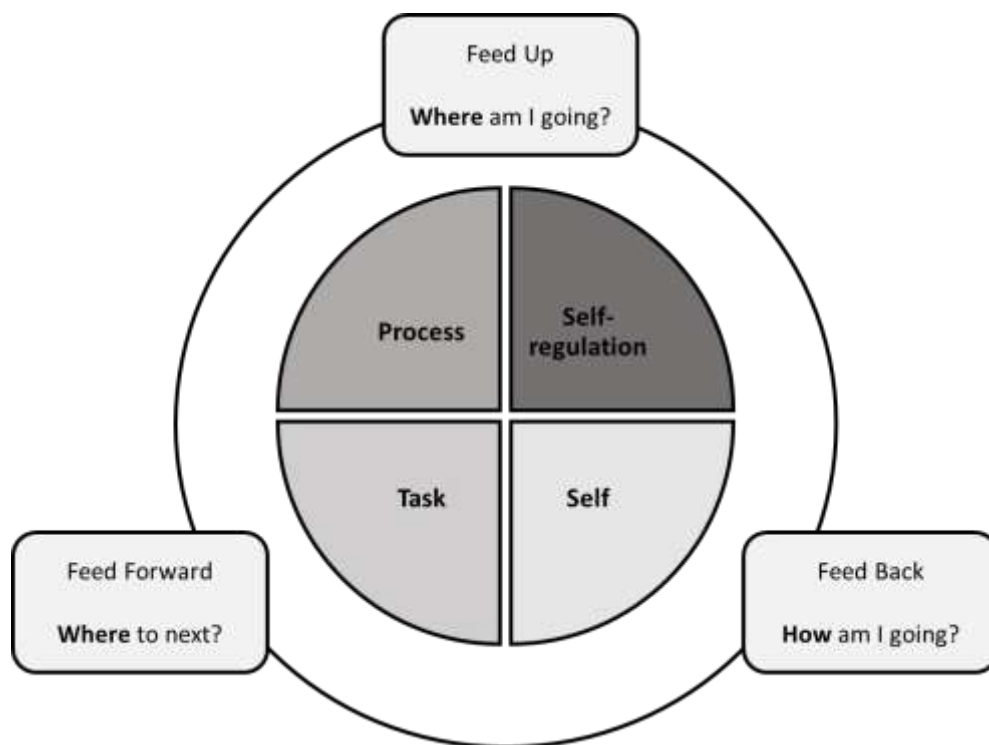


Figure 13: Model of feedback, according to Hattie and Timperley (2007); figure adapted from Zierer (2019).

⁶⁴ Kluger and DeNisi (1996, p. 255) offer a similar definition.

Feedback can be aimed at different levels: *Task*, *Process*, *Self-regulation* and *Self*. The task-level focuses on the product, the process-level on the activities taken to achieve the product, and the self-regulation-level on the meta-cognitive strategies that can be used to direct learning behavior, while feedback on the self-level relates to the person directly (Hattie & Timperley, 2007, p. 87). These levels can be targeted with specific questions that have particular aims: *Feed Up* – *Where am I going?*, i.e., what is the desired state, *Feed Back*⁶⁵ – *How am I going?*, i.e., how does the current state relate to that desired state, and *Feed Forward* – *Where to next?*, i.e., what steps need to be taken in order to achieve the desired state⁶⁶.

More recent perspectives on feedback, which originated from research in higher education, reject the notion of unidirectionality and take a (socio-)constructivist approach, based on the ideas of Vygotskij (1978) (Ajjawi & Boud, 2017; Carless, 2020; Carless & Boud, 2018; Carless & Winstone, 2020). Learners are included as active co-constructors in the process and feedback should thus be dialogic and bi- or multidirectional (De Florio-Hansen, 2023, p. 107; Schluer et al., 2023, p. 156). Providers of feedback should, therefore, account for the learners' individual knowledge level, needs and wishes and be able to adapt accordingly. The definition by Hattie and Timperley (2007) is also not exclusive to a cognitivist approach, if the feedback context is interactional. In order for feedback to be effective, e.g., aiding learners in reducing the gap between the current and desired state (Sadler, 1989, p. 142), learners need to be able to use the information and work with the feedback in order to learn from it (Lee, 2017, p. 78; Narciss, 2014, 77–78). Consequently, feedback involves "shared responsibilities" (Carless & Winstone, 2020, p. 150) between teachers and learners. Feedback as such could thus broadly be viewed as "interaction between teacher and student carried out for the purpose of student learning" (Hawe et al., 2008, p. 45). Schluer provides a definition of feedback which accounts for both the cognitivist as well as constructivist perspective of feedback as "an interactive process of exchanging information that persons can use in order to regulate their further learning and thus improve their performance" (2022, p. 17). This definition works well in the context of this thesis as it includes both the aspect of interaction, which is inherent to RobS, as well as the information and purpose of those interactions. This brief historical outline provides a first insight at what Lipnevich and Panadero highlight in their comprehensive review: There is a plethora of different

⁶⁵ Feed back is used in this thesis differentiation to feedback. The first relates to a specific level of the feedback model by Hattie and Timperley (2007), while the latter refers to the general notion.

⁶⁶ The model's components will be discussed in more detail in Chapter 7.3.1.

models and concepts of feedback, which might all be appropriate for their specific context (2021; Schluer, 2022, p. 17)⁶⁷.

6.1.2 Feedback as Literacies

Recently, the concept of feedback literacy has been introduced in research. The term literacy has evolved from its original focus on reading and writing; concepts such as that of *Multiliteracies* extend the term, adding that context and mode need to be included in meaning making and negotiation in modern communication (Kalantzis et al., 2016, pp. 1–5). Thus, being literate means being able to make, and participate in meanings so that an individual can partake in, and contribute to society (Kalantzis et al., 2016, p. 3). Consequently, in concepts of feedback literacy, feedback is predominantly regarded from the perspective of the learner, shifting away from the perspective of feedback as a teacher-centered transmission of information:

[Feedback] is defined as a process through which learners make sense of information from various sources and use it to enhance their work or learning strategies. This definition goes beyond notions that feedback is principally about teachers informing students about strengths, weaknesses and how to improve, and highlights the centrality of the student role in sense-making and using comments to improve subsequent work. (Carless & Boud, 2018, p. 1315)

As the idea of feedback literacy has gained momentum, different concepts have been proposed to grasp it more clearly (Lipnevich & Panadero, 2021, p. 18). Originally, the term stems from assessment and academic literacies (Schluer et al., 2023, p. 156; Sutton, 2012, pp. 32–33) and generally, a distinction between student feedback literacy and teacher feedback literacy is made. Carless and Boud, who developed a prominent model, define student feedback literacy as “understandings, capacities and dispositions needed to make sense of information and use it to enhance work or learning strategies” (2018, p. 1316). Their model consists of students *Appreciating feedback processes*, i.e., realizing its worth and that they need to play an active part in them, *Making judgements*, i.e., self-evaluation⁶⁸ as well as evaluating work of others, e.g., in peer-feedback situations, *Managing affect*, i.e., regulating one’s emotions and not interpreting negative feedback on a performance as negative, and *Taking action*, i.e., using feedback and working with the comments to improve (further) performance and close the gap (Carless & Boud, 2018, pp. 1316–1318).

⁶⁷ Panadero and Lipnevich (2022) also propose an integrative model of feedback elements, based on their review.

⁶⁸ See also Tai et al. (2018).

In order to foster student feedback literacy, teachers need to be equipped with the necessary resources to do so, i.e., teacher feedback literacy, which can be defined as “the knowledge, expertise and dispositions to design feedback processes in ways which enable student uptake of feedback and seed the development of student feedback literacy” (Carless & Winstone, 2020, p. 154). The authors name three dimensions of teacher feedback literacy:

- 1) The *design dimension* focuses on the teachers’ ability to create and foster an environment in which student uptake is fostered. It is clear that errors are opportunities to learn (Schluer, 2022, p. 2019), feedback is subsequently acted upon and students’ abilities to make reasonable judgment is fostered (Carless & Winstone, 2020, pp. 154–155).
- 2) The *relational dimension* addresses the interactional and personal facet and relates to the teachers’ abilities to create an environment in which students’ emotions are accounted for by providing feedback information in a “supportive [and] constructive way” (Carless & Winstone, 2020, pp. 155–156).
- 3) The *pragmatic dimension* represents the notion of compromise. Teachers must account for many different external factors such as resources of time or curricula and must make reasonable and pragmatic decisions (Carless & Winstone, 2020, pp. 156–157).

Importantly, for feedback processes to be effective, they must consider the specific discipline, which means that teacher feedback literacy is also domain specific (Carless & Winstone, 2020, p. 154). While the binary distinction between teacher and student feedback literacy highlights the responsibilities that teachers and students have in the feedback process, they remain shared (Carless & Winstone, 2020, p. 150). As teaching and learning are intertwined and responsibilities are shared, Tai et al. oppose this binary view and propose the use of the plural “feedback literacies” (2023, p. 203), which emphasizes the co-constructive nature of feedback processes (Schluer et al., 2023, p. 157). While feedback centers around the learner as they receive the information and must be enabled to use the feedback appropriately (student feedback literacy), it is still the responsibility of the teacher to create an environment in which this can happen (teacher feedback literacy). This includes the arrangement of an adaptive and dialogic setting, as well as the provision of appropriate information, which the learner can use to close the gap. After the concept of feedback has been discussed from a theoretical perspective, the relevance of feedback will be discussed.

6.1.3 Relevance of Feedback in (EFL) Teacher Education

The relevance of focusing on feedback in the RobS will be discussed from two perspectives: 1) the empirical evidence indicating that feedback has an impact

on learning, and 2) the statutory and theoretical demands of (EFL) teacher education.

While some scholars argue that “[the] positive effect of feedback on students’ performance and learning is no longer disputed” (Lipnevich & Panadero, 2021, p. 1), it is nonetheless worthwhile to discuss these effects as this further supports the choice of focusing the RobS on feedback. Several meta-studies of first and second degree focus on feedback and come to the similar conclusion that feedback can have a substantial impact on learning. In their broad review of feedback interventions, Kluger and DeNisi find that feedback interventions might improve performance with an average effect of $d = 0.41$, but about one third of interventions have a negative impact on performance, arguing that “feedback intervention (FI) effects on performance are quite variable” (1996, p. 254). It is thus not necessarily the case that FIs that positively influence the performance in one group have similar effects in another group. The authors single out FIs that contain praise and focus on the self (Shute, 2008, p. 169) to potentially decrease achievement (Kluger & DeNisi, 1996, p. 275). While the analysis conducted by Kluger and DeNisi does not differentiate for the context of feedback (1996, p. 257), only educational contexts are included in the analysis by Bangert-Drowns et al. (1991, p. 218). Like Kluger and DeNisi (1996) they still find a positive, although smaller effect size of $d = 0.26$ (1991, p. 224) and a similar share of negative feedback effects of about a third (1991, p. 232). Potentially the most influential meta-study on “The Power of Feedback” (2007) was published by Hattie and Timperley. Other than the previous studies, they include 12 meta-analyses on the effect that feedback in the classroom has on student achievement to conduct a second degree meta-analysis, finding an effect size of $d = 0.79$ ⁶⁹ (Hattie & Timperley, 2007, pp. 82–83). This effect size is large, especially when compared to other variables influencing student achievement, putting it in the tenth place of the most influential factors in Hattie’s initial *Visible Learning* study (Hattie, 2009, p. 173). Recently, these results have been revisited in a new meta-analysis (Wisniewski et al., 2020). This update was considered necessary as the methodological approach taken by the initial meta-analysis might have led to skewed results due to using a fixed-effect model and not accounting for overlapping studies (Wisniewski et al., 2020, p. 2). The adjusted effect size, without outliers, is described as $d = 0.48$, which is substantially less but still moderate (Wisniewski et al., 2020, pp. 10–11). For the original set of studies, a corrected effect size is described as $d = 0.53$ (Hattie, 2023, p. 320). All studies, however, confirm previous results which indicate that there is great variability in

⁶⁹ Hattie (2009, p. 173) describes an effect size of $d = 0.73$.

the effect of feedback, depending on its form (Hattie, 2023, p. 319; Wisniewski et al., 2020, p. 12).

While the impact that feedback has on learning and achievement has been detailed as one reason for focusing on feedback in the RobS, it should also be taken into account that feedback is included in guidelines for teacher education in Germany. As previously outlined, the KMK has outlined the objectives for teacher education in Germany. In the more general *KMK04/19*, the notion of feedback is not found explicitly, but is subsumed as part of the competence area "Assessing/evaluating" (*Beurteilen*) as is outlined in Chapter 3.3 (KMK, 2019b, pp. 11–12; translated by TJ). For EFL teacher education particularly, the *KMK08/19* states that at the university level, PSETs are supposed to know about specific approaches to diagnosis, assessment and support measures, which are based on results of the respective evaluation (KMK, 2019a, p. 46). What is more, they are supposed to know about specific measures to support the development of communicative, intercultural and text-related foreign language competence, and know the basics of subject-specific diagnosis and assessment (KMK, 2019a, p. 44). However, Helmke explicitly criticizes that diagnosis and assessment only play a minor role in German teacher education, stating the following:

Gemessen an der offenkundig großen Bedeutung der pädagogischen Diagnostik in der Schulpraxis, spielt dieser Bereich in der deutschen universitären Lehramtsausbildung eine untergeordnete Rolle. [...] Dieses Versäumnis kann auch von der 2. Phase der Lehrerausbildung nicht kompensiert werden, zumal diese ohnehin überfrachtet ist und verschiedentlich sogar noch verkürzt wurde. (Helmke, 2022, p. 115)

(Considering the obvious relevance of pedagogical diagnostics at school, this area plays a minor role in German university-based teacher education. [...] This failure cannot be compensated in the second phase, as it is already overburdened and has also been shortened occasionally.)

Helmke argues that the failure to include diagnosis and assessment at the university level of teacher education cannot be compensated in the second phase, which is already overloaded and shortened. As a consequence, university-based teacher education should include this aspect and not rely on the second phase to develop related competences.

Feedback is also repeatedly mentioned as a core practice for teachers. Schutz, Grossman and Shaughnessy, for example, state that the provision of feedback on writing could be fostered in a core practice approach through approximation (Schutz et al., 2018, p. 58). Providing feedback to learners during task phases is also framed as a core practice (Prilop et al., 2024, p. 188). The *TeachingWorks* consortium of the University of Michigan also specifically includes feedback in their list of core practices:

Providing oral and written feedback to students

Effective feedback helps focus students' attention on specific qualities of their work; it highlights areas needing improvement, and delineates ways to improve. Good feedback is specific, not overwhelming in scope, focused on the academic task, and supports students' perceptions of their own capability. Giving skillful feedback requires the teacher to make strategic choices about the frequency, method, and content of feedback and to communicate in ways that are understandable by students. (Grossman, 2018, p. 168; emphasis in the original)

While there is generally scarce research on core practices, specifically for the EFL classroom, feedback is consistently part of that discussion (Barahona, 2020; Pang, 2019, p. 265). Pang frames the core practice as providing "feedback on use of language" (Pang, 2019, p. 265), while Barahona, who specifically addresses the context of Chile, focuses more specifically on "[providing] oral corrective feedback to improve learner performance (Barahona, 2020, p. 6). It is not surprising that the focus is specifically on the linguistics side of the EFL classroom considering the long history of research on corrective feedback (Erlam et al., 2013).

In conclusion, findings indicate that feedback can have a substantial impact on student learning and achievement, however, this potentially includes both positive and negative effects. While the impact is, on average, moderate, there is great variability. It is therefore essential to look at the specific instructional context in which the feedback is provided, in order to make a sound argument of effective feedback methods, approaches, strategies and techniques (Schluer, 2022, p. 31). Additionally, both the statutory guidelines and research on core practices indicate that teachers need to be able to provide feedback and that this should be focused more in the first phase, i.e., university (Helmke, 2022, p. 115). A focus on feedback within the RobS can thus easily be reasoned, as teachers need to be able to master the practice of providing feedback. Additionally, the RobS could also contribute to counteracting the minor role that feedback plays in university-based teacher education. The next chapter will address the kind of performance that is addressed in the feedback: EFL learners' writing competence.

6.2 Writing in the EFL classroom

Writing is a central aspect of communicative competence and a complex skill that needs to be developed in order to participate in society. In a consistently more globalized world that uses English as a lingua franca (M. Dewey, 2007, p. 332), people must master writing in more than one language. However, to date, there seems to be no empirically-based approach to teaching writing in the second, or for that matter, any additional language (Schnell, 2018, p. 13).

Nonetheless, there are models which describe the cognitive process of writing, approaches that propose methods, strategies and techniques for teaching writing and guidelines that describe competence levels that learners should strive for.

6.2.1 Cognitive Process Models for (L1-) Writing

In order to provide a foundation to discuss writing, the model of Hayes and Flower will be presented in detail as it is regarded as the “the most influential model” (Zimmermann, 2000, p. 74). Other models for L1⁷⁰ writing by Kellogg (1996) and Scardamalia/Bereiter (1987) and those that consider the processes of writing in the L2/FL⁷¹ will be presented briefly⁷².

While the original model on the cognitive process during writing by Hayes and Flowers from 1980, as seen in Figure 14, has been revised and updated frequently (Hayes, 2012, p. 370), it has proven to be the most prominently used version (Schnell, 2018, p. 33).

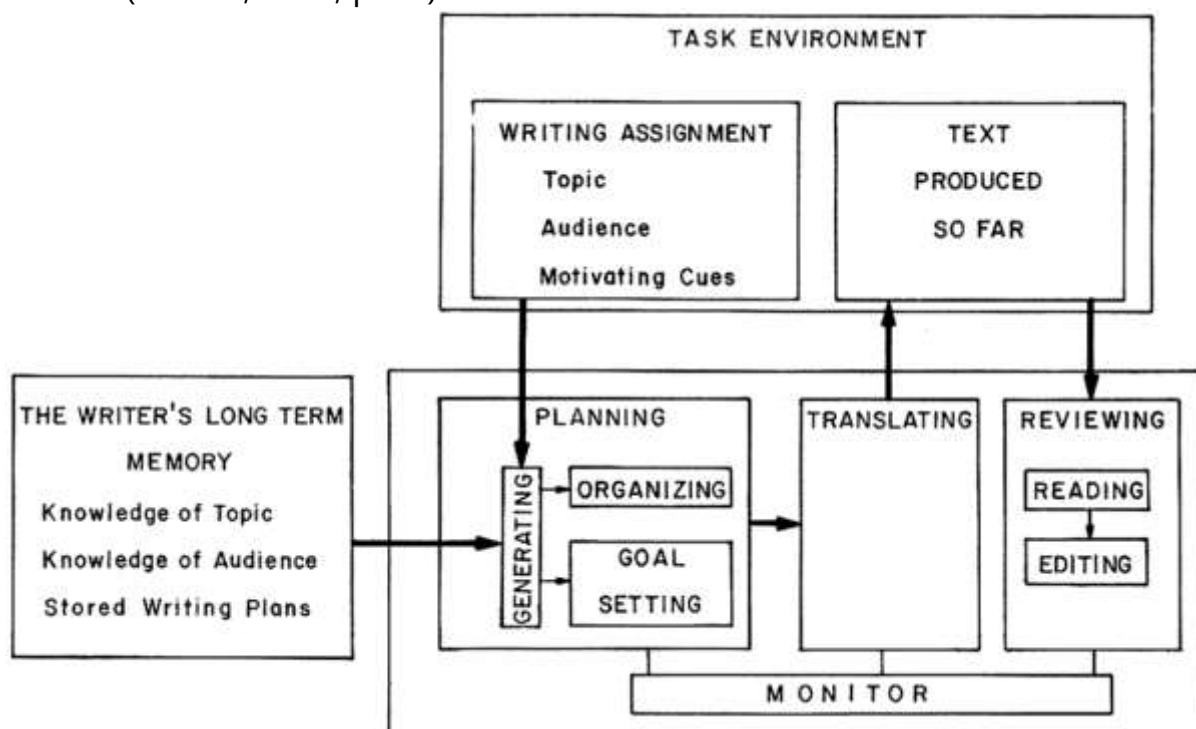


Figure 14: The Hayes-Flower model (1980, as cited in Hayes, 2012, p. 370).

⁷⁰ L1 describes the first language(s) of a learner.

⁷¹ L2 refers learners of a second language, i.e., a language in addition to L1(s) which they acquire in the environment in which the language is spoken, e.g., migrants, while FL relates to the learning of an additional language in the home country, e.g., through instruction at school (Scherer et al., 2024, p. 3).

⁷² See, e.g., Schnell (2018) for a more comprehensive overview.

The model consists of three basic units: “the task environment, the writer’s long term memory, and the writing processes” (Flower & Hayes, 1981, p. 369)⁷³. It does not present a fixed linear sequence in which writing takes place, rather, these different units present an iterative and complex process in which the units are drawn upon in an integrative manner (Flower & Hayes, 1981, p. 387). The task environment represents the demands that the learner has to fulfill in the written products, i.e., what the text is about, for whom it is written and what else is required for the task (Flower & Hayes, 1981, pp. 370–371). The written text that is produced is also relevant as it shows progress, but it also restricts the writer by limiting what can be written next (Flower & Hayes, 1981, p. 371). The long-term memory refers to the prior knowledge of the writer which can be drawn upon in the process in terms of content, but also with regard to audience orientation, e.g., through generic schemes. In later iterations, Hayes also adds the working memory, stating that it “repaired an obvious oversight in the original model” (2012, p. 370).

For the writing process, Hayes and Flores distinguish between three different groups of processes: *Planning*, *Translating*, and *Reviewing*, with the monitor as a fourth component regulating the processes (Flower & Hayes, 1981, pp. 372–375). In the planning process, writers create an „internal *representation* of the knowledge that will be used in writing” (Flower & Hayes, 1981, p. 372; emphasis in the original). This includes the non-linear, but recursive elements of generating ideas, based on the task and long-term memory, organizing ideas and setting goals (Schnell, 2018, p. 27). In later models, Hayes argues that, based on an understanding of planning as a writing activity, it is not a part of the writing process as “a separate planning process would simply duplicate an activity than can already be performed by the writing model” (Hayes, 2012, p. 376). The translating process then turns the mental representation into the written form (Flower & Hayes, 1981, p. 373). Translating does not specifically refer to the act of conveying meaning in a different language and can also be conceptualized as “formulating” (Zimmermann, 2000, p. 75). In later models, a transcription process, i.e., the actual act of writing or typing according to specific norms is added to account for the challenge that this process entails, especially for early learners (Hayes, 2012, pp. 371–372). The reviewing process consists of the writer reading, evaluating and, if necessary, revising, the produced pieces of text (Flower & Hayes, 1981, p. 374). Importantly, such reviewing processes can occur throughout the whole of the writing process, emphasizing the non-linear approach that

⁷³ The model in Flower and Hayes (1981) uses slightly different terms. For consistency, only the terms of Hayes and Flowers (1980) are considered when citing from updated versions that refer to identical elements.

the authors take in writing (Hayes & Flowers, 1980, p. 29). In the model, this revision process is focused on the writer, i.e., it focuses on internal feedback (Flower & Hayes, 1981, p. 374). Similar to the planning process, Hayes later argues that reviewing is a distinct writing activity, and thus merely a “special application of the writing model” (Hayes, 2012, p. 376). The monitor, as a last component, describes the regulation of these processes from a metacognitive perspective (Flower & Hayes, 1981, pp. 374–375; Schnell, 2018, p. 31). In its entirety, the model is similar to other cognitive processes of communicative skills (e.g., Goh & Burns, 2012, p. 36 for a model on speaking based on Levelt, 1995; Weir & Khalifa, 2008 for reading; Goh & Vandergrift, 2021, p. 25 for listening).

Other relevant models of writing that will briefly be outlined are that of Kellogg (1996) and Bereiter and Scardamalia (1987). Kellogg’s model is quite similar to the original Hayes and Flower model. It places a focus on the working memory and the systems that relate to these processes, the visuo-spatial sketchpad, the central executive and the phonological loop, and relates them to the distinctly described writing processes of *Planning*, *Translating*, *Programming*, *Executing*, *Reading* and *Editing*, which might also be activated simultaneously, if “the demands placed on the central executive do not exceed capacity limitations” (1996, pp. 58–59). Bereiter and Scardamalia (1987) describe the writing process in two models of strategies. The knowledge-telling model is predominantly used by novices like L2 learners and describes the strategy (or lack thereof) when writing in a nonorganized and linear matter, mostly to simply convey information using the available content and discourse knowledge, similar to oral language production (Bereiter & Scardamalia, 1987, pp. 9–10). The knowledge-transforming model is rather used by experts in writing, who also consider the purpose and genre of the text and try to compose the information in an appropriate manner to achieve their goals (Bereiter & Scardamalia, 1987, pp. 11–12). These strategies might be used to describe different levels of writing expertise, as the knowledge-telling is part of knowledge-transforming.

With regard to L2-writing, several models have been developed that are based on the presented models. Börner literally used the model by Hayes and Flowers (1980) and added L1/L2 specifications for the components of the long-term memory, planning, translating and the monitor, as well as an extension of the teaching context and the notion of interlanguage (1987, p. 1346; Schnell, 2018, p. 82). Furthermore, Börner argues that especially the processes of formulating/translating and reviewing are central for L2 learners, which makes sense, since the act of writing and high accuracy are of special concern for learners (1989, p. 49). Similarly, the model of Krings extends the previous model by focusing and elaborating on specific subprocesses that relate to both L1 and L2, as well as to L1 or L2 individually, during the translation/formulation process

(1989, p. 397). It specifically details that, after plans have been expressed in the L1 (mentally) and are being expressed in the L2, typical interruptions for L2 learners take place that consist of *Identifying L2 Problems*, *Activating L2 Strategies*, *Evaluating L2 Problem Solutions*, and *Deciding on L2 Problem Solutions* (Krings, 1989, p. 397; Zimmermann, 2000, p. 78). The model of text production by Chenoweth and Hayes (2003, p. 113) is based on a study that included L1 and L2 learners (Chenoweth & Hayes, 2001). It provides an alternative conceptualization of the formulator for L2 learners, like Krings, but focuses more on the linguistic process. In their model, the results of the translator are word strings which are evaluated and potentially revised before being transcribed into text, which is subsequently also open for (re-)evaluation (Chenoweth & Hayes, 2003, p. 113).

Naturally, these models have also drawn criticism. All these models focus merely on the cognitive and procedural aspect and can thus not be used to describe writing holistically and do not allow for a clear distinctive description of the mental and verbal processes (Molitor-Lübbert, 1996, p. 1024). The model of Hayes and Flowers, as well as the other cognitively-oriented models, are criticized for excluding the role of emotions (Schnell, 2018, p. 35) and motivation in general, although Hayes has tried to account for that in the updated model (Hayes, 2012, p. 372). Much of the criticism on the original model has been accounted for by Hayes in later models (2012), e.g., the undifferentiated translation process is amended by the transcription process (Hayes, 2012, pp. 371–372; Zimmermann, 2000, p. 75). Thus, the models themselves are also in a constant process and continue to develop.

In conclusion, the presented models describe the writing process in three basic phases: 1) Planning, be it mentally or in written form, for a text or a sentence; 2) Formulating, including both the overall structure as well as linguistic norms considering orthography, lexico-grammar and syntax; 3) Revision, as in reading the produced text, correcting errors and restructuring elements (Busse, 2022, p. 10; Hayes & Flowers, 1980, p. 12; Hyland, 2019, p. 12). These three phases are explicitly not arranged linearly, but depict an iterative process, which is regulated by a monitor. While Hayes does not consider planning and reviewing as part of the writing process but as specific writing activities, they remain relevant for writing and this reconceptualization of these terms should be framed as “re-representing” (Hayes, 2012, p. 385) not as an indication to neglect them. The following chapter will turn to how writing can be approached in the EFL classroom.

6.2.2 Approaches to Writing in the EFL classroom

Over the years, writing instruction has been conceptualized differently for the EFL classroom. Different approaches to writing also put a focus on different aspects that learners need to be capable of when it comes to writing, emphasizing the complexity of writing competence. Hyland has identified six guiding concepts to teach writing in the second language, which should not be considered exclusive, but “complementary overlapping perspectives, representing potentially compatible means of understanding the complex reality of writing” (Hyland, 2019, p. 3). While initially developed for L2 learners, the basic tenets can be applied to the EFL classroom as well. The theories focus on

- 1) Language Structures,
- 2) Text Functions,
- 3) Creative Expression,
- 4) the Writing Process,
- 5) Content, and
- 6) Genre (Hyland, 2019, pp. 3–21).

A focus on language structures views text predominantly as a string of “grammatical structures” (Hyland, 2019, p. 4) which are supposed to follow a certain norm. Writing is thus a means to an end of acquiring specific linguistic patterns and producing language accurately (Hyland, 2019, p. 4). A common approach to teaching grammar through writing that focuses on structure is the “presentation, practice, and production approach (PPP)” (Müller-Hartmann & Schocker-von Ditfurth, 2014, p. 104). A text can thus also be used to make inferences about the L2 capabilities of a learner with regard to their linguistic knowledge, allowing to assess and provide feedback on language accuracy. Although such an approach might contribute to L2 acquisition, Hyland argues that, while “many L2 students learn to write in this way, a structural orientation can create serious problems” (2019, p. 5). The main issue is that such a structural focus does not necessarily take into account the communicative purpose of texts, which both the output and noticing hypothesis include as necessary (Lightbown & Spada, 2013, p. 115). Learners may concentrate on expressing single grammatical issues accurately, but cannot use them functionally (Hyland, 2019, p. 5).

The functional approach to writing can be viewed as an amendment that aims to address these gaps. Structures are not taught in isolation but are considered in tandem with their meaning and how they can be used to communicate effectively (Hyland, 2019, pp. 6–8). The structural and functional approach are both reminiscent of “writing to learn language” (Manchón, 2016, p. 139) approaches, in which writing is seen as a means to acquire skills and competencies

in the L2 or FL. Such a perspective is in line with second language acquisition theories that focus on comprehensible output (Swain, 1985). The approach assumes that

[learners] must be pushed to produce the language because this forces them to process the L2 syntactically, as they have to produce the language that expresses their needs most precisely and efficiently (rather than a so-called 'semantic' type of processing which is minimally required for comprehension and can be done by a cognitively mature listener by using a range of linguistic, discourse and situational cues). (Behney & Marsden, 2020, p. 42)

Creating written output is one way to produce language by which learners want to communicate. In the process, learners may also notice gaps in their capabilities, which is essential for the process of actually learning how to overcome that gap and produce the desired output, e.g., convey the meaning accurately and effectively (R. Schmidt, 2001).

Focusing on creative expression in writing puts the learner, their perspective, opinion, and identity at the forefront. This approach stems from traditional higher education, for example, classes on English literature in the UK and North America that consider writing as a "creative act of self-discovery" (Hyland, 2019, p. 9). The role of the teacher is minimized and there is little structure in instruction. For L2 learners, especially younger ones, such an approach is, however, not necessarily suitable as it might neglect the learners' individual cultural differences, ignore different potentials for creativity, as well as their linguistic capabilities (Hyland, 2019, p. 10).

While also putting the focus on the learner, the process approach to writing provides organizational guidelines. This approach is based on the already-outlined cognitive models of the writing process by Hayes and Flowers (1980) and Bereiter and Scardamalia (1987) (see Chapter 6.2.1) and adopts them for writing instruction. The teacher's main task is to guide the learners through the writing process, which, based on the previous research, can be described as "[planning], drafting, revising and editing" (Hyland, 2019, p. 11). Using guided activities, learners can be supported in the planning and drafting process, while teachers can provide feedback during revision, which can and should be used when editing the text. It is the revision phase which

is potentially one of the most influential aspects of a process writing class, and the point at which the teacher's intervention is most obvious and perhaps most crucial. Not only does this individual attention play an important role in motivating learners, but it is also where overt correction and explicit language teaching are most likely to occur. (Hyland, 2019, p. 12)

The last two approaches seek to focus on the text and its substance matter. A focus on content is usually associated with the topic of the text (Hyland, 2019, p. 14). There are different strategies to choose from when deciding on a topic for a text, but generally, it is beneficial, if it is of interest to the learner. The content should relate to the learners' lives, situations and contexts because it allows them to use their prior knowledge. Potentially, this allows learners to feel competent and learners could even decide themselves what they want to write about to increase a sense of autonomy, and thus potentially motivation (Ryan & Deci, 2000). However, government guidelines often prescribe the topics that must be taught at school, leaving less room for autonomy.

In the same vein, a focus on genre goes "beyond content, grammar and composing processes to see writing as attempts to communicate with readers" (Hyland, 2019, p. 17). Providing a definition of genre has proven difficult and is beyond the scope of this thesis due to different constraints based on cultural and disciplinary contexts (Hallet, 2016, p. 32). On a general level, genres can be defined as "staged, culturally purposeful activities that [...] users of language draw on to get things done" (Paltridge, 2001, p. 11, as cited in Surkamp, 2017, p. 106). Hence, genres try to achieve certain goals of communication for a target-audience through specific structures and functions (Parkinson, 2017, p. 2). Genres are thus socially constructed and serve as habitualized forms of communication within a community (Hyland, 2007, p. 149; Martin & Rose, 2008, p. 18). Within genres, a distinction can be made between the macro and the micro level (Hallet, 2016, p. 40). The macro level considers the text as a whole, especially the sequencing of specific elements. For example, a news report consists of elements such as a headline, a nucleus or lead with basic information (i.e., *wh*-questions), lead development with further information, and a wrap up (Martin & Rose, 2008, pp. 76–77; Watanabe, 2017, p. 68; White, 1998, p. 266). On the micro level, specific linguistic patterns are found in specific genres, for instance, a news report that focuses on past events usually uses past tense forms (Hallet, 2016, p. 39). Commonly, these are also labeled as discourse functions, reflecting the focus of text function and emphasizing that the different writing approaches need to be considered in an integrative manner (see Halliday, 2014; Hyland, 2019, p. 18). As genres are a relevant communicative resource, learners must be enabled to produce texts in different genres, considering the respective purpose and audience. One approach to teaching genre writing in the classroom is an adaptation of the teaching-learning cycle, as depicted in Figure 15:

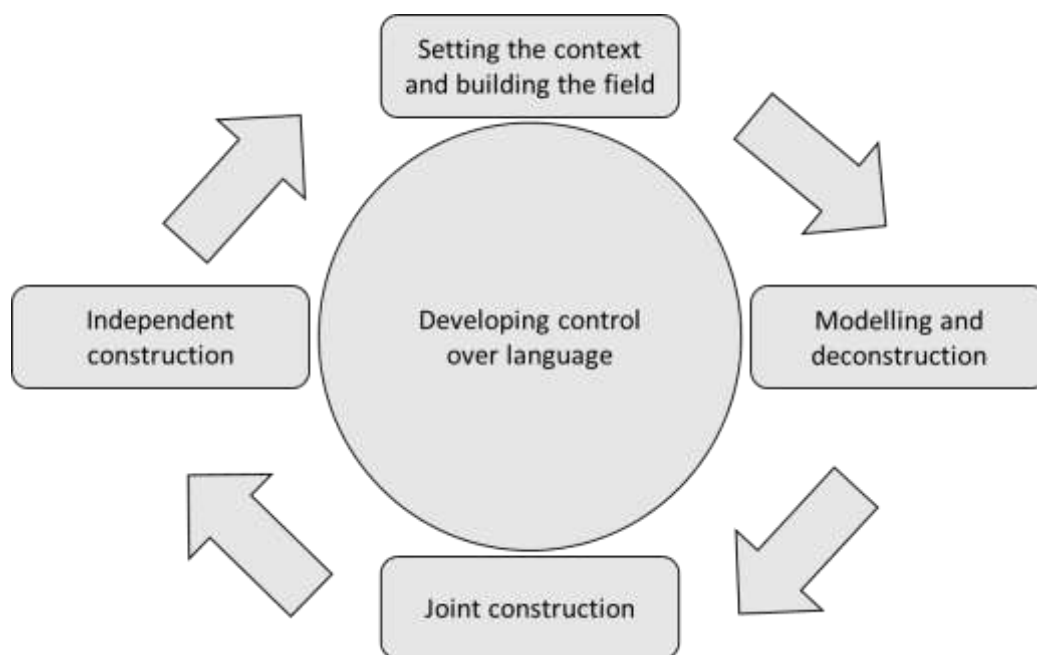


Figure 15: *Teaching-learning cycle for teaching genre writing (Hyland, 2019, p. 20).*

After setting the context and activating prior knowledge, a model text of the genre is discussed before the genre is constructed, first jointly, and finally, independently (Hyland, 2019, p. 20). A critical aspect which should be avoided is viewing genres as strict schemata or recipes, as writing within a genre still leaves space for creative expression (Hyland, 2019, p. 21).

The process and genre foci can most overtly be linked to “learning to write” (Manchón, 2016, p. 139) approaches, as they focus on how learners can be guided in the process of improving their writing skills. Consequently, these approaches also prevail in writing classrooms (Hyland, 2019, p. 22). A combination of both the process and genre approach is a suitable as “the strengths of one might complement the weakness of the other” (Hyland, 2019, p. 22). The genre approach allows a goal to be set for the text, providing clear guidance for the learner on specific characteristics for the text, while the process approach provides the structure for how to write the text. The other approaches then inform this procedure, e.g., by providing a normative and functional basis for genre-specific discourse functions. As this brief overview has shown, teaching writing is a complex matter that subsumes many different facets that a teacher could or is supposed to consider, in order to foster writing competence. The implementation is discussed from a statutory perspective in the next chapter.

6.2.3 Statutory Guidelines for Learning and Teaching Writing

A final relevant perspective on this issue is represented by the guidelines which not only indicate that writing is a required skill to focus on in the EFL classroom,

but also clearly state the skills and competencies that must be fostered⁷⁴. This chapter will first present the supranational guidelines in the form of the CEFR (Council of Europe, 2001, 2020), and then the national *BISTAs* (KMK, 2004, 2023a). The more specific guidelines on the level of the federal state (e.g., North Rhine-Westphalia or Rhineland-Palatine) are not included, as they are a further specification of the superior guidelines and do not offer drastically new approaches to writing, but rather specifications for the classroom. The discussion will also be restricted to the ninth grade of German comprehensive schools, as this is the targeted grade in the RobS. These students usually aim for the “Mittlerer Schulabschluss” after year 10, which is equivalent to B1 in the CEFR (KMK, 2004, p. 11, 2023a, p. 10)⁷⁵. Thus, the reported CEFR descriptors refer to B1 if not indicated otherwise.

In the CEFR, writing is discussed from the perspective of interaction (Council of Europe, 2020, pp. 82–83) and production (Council of Europe, 2020, p. 66). As the focus in the RobS is on production, only these relevant descriptors will be presented. For “overall written production” it states that learners “[can] write straightforward connected texts on a range of familiar subjects within their field of interest, by linking a series of shorter discrete elements into a linear sequence” (Council of Europe, 2020, p. 66). Two further relevant set of descriptors refer to B1 “creative writing”, which states that learners can, for example, “give a description of an event, a recent trip – real or imagined” (Council of Europe, 2020, p. 67). For the level of B1+⁷⁶, learners can, for example, “clearly signal chronological sequence in narrative text” (Council of Europe, 2020, p. 67). Lastly, the CEFR explicitly mentions “reports and essays”, with regard to which learners at the level of B1 “[can] produce very brief reports in a standard conventionalised format, which pass on routine factual information and state reasons for actions” and “[can] present a topic in a short report or poster, using photographs and short blocks of text” (Council of Europe, 2020, p. 68). The CEFR also states B1+ descriptors with regard to which learners

⁷⁴ This section will focus on writing. Naturally, writing is not isolated, especially the notion of linguistic competence, as outlined in the CEFR, factors into writing as well (Council of Europe, 2020, p. 130)

⁷⁵ The *BISTAs* state that after the ninth grade, students should achieve the CEFR level A2 by the first school leaving certificate (*Erster Schulabschluss*), and B1 is to be achieved by the secondary school certificate (*Mittlerer Schulabschluss*) after the 10th grade (KMK, 2004, 2023a). As learners in comprehensive schools tend to aim for the latter certificate, and considering the leap from A2 to B1 cannot be achieved within one school year and must be developed earlier, B1 is focused on as the target level in the RobS.

⁷⁶ These descriptors are usually situated between two CEFR levels (Council of Europe, 2001, p. 32). B1+ is thus situated between B1 and B2.

[can] produce short, simple essays on topics of interest.

[can] produce a text on a topical subject of personal interest, using simple language to list advantages and disadvantages, and give and justify their opinion.

[can] summarise, report and give their opinion about accumulated factual information on familiar routine and non-routine matters within their field with some confidence. (Council of Europe, 2020, p. 68)

Furthermore, the CEFR also lists specific production strategies which apply to both oral and written production. Specifically, it argues that L2/FL learners should master strategies that include “planning, execution, monitoring and repair” (Council of Europe, 2020, p. 68), resembling the cognitive models on the writing process described previously (see also Porsch, 2020, p. 69).

The national *BISTAs* represent the implementation of the CEFR on a national level. Coincidentally, the KMK revised these guidelines during the project which resulted in a few relevant reevaluations with regard to writing. Thus, both versions will be addressed. In the original version, learners were required to be able to present basic information in personal letters; formulate standardized letters and e-mails in an audience-oriented way; produce uncomplicated texts on different topics of their interest, e.g., reports, stories, descriptions; write short essays on generally relevant topics (KMK, 2004, p. 14). With regard to reports, it specifically states that for B1, learners need to be able to write brief reports in a standardized format, stating information and reasoning actions (KMK, 2004, p. 78). For B1+, it is added that learners should be able to write down notes if someone asks for information or explains a problem and write short reports on familiar topics, including the provision of reasons and taking a stance (KMK, 2004, p. 14). It is also explicitly stated as part of text and media competence that learners can independently navigate the writing process, including “drafting, writing, revising” (KMK, 2004, p. 17; translated by TJ). Thus, the original version makes an explicit reference to previous process models on writing, while also prescribing specific genres like the report that must be focused on.

The updated version describes that learners should be able to produce structurally uncomplicated, coherent texts on familiar and relevant topics; describe real or fictional events; write reports in a typical standardized format and provide reasons for actions; use (digital) means for the production and revision of texts; independently extend their linguistic repertoire and draw upon their multilingual resources for text production (KMK, 2023a, p. 16). It moves on from providing explicit genres and types of texts like TV shows due to the rapidly changing media landscape (Rossa & Wilden, 2024, p. 14). Also, the writing process is no longer explicitly referred to in the competence area of text and media competence. Rather, a broader argument is made that learners should be able to draw upon knowledge about text production, also arguing that other subjects should

contribute to text and media competence (KMK, 2023a, p. 21). After the relevant notions on writing are presented, it will be connected to feedback.

6.3 Synthesis: Feedback on Writing

After feedback and writing have been focused on individually, this section will provide a synthesis, i.e., argue why feedback on writing is a suitable issue for the RobS. First, a conceptual argument will be made based on the previously outlined concepts of writing and feedback. Second, an empirical perspective will provide insights into the impact that feedback can have on writing and the extent to which this is implemented in German schools.

On a conceptual level, there is a clear connection between the writing process and feedback. As displayed in the cognitive process models by Hayes and Flowers (1980), for example, the revision stage relies on feedback which initiates the editing process. This aspect of the writing process is, however, not exclusively focused on external feedback, as revision can naturally also happen internally. Also, it must not necessarily occur after a version of a text is finished, but can happen during planning or translating. However, Hyland also states that feedback “emphasizes a process of writing and rewriting where the text is seen as pointing forward to other texts the student will write” (2019, p. 171), highlighting the purpose of feedback in the sense of learning to write. While there are different sources of feedback, e.g., self-feedback, peer feedback, computer or AI-generated feedback, students do value the feedback from their teachers for the learning progress (Hyland, 2019, p. 172) and it is “the most influential [aspect] of a process writing class, and the point at which the teacher’s intervention is most obvious and perhaps most crucial” (Hyland, 2019, p. 12). Teacher feedback can thus be used by the learners to improve on their writing capabilities. What is more, the genre approach for writing provides a clear goal orientation for feedback as the target value or benchmark. By using model texts as exemplars, as described in the teaching-learning cycle, a possible standard is set and discussed with the learners as to what they should strive for. Feedback can thus assist learners in achieving this goal when it is included in the writing process, and actually implemented, e.g., through multiple drafting (Lee, 2017, p. 66).

While there is conceptual synergy, empirical data also supports the notion that teacher feedback is generally effective on writing and in the writing process. A main focus of feedback on L2/FL writing has been on written corrective feedback (WCF), e.g., written error correction aiming at language accuracy. Truscott initially argues against providing WCF, claiming that “grammar correction has no place in writing courses and should be abandoned” (1996, p. 328). He bases this claim on a review of studies which do not provide evidence that grammar correction leads to an improvement in accuracy (Truscott, 1996, pp. 329–333).

Since, there has been a debate around the issue, calling Truscott's contribution a "polemical article" (Lee, 2017, p. 66) and his thesis "premature and overly strong" (Ferris, 1999, p. 1). Ferris argues that an undifferentiated perspective on how error correction is conducted, overemphasis of particular findings and heterogeneous contexts of the primary studies show that his claim does not withstand appropriate scrutiny (Ferris, 1999, pp. 4–5). Subsequently, there has been a plethora of further studies that show that corrective feedback can have a positive impact on writing.

Several meta studies and analyses have discussed feedback on writing or in the writing process⁷⁷. A meta-study by Graham and Perin ($N = 123$ studies) finds quite variable effect sizes for the impact of feedback on writing quality for L1 learners due to heterogeneous feedback contexts, leading them to not "draw any reliable or meaningful conclusions" (2007, p. 464). A later meta-study by Graham and Harris ($N = 22$ studies) finds that, considering studies with experimental designs, feedback has a positive impact on writing quality in grades 2 to 12, with adult feedback ($ES = 0.87$) having a higher effect size than peer, self, and machine feedback (2018, p. 30; see also Graham et al., 2015). However, the study does not differentiate between L1 and L2 learners. Focusing on language accuracy of L2/FL learners, Truscott's meta study identified a negative effect of corrective feedback on accuracy, $d = -0.155$ (2007, p. 262; see also Scherer et al., 2024, p. 3). The small number of included studies ($N = 6$), however, does limit the validity of this study. The meta-analysis of $N = 25$ studies⁷⁸ by Biber et al. shows that feedback does have an effect on writing development for L1 and L2 learners of English (2011, p. 50). For L2 English learners, the effect ranges from $d = 0.92$ ⁷⁹ ($SD = 0.84$) in pre-/post-test designs, to $d = 0.66$ ($SD = 0.84$) in treatment/control group designs (Biber et al., 2011, p. 37). Possibly, these relatively large gains are due to the fact that these studies measure improvement mostly as accuracy. The effect of feedback is also larger for weaker L2 learners than for stronger learners (Biber et al., 2011, p. 38). Additionally, studies indicate that feedback by peers, self, or computers is more effective than from teachers (Biber et al., 2011, p. 39). However, this is not too surprising, considering that these studies also focus more on surface-level features as the quality indicator, i.e., spelling and grammar, which are also more easily identifiable by peers, self and

⁷⁷ This section will primarily focus on a general overview as not to pre-empt a detailed discussion on effective feedback practices in the writing context in the chapter focusing on the feedback framework, Chapter 7.

⁷⁸ In this analysis. 281 studies are excluded as they do not meet the criteria (Biber et al., 2011, pp. 24–25).

⁷⁹ Biber et al. calculate Cohen's d as the coefficient for their analysis (2011, p. 27).

computers than deep-level features, i.e., structure or cohesion, which might be more focused on by teachers, especially in the context of higher education. With regard to the mode, results are rather inconclusive as generally, learners seem to prefer oral feedback on their texts, but the effect sizes for L2 learners are not based on a large number of studies, complicating such a statement (Biber et al., 2011, pp. 39–40). In their re-analysis, Wisniewski et al. state that, overall, “[findings] by Biber et al. (2011) that written feedback is more effective than oral feedback could not be confirmed” (2020, p. 12). A meta-study by Scherer et al. ($N = 95$ studies; 253 effect sizes) differentiates between the effect of feedback on L1, L2 and FL learners with regard to surface- and deep-level feedback (Scherer et al., 2024, p. 9). They find that secondary school FL learners benefit regarding feedback on surface- ($g = 1.15$, $SE = 0.31$) and deep-level features ($g = 0.61$, $SE = 0.23$) with a moderate to large effect size in posttest measures (Scherer et al., 2024, pp. 11–15). Thus, it can be assumed that feedback does have a positive impact on writing outcome for FL or L2 learners, but, similar to the discussion on feedback overall, effect sizes are quite variable.

Against this backdrop, it is interesting that feedback practices in writing classes do not reflect the reported findings. Porsch investigated the extent to which EFL teachers implement process-oriented approaches to writing in grades eight to ten by asking $N = 1125$ learners and $N = 121$ teachers in 14 out of 16 German federal states (2010, pp. 58–59). Most of the writing activities in the EFL classroom are based on copying texts from the board or from projected transparencies, while actual writing happens mostly in cooperative situations (Porsch, 2010, p. 64). Strikingly, explicit revision activities are rare, as 44.1 % of teachers never require students to write drafts (which are to be revised) and only 17.1 % of learners indicate that they had to revise a text after receiving feedback, meaning that the vast majority of learners states that they have never actually used feedback to improve a text they have written (Porsch, 2010, p. 64). While teachers have a different perception, stating that they provide feedback regularly, it remains unclear if they include graded texts in class tests as a form of feedback in their answers (Porsch, 2010, p. 66). Learners state that they generate self-feedback by reading their texts and make adjustments, but do not receive external feedback (Porsch, 2010, p. 65). Especially written feedback, and to an even lesser extent, oral feedback occurs rarely when texts are written for practice purposes (Porsch, 2010, p. 66). These findings are generally confirmed in a study with $N = 166$ learners and $N = 6$ teachers conducted by Siekmann and Busse (2022, pp. 276–277). They find that planning activities, for 27.6 % of the learners, and revision activities, for 21 % of the learners, occur less than once a month (Siekmann & Busse, 2022, p. 280). Thus, while writing occurs regularly (i.e., at least two to three times a month), planning and revision activities do not occur as

frequently (Siekmann & Busse, 2022, pp. 280–281), and they “seem to play only a minor role in EFL classes” (2022, p. 285). The most frequent feedback agents are teachers, from whom 71 % of learners receive feedback on their texts at least once a month (Siekmann & Busse, 2022, p. 282). What is more, both teachers and students consider teacher feedback to be rather useful, but taking into account other findings from their project, the authors argue that “although students perceived feedback as effective [...], feedback seemed to have little effect on students’ writings” (Siekmann & Busse, 2022, p. 286; see also Siekmann et al., 2023). Reporting on the effects that a teacher intervention on feedback has, among other variables, on secondary EFL learners’ writing performance of argumentative texts, Siekmann et al. state that “[students’] ability to write well-structured and coherent texts [...] was relatively low at T1 [...] and did not improve over the course of 8 months despite the feedback intervention” (2023, p. 11), indicating that the teacher intervention did not necessarily translate into improved learner performance.

Coming to a conclusion, the relevance of focusing on feedback on writing for the RobS has been discussed by looking at the two components first individually, then integratively. Feedback can have a large impact on the learning process, but the impact is variable and depends on the way feedback is provided. Writing, which can be modelled as an iterative process that includes planning, translating and revising, is a relevant aspect of communicative competence and must be learned in the EFL classroom. Feedback can be regarded as a relevant aspect of the revision process and can have a positive impact on text quality on surface- and deep-level features. A genre approach in the writing classroom also provides a benchmark against which the feedback can be given. Actual practices in German schools, however, reveal that revision rarely occurs, despite teacher feedback occurring relatively regularly. While the teacher intervention in the third phase of teacher education by Siekmann et al. (2023) did not seem to improve text quality of learners, it could be argued that teachers’ capabilities to provide appropriate feedback should already be implemented in the first phase, i.e., at the university level, and the RobS could be one way to measure and assess future EFL teachers’ feedback competence on writing if a performance-oriented perspective is pursued. The next chapter will discuss the design of a framework, that can be used as a basis for constructing and developing a RobS, which assesses teachers’ competence on providing feedback on writing in the EFL context.

7 Developing a Framework for Feedback Competence on Writing in the EFL context⁸⁰

This chapter will present the framework which has been designed as a foundation for the development of the RobS. First, a brief reflection on theoretical research will provide a framework for the framework's conception. Second, existing instruments that are used to investigate feedback behavior are summarized as a first indication of relevant aspects to consider. Third, the framework is presented and discussed in its entirety. Last, the framework and especially its limitations are discussed.

7.1 Methodological Reflections on Theoretical Research

In order to design a framework, some methodological deliberations are required to ensure that the design process is soundly situated and contextualized. The approach taken to develop the framework can be classified as theoretical, as existing findings are compiled and combined (Legutke, 2022, p. 40). Following Legutke's definition of theoretical research that focuses on the development of constructs or models for foreign language education, it is necessarily interdisciplinary (Legutke, 2022, p. 41). This notion is also reflected in a post-theory approach as proposed by Delanoy et al. and defined as a stance that "aims to combine different perspectives" (2015, p. 8) rather than argue for the supremacy of a single theory. It is a "multi-perspective, almost dialogical approach to engaging with theories [which] allows for a more comprehensive understanding of the complexity of the examined phenomena" (Römhild, 2023, p. 263). Consequently, in order to develop a framework which seeks to define feedback competence on writing in the EFL context, different relevant theories and concepts from different influential disciplines must be considered and put into a dialogue. In that vein, a conscious reflection on potential biases is required as the declaration of *relevant* theories is, despite providing potentially appropriate reasons, at its core, rather subjective. As a means to counter this subjectivity, the framework was presented at different conferences to different groups to receive feedback, and implicitly included in the investigations by its operationalization as a scoring instrument, specifically in expertise interviews (see Chapter 15.3).

For the creation of the framework, boundaries with regard to content and form must be set. Concerning content, disciplines of general educational

⁸⁰ A prior version of this framework is published in Janzen, T., Vogelsang, C., & Rumlich, D. (2024). Feedbackkompetenz handlungsnah prüfen: Die Entwicklung einer rollenspielbasierten Simulation als Prüfungsformat für angehende Englischlehrkräfte. In D. Gerlach (Ed.), *Wissen, Können und Handeln von Fremdsprachenlehrpersonen: Professions- und Wissensforschung in den Fremdsprachendidaktiken* (pp. 221–234). Peter Lang.

sciences, specifically subdisciplines focusing on feedback, psychology, (L2/FL) writing and English language education/teaching English as a foreign language are considered relevant for the framework. Additionally, the statutory context must be considered as specific norms that teachers and teacher educators are required to adhere to must be included. As an initial guide, previous research that has already created research instruments can be used as a compass, in order to navigate the research landscape. With regard to form, the specifics of the RobS serve as a north star. As the test seeks to measure the performative aspect of competence, the framework must be designed for that purpose. Thus, only aspects that can be assessed by observed behavior in the simulated feedback conversation can reasonably be included in the framework. This inevitably leads to the exclusion of aspects that might be relevant otherwise, e.g., aspects of student uptake. Also, aspects that are written, i.e., WCF, are excluded, as only the verbal and physical behavior in the RobS is considered to be relevant observable behavior for the feedback conversation. Most importantly in the context of feedback, only the actions of the teacher are relevant for the framework as the learner is depicted by an actor in a standardized manner. Their actions are therefore not necessarily indicative of the quality of the feedback conversations. Against this methodological backdrop, the framework will be presented after a brief review of existing research instruments is conducted in the next chapter.

7.2 Instruments for the Evaluation of Feedback Performance

Several previous studies have designed research instruments that are used to explore feedback practices and performances. This section will provide a selective overview of these as they can be used as a first indication of the relevant facets for the framework. The focus will not be on the empirical results of these studies, but on the operationalization of feedback.

A useful first document for this framework is the coding manual designed by Kürzinger and Pohlmann-Rother (2019), which was developed in the context of the PERLE⁸¹ study (Lipowsky, 2013). In the study, primary school learners' development with regard to personality and learning is examined longitudinally by using multiple sources of data, including videotaped lessons of instruction in the subject of German (Lipowsky et al., 2013, p. 9). In this context, a coding manual that focuses on teacher feedback in the German writing classroom in primary schools was developed (Kürzinger & Pohlmann-Rother, 2019). It relies primarily on the model of feedback by Hattie and Timperley (2007) and adds specifics with regard to the actual feedback content by self-developed items as outlined

⁸¹ PERLE is the abbreviation of *Persönlichkeits und Lernentwicklung von Grundschulkindern* (Development of personality and learning in primary school children).

by Lotz (2016) (Kürzinger & Pohlmann-Rother, 2019, pp. 391–402). A similar study by van den Bergh et al. (2013) focuses on feedback in the primary history classroom. They conduct a literature review and identify the levels of feedback according to Hattie and Timperley (2007) as relevant, add a social level, and relate each level to *goal relatedness*, *nature* and *way* for their coding system (van den Bergh et al., 2013, p. 347). Buholzer et al. base their framework for the analysis of classroom videos in primary mathematical education on previous projects like the PERLE study (2020, p. 638). Consequently, they initially rely on the model of Hattie and Timperley (2007) and add strategical dimensions of *adaptivity*, *motivation & emotion*, and *reciprocity* (Buholzer et al., 2020, p. 642). Holstein et al., in a study that uses microteaching scenarios of feedback in which actual students between the ages of eleven and 13 act as SLs, investigate the difference in feedback performances between pre- and in-service teachers in the context of a writing task (2022). For analysis of the videotaped scenarios, they also rely on the model by Hattie and Timperley (2007) by the means of the operationalization of Kürzinger and Pohlmann-Rother (2019) and amend it for their specific research purpose (Holstein et al., 2022, p. 5).

Some inferences can be made based on these previous studies. Evidently, the model by Hattie and Timperley (2007) has been quite influential in conceptions of feedback research⁸². The contexts of most of the presented studies, however, differ from that found in this thesis regarding subject, age group, setting, and, most importantly, the normative nature that is inherent to summative assessment. The greatest discrepancy lies in the general research design, as these studies mostly analyze feedback occurring in action in classroom scenarios, while the RobS aims at an approximation of practice, especially in terms of a reduction of complexity and standardization for assessment purposes. The study by Holstein et al. (2022) is closest in terms of this setting, but includes a different age group and subject and, what is more, does not provide a subject-specific perspective on the described practice. Thus, the instruments cannot be adapted as is. Still, the broad strokes of the manuals serve well as a first resource and structuring device, in that they all consider the *function* of feedback, usually based on the model by Hattie and Timperley (2007), the *strategies* used to provide feedback, e.g., the way, or the notion of adaptivity (Buholzer et al., 2020, p. 642; Kürzinger & Pohlmann-Rother, 2019, p. 391; van den Bergh et al., 2013, p. 346), and the *topic* of feedback, which is specific to the context. While this three-pronged approach is not explicit in all these instruments, the different aspects can be subsumed under these basic dimensions and, consequently, the

⁸²See also Marty (p. 92); Ruelmann et al. (2023, p. 87) for further examples.

framework for this thesis will be structured according to the *Function*, *Strategy*, and *Topic* of feedback. These three basic dimensions serve mainly as a structuring device for the framework as they facilitate a systematic discussion of the individual dimensions. The framework is shown in its entirety in Table 1 and will be elaborated on in detail in the subsequent chapter.

7.3 The Framework

Table 1: Framework for a performance-oriented assessment of feedback competence on writing

Function		Strategy		Topic	
Aim	Level	Communication	L2/FL writing pedagogy	Diagnosis	Information
Feed up	Task	Dialogue orientation	Adaptive error correction	Language appropriateness	Detailedness
Feed back	Process	Adaptation of content	Focused error selection	Language correctness	Professional accuracy
Feed forward	Self-regulation	Adaptation of language	Target-orientation	Structure	Language accuracy
Positive feedback	Self	Descriptiveness	Process orientation	Content	
Self-evaluation		Specificity and Clarity	Prioritization of deep-level features	Process	

7.3.1 Functions of Feedback

The basic dimension of feedback *Function* describes the specific purpose of the feedback information or interaction, i.e., the specific aspect that is being targeted or needs improvement. It includes the aims of the subdimensions and the levels of feedback. It is mostly based on the findings and the feedback model by Hattie and Timperley (2007).

7.3.1.1 Aim

The respective aims for feedback relate mostly the model of Hattie and Timperley (2007). Aspects are added which are not explicitly presented in the model and do not focus on the information, but on the learner and how they process the feedback, relating to aspects of feedback literacies.

Feed Up

This aspect of the framework is taken directly from Hattie and Timperley (2007, pp. 88–89). It seeks to answer the question of *Where am I going?*, i.e., provide the learners with information on the desired state. These goals can be described with different degrees of detail, ranging, for example, from naming the news report as a target-genre, to describing certain moves or discourse functions that constitute the genre of a news report. Considering that it is the goal of feedback to reduce a discrepancy between the actual and desired state, it is essential that the goal is well-defined, so that the gap can be expressed more exactly (Hattie & Timperley, 2007, p. 89). Goals do not have to be prescribed by the teacher but can be set by the learners individually or together with the teacher, so that learners feel more committed to achieving them (Hattie & Timperley, 2007, p. 89). The notion of goal setting is also found in other overviews that provide criteria for feedback practices, many referring back to Hattie and Timperley (2007) and other meta-analyses (Busse, 2019, p. 24; Grotjahn & Kleppin, 2017, p. 261; Hawe et al., 2008, p. 45; Kluger & DeNisi, 1996, p. 278; Kopp & Mandl, 2014a, p. 151; Lee, 2017, pp. 4–5; Narciss, 2014, 68, 2017, pp. 180–181; Schluer, 2022, p. 18; Siekmann et al., 2022, p. 271).

Feed Back

This aspect of the framework is taken directly from Hattie and Timperley (2007, p. 89). It seeks to answer the question of *How am I going?*, i.e., providing information to the learner about their actual state of performance in relation to a goal. Naturally, a preceding diagnosis is necessary to identify the actual state of the learner, using the goals as a benchmark. This includes not only information on the gap between the actual and desired state, but also on the aspects in which the learner has actually achieved the goals (see *Positive feedback* below).

Feed back is also often denoted with a process component, e.g., potentially comparing the performance in focus to a previous performance, allowing for an individual comparison and feedback on the individual progress (Voerman et al., 2012, p. 1108).

Feed Forward

This aspect of the framework is taken directly from Hattie and Timperley (2007, p. 90) and has subsequently been framed as the most-valued aspect of feedback by learners (Mandouit & Hattie, 2023, p. 6; Zierer, 2019, p. 7). It seeks to answer the question of *Where to next?*, i.e., what are the specific next steps that have to be taken in order to achieve a goal and improve performance. This includes

information to students and teachers that leads to an adaption of learning in the form of enhanced challenges, more self-regulation over the learning process, greater fluency and automaticity, more strategies and processes to work on the tasks, deeper understanding, and more information about what is and what is not understood. (Wisniewski et al., 2020, p. 2)

The provided information is supposed to “[enable] students to close the gap between current and desired performance, helping them improve their future performance” (Lee, 2017, p. 56) and reflects notions of feedback literacy, in that students need to be able to use this information to further their learning (Carless & Boud, 2018, p. 1315). Although many students consider this aspect as valuable for their learning progress, research indicates that this kind of feedback is rare in classrooms (Brooks et al., 2019, p. 22). Considering the process models of writing (Bereiter & Scardamalia, 1987; Hayes & Flowers, 1980), feed forward is essential to improve on drafts of a text and close a diagnosed gap (Ferris, 2013, p. 8; Lee, 2017, p. 15; Sadler, 1989, p. 121). The relevance of feed forward is also imminent when feedback is considered as formative assessment, i.e., used to support learners in the learning process and improve further learning (Black & Wiliam, 1998).

Positive feedback

A facet that is mostly excluded in the model by Hattie and Timperley (2007), is how feedback affects learners’ “motivation, efficacy, and engagement” (Mandouit & Hattie, 2023, p. 2). The aspect of *positive feedback* is therefore included as a specific form of feed back which describes the provision of “information about learners’ correct task fulfillment, the progress they have made or useful strategies they have applied” (Schluer, 2022, p. 19). Consequently, it describes aspects in which the learners’ actual state and desired state are identical, i.e., there is no need for improvement (Narciss, 2006, p. 16). Such information might be beneficial for learners’ self-efficacy, as it potentially allows for an “enactive

mastery experience" (Bandura, 1997, p. 80), i.e., a moment of success (Mory, 2004, p. 766). Self-efficacy, defined as the "learner's perception of how well he or she [sic!] can perform the learning tasks to achieve his or her [sic!] goals" (Mory, 2004, p. 766) is relevant, as "it influences how much effort a learner will invest in a task" (Mory, 2004, p. 766). Providing positive feedback could thus bolster self-efficacy, which might benefit learners' efforts for future tasks⁸³. Another potential argument for the inclusion of positive feedback can be made based on the self-determination theory of motivation by Ryan and Deci (2000). They postulate three basic needs which are to be fulfilled: autonomy, relatedness and competence. With regard to the latter, the authors state that "events such as positive feedback that lead to perceived competence⁸⁴ by supporting the competence need will enhance intrinsic motivation" (Deci & Ryan, 2011, p. 418)⁸⁵. Importantly, in order to have a positive impact on achievement, positive feedback should not relate to the person themselves, but rather to an actual accomplishment (Hattie & Timperley, 2007, pp. 96–97; Kopp & Mandl, 2014a, p. 154) (see also section on *Level* below). The necessity to provide learners with information on their achievement is also addressed in other overviews of describing 'good' feedback practices (Grotjahn & Kleppin, 2017, p. 261; Narciss, 2017, p. 183).

Self-evaluation

This aspect of *self-evaluation* is based on the interactive tutoring feedback (ITF) model by Narciss, which is based on the assumptions of systems theory (2006, pp. 69–70, 2017, p. 177)⁸⁶. In the model, two feedback loops that interact with each other are described. On the one hand, there is the external feedback loop, focusing on information on performance provided by an outside agent, like the teachers or peers (Narciss, 2017, p. 176). On the other hand, there is the internal feedback loop, focusing on feedback generated by the learners themselves (Narciss, 2017, p. 176). In the internal feedback loop, learners self-evaluate their performance and generate a subjective representation of the extent to which they believe to have achieved the goal, by comparing their performance to what they

⁸³ This causal chain is depicted in a simplified manner as, naturally, other aspects might have an impact on self-efficacy and self-efficacy is only one aspect contributing to effort.

⁸⁴ Perceived competence is also described as self-efficacy by Ryan and Deci (2000, p. 58).

⁸⁵ Deci and Ryan point to two further conditions, which shall not be ignored: "However, the positive feedback must be for an autonomously motivated activity [...] or within an autonomy-supportive context [...] in order for it to enhance intrinsic motivation" (2011, p. 418).

⁸⁶ In Narciss (2008, p. 128, 2014, p. 62), the mode is referred to as the "interactive, two-feedback-loop (ITFL) model."

consider the goal to be (Narciss, 2017, pp. 176–178). This internal representation is essential, as it serves as a backdrop against which the external feedback message is compared and is necessary to actually make use of external feedback (Narciss, 2006, p. 72, 2017, pp. 179–180). Learners' "self-assessment skills" (Narciss, 2017, p. 180) are especially relevant as part of self-regulated learning (Nicol & Macfarlane-Dick, 2006, p. 205). In a meta-study with $n = 19$ studies, Panadero et al. found that self-assessment can have a positive influence on the use of self-regulated learning strategies as well as self-efficacy (2017, p. 96). Furthermore, the generation of internal feedback might also facilitate learners' appreciation of negative feedback (Laudel & Narciss, 2023, p. 10). Consequently, learners need to be given the opportunity to explicitly self-evaluate their performance before external feedback is given so that they can better use teacher comments and can improve on their self-assessment skills:

Der interne Feedback-Loop ist von zentraler Bedeutung für selbstreguliertes Lernen und sollte daher vor dem externen Loop aktiviert werden. Zu Beginn eines Feedbackgesprächs sollten die Lernenden dazu angehalten werden, einerseits festzustellen, was sie mit Blick auf die Standards bereits gut erfüllt haben, andererseits mögliche Defizite zu erkennen. Erst nachdem die Lernenden ihr Lernergebnis oder ihren Lernprozess selbst mit Blick auf die Standards evaluiert und internes Feedback generiert haben, sollte also das externe Feedback angeboten bzw. in selbstregulierten Lehr-Lernszenarien aufgerufen werden. (Narciss, 2014, 77)

(The internal feedback loop is of central importance for self-regulated learning and should thus be activated before the external loop. At the beginning of a feedback conversation, the learners should be asked to on the one hand, state which aspects of the standards have been achieved, and, on the other hand, to diagnose possible deficits. Only after learners have evaluated their learning product or process with regard to the standards and generated internal feedback, should external feedback be offered or provided in self-regulated learning scenarios.)

Narciss states that learners should be encouraged to state aspects in which they have succeeded, as well as deficits (2014, 77). Only after that, the external feedback message should be offered and provided. This also aligns with the goals of student feedback literacy, which includes the ability to self-evaluate performances, i.e., *Making judgements*, and making use of external feedback, i.e., *Appreciating feedback processes* (Carless & Boud, 2018, p. 1317; see also Tai et al., 2018 and Chapter 6.1.2).

7.3.1.2 Level

The subdimension of level relates to the model of Hattie and Timperley and their findings on the respective impact that feedback on each level has on achievement, i.e., *Task, Process, Self-regulation* and *Self* (2007). To a greater or lesser

extent, the model has been used to observe, describe and evaluate feedback practices in previous studies (Buholzer et al., 2020; Holstein et al., 2022; Kürzinger & Pohlmann-Rother, 2019; van den Bergh et al., 2013).

Task

Feedback at the task-level considers the quality of a product that resulted from an activity, i.e., a text that was written following a prompt (Hattie & Timperley, 2007, p. 91). If the result of the task is a certain process, i.e., planning, writing and evaluation of a text, the process can also be the focus of task-level feedback, but it mostly relates to surface-level features (Brooks et al., 2019, p. 25). The task-level is the most addressed level in classrooms (Brooks et al., 2019, p. 22; Hattie & Timperley, 2007, p. 91), a finding which is corroborated in the results of the laboratory study by Holstein et al. for both pre- and in-service teachers (2022, p. 7). Feedback on the task-level is essential, because it is the basis for further and deeper feedback (Hattie, 2023, p. 321) and “[having] correct information is a pedestal on which the processing and self-regulation is effectively built” (Hattie & Timperley, 2007, p. 91). Task-level feedback can be described as the foundation for high-information feedback, which also includes feedback on *Process* and *Self-regulation*, and has a substantial effect size of $d = 0.99$ (Wisniewski et al., 2020, p. 11). However, task-level feedback on its own is only powerful “when learning content, the ideas or surface level knowledge” (Hattie & Clarke, 2019, p. 76).

Process

Feedback that is provided at the process-level focuses on the strategies that are used to complete a task and is thus “aimed at the processing of information that is necessary to understand or complete a certain task” (Wisniewski et al., 2020, p. 2). While being directed at specific tasks, process feedback can include “strategies for error detection” (Hattie & Timperley, 2007, p. 93), i.e., the identification of challenges in the writing process which lead to modifications to the process (Hattie et al., 2017, p. 304). It is part of high-information feedback and leads to deeper learning, if provided with feedback on the task (Hattie et al., 2017, p. 304; Hattie & Timperley, 2007, p. 93; Wisniewski et al., 2020, p. 12). However, process-level feedback is naturally more complex than task-level feedback, as it moves beyond the surface-level (Hattie, 2023, p. 321). Thus, it is more effective for familiar tasks and for learners who are more proficient (Hattie, 2023, p. 322; Hattie & Wollenschläger, 2014, p. 141).

Self-regulation

At the level of self-regulation, feedback is aimed at “the individual’s regulation of the strategies they are using to their performance” and how “to monitor the

use of strategies in the learning process" (Wisniewski et al., 2020, p. 2). Consequently, it is not directly related to a task and how it has been completed. It includes aspects such as "skills for self-evaluation, expanding effort in task engagement, or seeking further feedback information" (Hattie et al., 2017, p. 304), i.e., the active choice of seeking feedback can be described as a facet of self-regulation and thus feedback can be provided on it. It is also a potential facet of "high-information feedback" and thus quite influential with regard to student achievement (Wisniewski et al., 2020, p. 12). The effectiveness of self-regulation strategies themselves are well documented. Graham and Perin, for example, report that the teaching of self-regulation strategies (e.g., planning, writing, revising) has a large effect size of 1.14 to 1.47 with regard to writing quality (2007, p. 446). Therefore, feedback on these strategies should be included in instruction. However, it appears that feedback rarely occurs on this level (Brooks et al., 2019, p. 22; Holstein et al., 2022, p. 7). It should be noted that feedback on this level exceeds the complexity of the previously addressed levels and is more effective when learners are at an advanced stage (Brooks et al., 2019, pp. 26–27; Hattie, 2023, p. 323; Hattie & Clarke, 2019, p. 78).

Self

When feedback is given on the level of the self, it contains praise or criticism on a personal level and "usually contains little task-related information and is rarely converted into more engagement, commitment to the learning goals, enhanced self-efficacy, or understanding about the task" (Hattie & Timperley, 2007, p. 96). Praise in the form of generic comments, e.g., "You are a good student," might actually deflect from the actual task and is thus not effective in promoting learning (Hattie & Timperley, 2007, p. 96). Kluger and DeNisi report a small effect size of praise in feedback interventions of $d = 0.09$ (1996, p. 273), while Hattie reports an average of $d = 0.14$ (Hattie, 2023, p. 320). What is more, feedback on the self-level might even have a negative effect on achievement, even in the form of praise (Wisniewski et al., 2020, p. 12). Although especially younger learners appreciate the praise (Hattie & Timperley, 2007, p. 97), it seems likely that older learners might perceive praise "as an indication that the teacher perceived their ability to be low" (Hattie & Timperley, 2007, p. 97). With regard to motivation, Wisniewski et al. state that "effects of uninformative forms of feedback on motivation are low or even negative" (2020, p. 12). Consequently, feedback on that level should be avoided.

7.3.2 Strategies of Feedback

The basic dimension of *Strategy* focuses on the mode in which the feedback is provided. On the one hand, this concerns the communicative aspect, i.e., the

extent to which the information is presented appropriately to the learner by the teacher. On the other hand, this dimension also refers to relevant practices that are essential for effective feedback in the L2/FL writing classroom, i.e., deciding which aspects to provide feedback on, how relevant they are, and how that feedback is used.

7.3.2.1 Communication

Dialogue orientation

This aspect is based on a dialogic understanding of feedback and a constructive approach to learning (Aljaafreh & Lantolf, 1994; Carless, 2020; Carless & Boud, 2018; Carless & Winstone, 2020) (see also Chapter 6.1). Feedback should not be considered as a monologue in which the teacher provides information to a learner. Rather, the learner should be considered as an active (co-)constructor of their learning process. Thus, feedback situations should involve the learners constructively, i.e., by asking learners what they would like to focus on in the feedback conversation, allowing them to ask questions, or providing opportunities to explain certain decisions they made. This aspect, thus, is also reflective of the relational dimension of teacher feedback literacy (Carless & Winstone, 2020, pp. 155–156). Involving the learner in the conversation seems especially appropriate in the writing context. As illustrated in the knowledge-transforming model by Bereiter and Scardamalia (1987, pp. 12–13), the final written product does not allow for making inferences about the process itself (Schnell, 2018, p. 43) (see also Chapter 6.2.1). By taking a dialogic approach to a feedback conversation on a text, then, more insights can be gained into the writing process, enabling the feedback providers to include feedback on the writing process. What is more, engaging the learner in the conversation increases the likelihood of uptake as the information is not simply consumed passively. Furthermore, dialogue is also dependent on non-verbal aspects of communication. Thus, the body language and facial expressions should support the dialogue and indicate attentiveness to the learner, valuing their contributions⁸⁷.

Adaptation of content

This aspect is based on the notion of the usability of feedback. In order for learners to use the feedback, they need to understand the information that is provided to them on a content level (Hattie & Clarke, 2019, p. 51; Price et al., 2010, p. 279; Schluer, 2022, p. 42). Using Vygotskij's model of the ZPD, feedback should be situated above the level of what learners can do on their own, and

⁸⁷ The aspect of body language was added to the framework after the expert interviews, see Chapter 15.3.4.

below the level that is out of their reach, so, within the zone in which learners can further develop their competence, if appropriate scaffolding is provided (1978). Besides the information that is available to the teacher due to previous interactions, the learner text also provides information on the learner's competence level and teachers might provide feedback based on their text-related diagnosis. However, teachers should confirm that the input they have provided is actually understood by the learner by actively checking understanding (Mandouit, 2020, p. 15). This allows teachers to adapt their feedback to the learners by reformulating, providing different examples, etc., if learners indicate a lack of understanding.

Adaptation of language

This aspect also results from the goal of usability, but takes a genuinely EFL-related perspective, that of language. Naturally, EFL teachers serve as language models for their learners in the EFL classroom, supposedly providing comprehensible input, which learners can use to acquire the target language (Lightbown & Spada, 2013, p. 106). The normative principle of functional monolingualism accounts for this, describing that teachers should speak English as much as possible, and as little German as needed (Matz et al., n.d.; Schulentwicklung NRW, n.d.). As a consequence, EFL teachers are expected to have excellent competence in the target language (Freeman, 2017, p. 32). While, for a long time, the native-speaker ideal has prevailed in the profession, claiming that EFL teachers should adhere to a certain variety as a norm of English as a target, there are doubts that such an ideal has an impact on the teaching quality or student achievement (Rütli-Joy, pp. 25–26). This is not to say that language accuracy should be disregarded (see the aspect of *Language accuracy*), but that language is also a tool that EFL teachers need to be able to use appropriately and in an adaptive manner. A concept that seeks to address this issue is teacher language competence, which also includes the communicative nature, i.e., the notion of actually transmitting the desired information to the learner (Rütli-Joy, p. 28). Kuster et al. have, for that purpose, created a set of profession-related language competence profiles, in which they describe “the communication-oriented linguistic skills required by teachers of foreign languages at lower secondary level (‘Sekundarstufe I’) in order to practice their profession” (Kuster et al., 2014, p. 2). As an example, they state that “[in] the target language, the teacher is able to give oral feedback on a learner’s performance” (Kuster et al., 2014, p. 14).

In the context of feedback, however, especially in individual situations that do not directly take place in an instructive setting, the primary goal is not to function as a language role model, but rather to provide information to the learner in an appropriate and understandable manner, as discussed above. This

could be done in the target language, if the individual learner's linguistic competence is sufficient, adhering to the principle of functional monolingualism. As it can be assumed, however, that learners in the EFL classroom, especially in the lower secondary level have not (yet) achieved the language level of their teachers, teachers still need to adapt their output appropriately and reduce linguistic complexity (Bleichenbacher et al., 2017, p. 17). EFL teachers should thus, on the one hand, be able to theoretically perform classroom activities in the target language, but on the other hand, be able to adapt the level of English output they produce to satisfy their learners' needs. As a final consequence, as understandability is key for learners being able to use feedback, this would require teachers to actually switch to the learner's L1, if possible.

Descriptiveness

This aspect focuses on the tone of the feedback, specifically with regard to affect (Lotz, 2016, p. 259). When information is provided on the current state of the learner, it should be done in a descriptive, nonevaluative manner (Kopp & Mandl, 2014a, p. 158; Shute, 2008, p. 153). This is not to say that statements made by the learner cannot be classified as false, rather, that the presented discrepancy is not connotated with a value judgement. For example, a more appropriate statement is: "You misspelled this word here", while a statement like: "It is bad that you misspelled it," may be inappropriate. The foundation of this subdimension lies within Rosenberg's concept of nonviolent communication (Rosenberg & Holler, 2016). This concept provides a coherent argument and guidelines for communication and has also been used in other conceptualizations of feedback (Hümbert-Schnurr et al., 2019, p. 851). One notion that Rosenberg highlights is the distinction between description and evaluation (Rosenberg & Holler, 2016, p. 38). A differentiation between both is necessary, so that the addressee is less likely to perceive the feedback information as criticism (Rosenberg & Holler, 2016, p. 38). As the context of the RobS is situated within the learning process, and thus of a formative and not summative-evaluative nature (see Chapter 4.4.1) the discrepancies should be presented in a descriptive manner so that the learners can use the feedback appropriately.

Specificity and Clarity

Another relevant aspect focuses on the precise presentation of feedback information, i.e., its specificity and clarity. Feedback should be specific, i.e., precisely locating the issue and stating the goal to be strived for (Hattie & Timperley, 2007, pp. 85–86; Kluger & DeNisi, 1996, p. 260; Shute, 2008, p. 167, p. 174). The message should, for example, directly refer to an error or error pattern in a text and not be provided vaguely (Ferris, 2013, p. 8). This allows learners to distinctly

identify an issue and improve on it. Concerning clarity, the learners should be provided with unambiguous statements to avoid confusion or uncertainty (Hattie & Timperley, 2007, p. 104). Clarity, or unambiguity, is especially important regarding actions that learners need to take to improve (i.e., *feed forward*). If learners receive ambiguous guidance, their subsequent actions might not lead to the potential improvement. Rosenberg reflects the notion of specificity and clarity as he describes the necessity of formulating “concrete actions” (Rosenberg & Holler, 2016, p. 77) and the use of “positive action language” (Rosenberg & Holler, 2016, p. 75). When stating requests, i.e., in the form of actions that are supposed to be taken to improve on a text, these requests should contain a clear action and be described in a way that informs learners on what to do, rather than on what to avoid. For example, a statement like: “You need to revise the text by checking each verb with regard to the correct tense” is actionable, concrete and positively formulated, while a statement like: “You need to avoid using the present simple” is not.

7.3.2.2 L2/FL Writing Pedagogy

Adaptive error correction

Similar to *Adaptation to content*, which focuses on checking understanding after providing feedback, adaptive error correction seeks to describe teachers’ consciousness towards meeting the learners where they are, in terms of their current knowledge. This aspect, however, considers the approach that teachers take concerning errors in the learner text. Aljaafreh and Lantolf propose a graduated approach to error correction, based on Vygotskijan socio-cultural theory:

Graduation and contingency work in tandem in such a way that the expert, together with the novice, tries to discover the ZPD of the novice in order to determine if help is required and if it is, to jointly work out the appropriate level at which to provide it. The process is thus one of continuous assessment of the novice's needs and abilities and the tailoring of help to those conditions. This process can be accomplished only through the collaborative interaction of the expert and the novice [...]. Discovering the potential developmental level of the novice and providing appropriate help accordingly is at its core a dialogic activity that unfolds between more capable and less capable individuals. Dialogue is an essential component of Vygotskian theory [...] and hence of the ZPD. Without dialogic negotiation, it is virtually impossible to discover the novice's ZPD. (1994, p. 468)

Following this approach, teachers should first allow learners to self-correct after being made aware of potential errors. If learners are unable to self-correct, additional scaffolding should be provided, which could assist learners in self-correcting. Only in a last step should learners be provided with a correct solution. Another approach to frame adaptivity is through the different approaches to

(written) corrective feedback. For instance, Ferris et al. describe *indirect* corrective feedback as a situation in which “the error is called to the writer’s attention, but the correct form is not given” (Ferris et al., 2013, p. 309), while *direct* feedback includes providing the correct form (Ferris, 2013, p. 8). Another distinction can be made according to the degree to which information is provided. *Explicit* corrective feedback includes “labels, codes, or other metalinguistic explanation” (Ferris et al., 2013, p. 309), while unlabeled corrective feedback does not provide additional information (Ferris et al., 2013, p. 309). Especially for L2 learners of English, explicit corrective feedback seems to be beneficial (Ferris et al., 2013, p. 309). While these concepts derive from WCF, they can be applied to the oral context and the idea of graduated feedback. The scaffolding that teachers provide, if initial learner self-correction is not successful, could thus be indirect and explicit first, e.g., providing the appropriate amount of information for the individuals learner’s ZPD, before error correction is more direct. In a small study with $n = 15$ L2 learners of English aged between 18 and 29 years, focusing on oral feedback on texts, Erlam et al. found that “Graduated Feedback [sic!] was effective in enabling learners to self-correct. In contrast, Explicit Correction [sic!] was less effective in this respect and also the learners did not consistently uptake the correction provided by the teacher” (Erlam et al., 2013, p. 266), indicating that a graduated approach to feedback might be beneficial for learners. At the end, only a dialogic-explorative approach allows teachers to provide learners with the necessary and appropriate extent of corrective information.

Focused error selection

When providing feedback on a text, teachers need to be able to focus and thus make a conscious selection of aspects to provide feedback on. While unfocused or comprehensive error feedback, i.e., addressing all the errors found in a text, might be beneficial for L2/FL learners’ writing accuracy, especially in terms of grammar (Mao & Lee, 2020, pp. 7–8; van Beuningen et al., 2012), there is a general consensus that this is not conducive to learning (Ellis, 2009, pp. 6–7; Ferris, 2013, p. 8; Hyland, 2019, p. 199; Lee, 2017, p. 75, 2019; Nicol & Macfarlane-Dick, 2006, p. 210; Shute, 2008, p. 159). One theory that is used to argue for focused error correction is that of cognitive load (Siekman et al., 2022, p. 280; Sweller, 1988, 2011). Possibly, unfocused feedback contains too much information, i.e., the extraneous cognitive load is too high for the working memory, leading to an ‘overload’ of the overall learning process and the possibility that the feedback is not used (Siekman et al., 2022, p. 280; Sweller, 2011). Hence, in order to not overwhelm learners, teachers should select a limited amount of errors to focus on.

While consensus exists on the matter that feedback should be provided on selected aspects, a certain number of pieces of information is still unsettled, as Ferris points out for WCF:

From a L2 writing perspective, it seems fair to say that 1 or 2 features are too few and 15–20 features are probably too many. However, it has not been established what the most appropriate number of written CF categories should be. These categories might change over time; for instance, an instructor or researcher could focus on several categories early in a course or a study but then emphasize different features later. (2010, p. 196)

Other researchers propose providing feedback on a maximum of three different aspects or skills in the context of screencast feedback (Whitehurst, n.d.; as cited in Schluer, 2022, p. 21). Naturally, the appropriate amount of feedback is also strongly context-dependent as the task complexity, the learner, the allocated time, the mode in which the feedback is provided (written or oral), etc. all influence how much information the learner can actually make use of. Moreover, Lee, for example, argues that the number of items should also be increased with learners' proficiency:

[Instead] of targeting only one or two grammar items, teachers should provide WCF on a range of error types based on the genre/writing task as well as students' needs, perhaps gradually increasing the number of target grammar items as their students' progress in written accuracy. (2019, p. 532)

In the end, the amount of feedback needs to be appropriate to the context and it must be evident that a reduced number of aspects is focused on. An additional challenge in specifying the amount, lies in the definition of feature or aspect. For example, similar errors could be grouped as patterns in a feedback conversation, thus representing one aspect. However, the same features could also be addressed individually, representing several different pieces of information. Identifying error patterns, for that matter, is also considered good feedback practice when setting a focus because "[selective] error feedback on several patterns of error is more beneficial than comprehensive error correction" (Ferris, 2013, p. 8). Identifying and grouping errors into patterns indicates areas in which the learner has a systematic misunderstanding, other than just a single mistake, effectively leading to a focus and reducing the likelihood of overload (Busse, 2019, p. 25; Lee, 2017, p. 76; Schluer, 2022, p. 204; Siekmann et al., 2022, p. 280).

Besides frequency, the kind of information also needs to be reflected upon when considering focus. A focus on several, but not too many, different aspects of a text is preferable to focusing on one specific area (Ferris, 2013, p. 8; Lee, 2017, p. 532). A useful distinction can be made between surface-level features of a text (e.g., accuracy in the form of spelling and grammar) and deep-level

features (e.g., content, structure and cohesion of a text) (Scherer et al., 2024, p. 2). With regard to their meta-study, Biber et al. conclude that “[a] combined focus on content + form results in greater gains in writing development than an exclusive focus on form. [...] If we consider only L2-English learners, the differential influence of a focus on content + form versus form becomes even greater” (Biber et al., 2011, p. 50). In a later meta-analysis, Scherer et al. addresses a limitation of the analysis by Biber et al. in that they differentiate between L2 and FL Learners:

L2 learners typically comprise students acquiring the language of instruction after their L1 with the language of instruction being spoken in the immediate environment (e.g., first-generation migrants at school but also international students). In contrast, FL learners study a foreign language in their home country; that is, the FL is not spoken in the direct environment (e.g., learning English in non-English speaking countries), and learners are mainly exposed to the FL in the classroom [...]. (2024, p. 2)

Scherer et al. find that, overall, FL learners seem to benefit more from surface-level feedback than L2 learners, if the measure is focused on surface-level features, while a focus on surface-level aspects has a negative impact on deep-level features (2024, pp. 12–13). Both, L2 and FL learners do not seem to benefit from feedback on deep-level features, possibly “due to reduced linguistic resources” (Scherer et al., 2024, p. 13). A combination of both leads to an improvement on both levels with a moderate positive effect size, with deep-level features having a slightly larger effect (Scherer et al., 2024, p. 13). Consequently, teachers should strike an appropriate balance and provide feedback on both surface- and deep-level features of the text.

Target orientation

When deciding on which aspects to focus on in the feedback conversation, aspects that relate to the overall learning objectives are essential (Hattie & Timperley, 2007, p. 104; Lee, 2017, p. 76). While L2/FL learners may make different kinds of errors when writing a text, a portion of those errors may relate directly to the learning objectives, while others do not. For instance, if the goal is to write a news report on a past event, an essential objective is to use the correct tense for the genre, e.g., past tense. Within that text, other errors or mistakes may occur which are not directly relevant for the learning objective, like spelling mistakes. Considering the relevance of focus so as not to overwhelm the learners, teachers should prioritize those aspects in the text which are relevant for the learning objectives, because “[if] irrelevant variables are measured and their values are fed back to a system (e.g., a learner), this information will be useless if not harmful” (Narciss, 2017, p. 182).

Process orientation

As has been outlined in detail in Chapter 6.2, writing is not a linear process but an iterative one, consisting of multiple drafts and cycles of feedback on a text, before settling on a final version (Ferris, 2013, p. 8; Hyland, 2019, p. 11; Lee, 2017, pp. 14–15; Porsch, 2010, p. 57). For feedback to work, it has to be applied by the learner on a subsequent draft, e.g., a form of revision has to take place, reflecting the notion of taking action in student feedback literacy (Carless & Boud, 2018, p. 1318). Teachers should explicitly provide space in the classroom to do so, as to legitimize revision as a valid aspect of writing competence and not only focus on the product, thus also relating to the design dimension of teacher feedback literacy (Carless & Winstone, 2020, pp. 154–155). This aspect is, thus, relevant, as it speaks to teachers' general understanding of the value of feedback, also because errors should be understood as learning opportunities which can be accessed and addressed by appropriate feedback (Siekman et al., 2022, p. 280).

Prioritization of deep-level features

While a balance between surface- and deep-level features is advocated for, there is an argument made that, in the context of multiple drafting, especially in earlier drafts, deep-level features should be prioritized and addressed first (Ferris, 2013, p. 8; Lee, 2017, pp. 76–77; Zamel, 1985, p. 96). If a multiple-drafting approach is employed with differentiated foci, learners should be made aware of it so as not to be confused by the feedback (Ferris, 1995, p. 42). Interestingly, in a study with small groups ($n = 9 - 13$) of Japanese university students, Ashwell compared the effect that different feedback patterns have, i.e., a content focus on the first draft, a focus on accuracy on the second draft vs. the reverse pattern (2000). No significant differences with regard to the patterns in which participants receive feedback are found, using the respective facet (accuracy and content scores) as a measure. While strict differentiation of focus between drafts does not seem to be necessary, feedback should, nonetheless, prioritize deep-level features. A content-first approach allows learners to first create a text that aligns with the content- and structure-related goals and it might reduce writing inhibitions, if learners are made aware of that before the writing process starts. It thus emphasizes a general attitude towards the important facets of a text if deep-level features are prioritized. While revising the first draft, using the provided feedback on content and structure, learners might themselves find inaccuracies and self-correct. Also, errors that occur in first drafts might no longer occur in subsequent drafts, reducing the need for early corrections. Considering the required time investment for teachers, surface-level feedback and language accuracy is optimally focused on after the content and structure of a text are settled on.

7.3.3 Topics of Feedback

This dimension takes a closer look into the actual feedback information that teachers provide to the learners. This relates to the diagnosis of text dimension-related aspects, the extent to which the information is provided, as well as its accuracy.

7.3.3.1 Diagnosis

Teachers should provide feedback on “a range of issues, including content, organization, language, mechanics, and style” (Ferris, 2013, p. 8). In order to appropriately describe this range for the EFL context, the differentiation between surface- and deep-level features of a text offers a first indication of aspects. However, each level itself covers several different aspects of a text, e.g., content and structure are usually subsumed as deep-level features, and grammar and spelling are associated with surface-level features.

An influential and more detailed model for text analysis is the *Zürcher Textanalyseraster* (Zurich text analysis instrument; ZTA) (Nussbaumer, 1991, pp. 303–305), which, although being designed for the L1 German language classroom, has also been partially adapted for research in the EFL context (Keller, 2013, p. 104). The ZTA consists of three parts: the *zero part*, focusing on the description of relevant variables; the A-part, focusing on language accuracy and error correction; and the B-part, focusing on the appropriateness (Nussbaumer & Sieber, 1994, pp. 156–168). The ZTA has been tested, evaluated and optimized in a study with German L1 learners, but is primarily designed for research purposes (Hanser et al., 1994; Nussbaumer & Sieber, 1994, pp. 151–152). Another “widely used analytic [scale] in ESL” (Weigle, 2002, p. 115; see also Keller, 2013, p. 111) is the ESL composition scale by Jacobs et al. (1981). It is quite similar to the ZTA, describing the criteria content, organization, language use and mechanics, but adds an extra dimension of vocabulary, focusing on the appropriate and idiomatic use of lexis (Weigle, 2002, p. 116).

For this framework, the dimensions described by Becker-Mrotzek (2014) serve as a foundation. The model is designed for the L1 German language classroom and is based on previous work with Böttcher (2020, pp. 128–132; 2009, pp. 53–56) and the scoring instrument by Baurmann (2017, p. 136) (2014, pp. 507–509). In itself, the dimensions are derived from the ZTA but operationalized and reduced in its complexity for the evaluation of learner texts (Becker-Mrotzek, 2014, pp. 507–509). The model describes five dimensions, each with two to four different criteria that further describe the dimension, totaling twelve criteria. While the dimensions are more generic, the criteria need to be specified to the specific task and genre (Becker-Mrotzek, 2014, p. 508; Becker-Mrotzek &

Böttcher, 2020, p. 128). The five dimensions are “‘language correctness’⁸⁸, ‘language appropriateness’, ‘content’, ‘structure’ and ‘writing process’” (Becker-Mrotzek, 2014, p. 508; translated by TJ). Although similar in its dimensions to the ESL composition scale, this model is favored for two reasons. First, it includes the notion of the writing process, which is essential for developing writing competence. Second, while the ESL composition scale explicitly addresses the criterion of vocabulary, lexis should not be separated from grammar considering, for example, the lexical approach by Lewis (2009). Rather, the focus in the EFL classroom is supposed to be on lexical chunks, i.e., “grammaticalized lexis” (Lewis, 2009, p. 89; see also Hutz, 2018, p. 142). The framework by Becker-Mrotzek allows for that more integrated view, as the criterion of lexico-grammar can be subsumed in the dimension of *language appropriateness*. For better orientation and cohesion within the framework, the five dimensions will be related to the respective level of text, i.e., surface- or deep-level, and referred to as aspects.

The first aspect *Surface: Language correctness* covers the mechanical notions of language. This relates to notions of spelling/orthography, punctuation but also the correctness of morphological and syntactical constructions (Becker-Mrotzek, 2014, p. 508). For instance, learners might erroneously use **wined*⁸⁹ as a past simple form of *to win*, while the correct lexical item is *won*. Also, considering the EFL context, this dimension also extends to the correct use of lexico-grammar, in that the lexical chunks that are used convey the intended meaning. For example, a common error among L2/FL learners with German as their L1 is the use of *to become*, due to the false cognate assumption with the German *bekommen*, which translates as *to get*. The dimension thus considers context-independent errors that occur in the text. Feedback on language correctness is reflective of an approach to writing that focuses on language structures (Hyland, 2019, p. 4).

The second aspect *Surface: Language appropriateness* seeks to evaluate whether learners used language that is appropriate for the required genre and context. It includes the use of lexico-grammar and syntax with regard to the specific genre, context and audience of the text (Becker-Mrotzek, 2014, p. 508). The aspect is reminiscent of a writing approach that focuses on text functions and genre on a micro level (Hyland, 2019, 5, 19). For example, the chunk: “She was a great president.” is a grammatically correct statement, but might not be

⁸⁸ In order to better differentiate between the two categories that consider the mechanics of language, this text-related diagnostic dimension is coined *language correctness*, while teacher’s language capabilities are framed as *language accuracy*.

⁸⁹ Erroneous text fragments are indicated with an asterisk (*).

appropriate to use in a news report, which supposedly takes a neutral approach when reporting on events. Such a value statement can thus be considered inappropriate in that context. Furthermore, there are genre-specific conventions, which prescribe specific linguistic norms for texts. News reports that describe past events, for instance, should use the appropriate past tense forms that indicate the manner of time.

The third aspect *Deep: Structure* considers the extent to which the text follows the required pattern, i.e., genre conventions. This includes the appropriate use and sequencing of paragraphs, i.e., the sequence of genre-specific moves (like introduction, main part, conclusion) (Becker-Mrotzek, 2014, p. 508), and reflects the focus of the genre approach to writing from a macro perspective (Hyland, 2019, p. 19). For example, a headline is a relevant structural element of a news report, providing the reader with a first reference to the text's topic and should thus be addressed in a feedback conversation, if missing. Additionally, this dimension includes the extent to which a) the text coherently covers and develops the topic and b) the reader is guided through the text (Becker-Mrotzek, 2014, p. 508).

The fourth aspect is *Deep: Content* and addresses whether the text includes appropriate and sufficient information to fulfill the task requirements (Becker-Mrotzek, 2014, p. 508). This dimension is reflective of the content approach to writing, as described by Hyland (2019, p. 14). If a task, for instance, lists several pieces of information relating to a (fictional) past event, they are to be incorporated in the news report by the learner. If a relevant piece of information is missing, learners should be informed about it. As news reports are multimodal texts, this extends to the use of images (Jewitt & Kress, 2010).

The fifth aspect is that of *Deep: Process*, providing criteria to evaluate the extent to which the learner follows the writing process (Becker-Mrotzek, 2014, p. 508). It includes the extent to which revision marks are evident in the text and if the learner has explored new stylistic means of expression. For instance, a learner that has tried to use new syntactical structures can be praised for their risk, even though improvement is still necessary.

7.3.3.2 Information

After detailing the different aspects of the text that should be addressed, this subdimension focuses on how the information is supposed to be provided.

Detailedness

The detailedness of information that learners are confronted with is often differentiated on three levels. *Knowledge of Result* (KR) contains information only on whether the solution is correct or false, while *Knowledge of Correct Results* (KCR)

also includes information on the correct solution, if the previous solution was false (Lotz et al., 2013, p. 133; Narciss, 2008, p. 132). *Elaborated Feedback* (EF) includes any “additional information besides KR or KCR” (Narciss, 2008, p. 132; see also Kulhavy & Stock, 1989, p. 285). The exact nature of the additional information relates “to a large if not fuzzy set of meanings” (Lotz, 2016, p. 133), and is thus not well defined. For example, EF includes “reasons of (in-)correctness of a response [...], explanations about the task, suggestions, hints or strategic advice [...]” (Schluer, 2022, p. 21). Findings of meta-analyses indicate that the “[type] of feedback was also strongly related to effect size. When feedback merely indicated that a response was correct or incorrect, it resulted in a lower effect than when the feedback in some way informed the learner of the correct answer” (Bangert-Drowns et al., 1991, p. 232). Regarding elaborative feedback, the findings are rather inconclusive. While some analyses do not find that the amount of information influences the effects of feedback (e. g. Bangert-Drowns et al., 1991, p. 234; see Lotz, 2016, 133-132 for an overview), Wisniewski et al. argue that “[feedback] is more effective the more information it contains” (2020, p. 12). Arguably, many factors need to be considered here, especially the learner and their individual cognitive load, as has been discussed with regard to the general amount of information previously (see Chapter 7.3.2.2). Consequently, it seems more relevant that teachers do provide some sort of elaboration rather than considering the kind and amount of additional information.

Professional accuracy

A perspective which should not be ignored when developing a framework that essentially serves as a basis for assessment is the actual accuracy of information that is provided. One aspect is that of *professional accuracy*, describing the (in-)ability to provide correct and adequate information on non-language related aspects. It includes the extent to which information on deep-level features (i.e., genre-specifics, content of the text, and task demands) is stated correctly. For instance, it would be inaccurate to claim that a news report does not need a headline or to not correct but repeat or praise erroneous statements from the text. Moreover, this extends to the accuracy of learning strategies that are provided to the learners and the extent to which this inaccuracy might be harmful to their learning process. For example, while stating that learning vocabulary by simply memorizing bilingual vocabulary lists is *one potential* approach to take, it is not necessarily the optimal approach, considering the model of the mental lexicon (Hutz, 2018). Nonetheless, it is also not inherently harmful to the learning process, as it might lead to learning and other options are not explicitly excluded. Arguing that memorizing these bilingual lists is *the only* approach to learning could, however, be considered harmful to the learning process, as other

options that might be more aligned with the mental lexicon, focus on lexical chunks or have a more communicatively oriented foundation, and are thus potentially more beneficial, are explicitly excluded.

Language accuracy

The final aspect considers the accuracy concerning the English language. As English language role models, EFL teachers should have excellent language competence. While language complexity should be steered towards the learners' level of comprehension (see *Adaptation of language*), the language that is used should nonetheless be accurate. This refers to both language as the subject of conversation, as well as the means of communication. With regard to the former, this includes talking about language in the text, i.e., providing error corrections or explaining grammatical rules such as the formation of the past simple. Regarding the latter, this includes any instance, in which the pre-service EFL teacher speaks in English. Importantly, the accuracy of produced language must be evaluated in the context of the oral situation. Oral communication adheres to different norms than written communication (Goh & Burns, 2012; Levelt, 1995). Hence, features of spoken language which might be considered erroneous in written language, like minor mistakes (slips of the tongue), self-corrections or ellipsis, are natural to spoken language (Goh & Burns, 2012, pp. 86–94). Thus, the gravity of the error with regard to its potential impediment to learning and achieving the learning objectives should be considered in an evaluation. These aspects also include pronunciation, especially if it leads to potential complications in communication or shows lack of ability to formulate relevant phonemes for the English language, e.g., inadequate pronunciation of the <th> sounds /θ/ and /ð/.

7.4 Discussion and Limitations of the Framework

As with each framework, the one presented in this thesis has its own limitations. These derive primarily from the purpose that it serves; being created as the basis for a performance-oriented assessment format, specified as a role-play-based simulation, for PSETs.

The first constraint of the framework concerns its teacher-centered orientation. The framework focuses on the behavior of the teacher, describing the aim and content of the feedback that they provide, as well as the strategies they use. It does not include how the learner actually perceives and uses the feedback, which could be used as an indicator of the success of the conversation, because the learner is portrayed by an actor who is trained to behave in a standardized manner. As a result of the standardization, their behavior cannot be used as an indication of comprehension, as the actors' behavior is not necessarily a result

of teacher input, but is informed by the instructions they received. Thus, the framework only partially includes aspects of student and teacher feedback literacy. For instance, the relational dimension is addressed in *Dialogue orientation*, the pragmatical dimension in *Focused error selection*, and the design dimension in *Process orientation* (Carless & Winstone, 2020, pp. 154–157). Although a focus on the teacher already reduces the dimensions of the framework, some potentially relevant aspects are not included, especially timing, the structure of the conversation, and motivation.

Timing

Findings indicate that the effect of feedback is also dependent on the timing and the respective level (Bangert-Drowns et al., 1991, p. 226; Hattie & Timperley, 2007, p. 98). For instance, immediate error feedback might be more influential for feedback on the task- and surface-level, while delayed feedback seems to be more impactful for deeper levels, i.e., the process (Hattie & Timperley, 2007, p. 98). In the framework, the notion of timing has been excluded, because the RobS provides a specific time frame in which the feedback conversation happens.

Structure

Feedback conversations, like written texts, usually follow a certain generic structure, which is “partly conditioned by the modes and tools that are utilized” (Schluer, 2022, p. 24). One commonly known and widely used macro structure for feedback conversations is the so-called feedback sandwich, or feedback burger, in which the negative feedback is situated between two positive comments, with the last comment potentially describing a way forward (Schluer, 2022, p. 25). While such a structure allows for feedback providers to have a guideline that includes positive as well as negative feedback, “empirical or experimental evidence supporting the use of this method is lacking” (Henley & DiGennaro Reed, 2015, p. 322). Findings that exist indicate that the sequencing might not be of substantive relevance, while it is still more powerful than no feedback (Henley & DiGennaro Reed, 2015, pp. 331–332; Parkes et al., 2013, p. 406; Prochazka et al., 2020, p. 7). Potentially, learners that are used to this procedure might anticipate the negative comments and not appreciate the positive statements, considering them disingenuous (Parkes et al., 2013, p. 399; Schluer, 2022, p. 26). For the framework, a strict macro structure is avoided. Firstly, the empirical findings on the effectiveness of specific structures, such as the feedback sandwich, is not considered sufficiently sound for inclusion. Secondly, the situation of the RobS does not allow for the teacher to have prior knowledge about routinely implemented feedback practices, thus complicating

their evaluation of the appropriateness of such structures. Thirdly, the RobS is a dynamic situation in which the participants are not fully in control, as the standardized learner is not a passive listener, but an active participant who engages and potentially interrupts the dialogue, consciously steering the conversation in a specific direction. Hence, it could be considered unfair to include a fixed structure in the framework and consequently in the assessment instrument, knowing that participants might not be able to fulfill this criterion. Nonetheless, there are other structural elements which are considered important for a feedback message that are not included in the framework. For example, a personal greeting and address, a reference to the learner's progress, explaining the purpose of the conversation are elements that are often advocated for (see Schluer, 2022, p. 24 for an overview). These elements are not included for individual reasons. The situation of the specific RobS takes place after the lesson, thus a personal greeting has probably already occurred and is thus not obligatory (see Chapter 8). A reference to the learner's progress is complicated due to the fact that in this approximation of practice, the information on the learner is sparse and no previous work of the learner is provided for time constraints, making actual progress feedback difficult (Voerman et al., 2012). The purpose of the conversation is also not considered, as it is the learner in this situation who asks for feedback, making such a statement superfluous (see Chapter 8 and Appendix A for more information on the specific context of the RobS and material).

Motivation

Motivation is obviously a relevant aspect of feedback (Narciss, 2006, p. 32) and has been considered partially in the framework in the form of *Positive feedback* (see Chapter 7.3.1.1). Considering that motivation can be defined as "an internal state that arouses, directs, and maintains behavior" (Woolfolk, 2010, p. 376), *Positive feedback* might positively influence the internal state, if related to the task, process or self-regulation, while still relating to the text, which is the focus of the RobS. Other potential sources that might influence motivation are not necessarily related to the text or the task, like general interest in the subject or topic, as well as issues beyond the scope of the classroom. While in actual practice these issues certainly contribute to learners' motivation in the classroom and influence their behavior, their inclusion in the framework might shift the focus away from the text and towards a general consultation of the learner's situation. As the simulation is an "approximation of practice" (Grossman, Compton, et al., 2009, p. 2077), it allows for the reduction of complexity and placing of certain emphases. Therefore, aspects of motivation that are beyond the text are not included in the framework.

While a lot of care has been put into the design of the framework, and some of the excluded dimensions detailed and their exclusion reasoned, it is possible, if not likely, that potentially relevant aspects have not been considered or included. Consequently, the framework is of a preliminary nature and can function as a starting point for further discussion on the topic, certainly allowing for amendment and exclusion of aspects and other modifications. The next chapter will present the RobS that has been developed, based on this framework.

8 A RobS to Assess Feedback Competence on Writing in the EFL context

This chapter will lay out the properties of the RobS as the test instrument that was developed in this thesis and will be structured according to the three distinct phases of the RobS: The preparation, the simulation and the assessment (see Figure 16). Each section will display the final version of the material that was used and developed in this study.

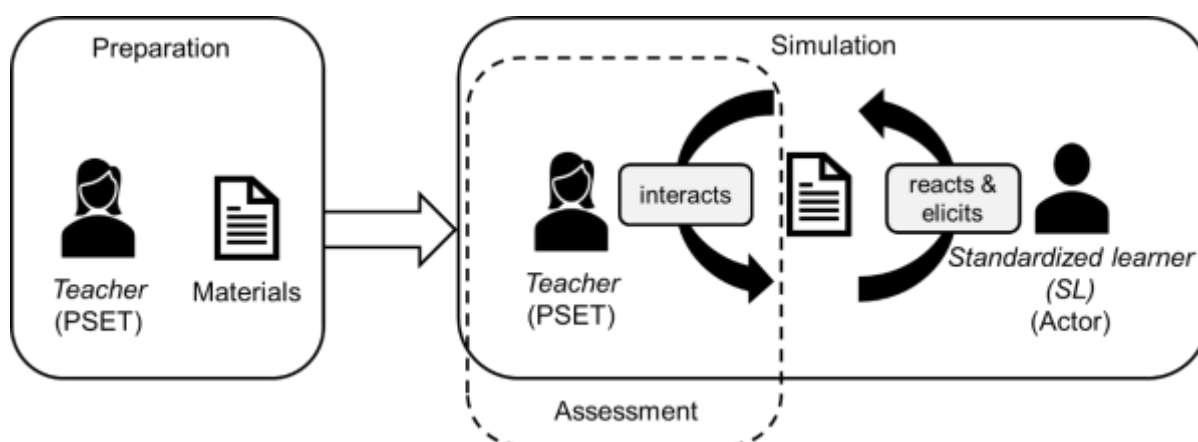


Figure 16: Schematic procedure of the RobS (Adapted and translated from Wotschel & Janzen, in preparation).

8.1 Preparation

The preparation phase has a duration of 15 minutes and allows participants to get ready for the simulation. Participants sit at a table and are provided with writing utensils, a timer, water and fruit lozenges. Most importantly, participants have time to work through the material package, which includes the following documents (see Appendix A):

- M0 Vignette (e.g., context and task description);
- M1 The learner text;
- M2 The task for the text;
- M3 Excerpt from the series of lessons on the topic;
- M4 Sample text from the series of lessons;

- M5 Empty sheet for notes.

The material package is predominantly in English, except for M3 and a fictional quote from the SL in M0.

In M0, the participant is provided with the task. The following information on the situation is given:

You have been working as an English teacher at the Heinz-Nixdorf-[comprehensive school] for 4 months now. In your 9th grade you made an offer to provide your students with feedback in the writing process. One of your students has thus handed in her first draft of a newspaper report after the 6th lesson in your plan. She stated the following: "Können Sie bitte schauen, ob das so in Ordnung ist? Ich bin mir nicht sicher, ob das alles so passt." (M0; adaptation made by TJ)

The learner in that situation asks, in German: "Could you please have a look at this? I am not sure if all of this is okay." In addition, participants receive the following information about the learner: "Mia is 15 years old and rather average in English, her grades are between 3 and 4, though the last class test was a 5+" (M0). The participant then receives the following task:

You have 15 Minutes to prepare for a feedback conversation, based on the learner text as well as the additional material. The conversation is to take place during the 10-minute-break after the lesson.

You can also add some written remarks for Mia to the text. (M0)

The situation has been purposefully designed so that Mia approaches the teacher. This behavior indicates that Mia appreciates the feedback process and is willing to take action, both are aspects of feedback literacy (Carless & Boud, 2018, pp. 1316–1318; see also Chapter 6.1.2). Furthermore, seeking feedback is a self-regulation strategy, and an aspect which the teacher can provide positive feedback on (see Chapter 7.3.1.2). It also specifically states that the text is a first draft of the news report, reflecting the learning objectives in M3, in that multiple drafting is planned for in this series of lessons. This aligns with a process approach to writing as indicated in Chapters 6.2 and 7.3.2.2.

M1 contains the learner text, which is the Mia has created during the lesson. In order to facilitate commenting, an extra margin is provided on the right side (M1). The text is supposed to be a news report, the report being a genre which learners need to master to achieve the B1 level, according to the CEFR (see Chapter 6.2.3). This genre has also been chosen as it can be assumed that participants have sufficient knowledge to evaluate this type of text as it mostly relates to general knowledge. The text shows instances in which the learner has already succeeded, as well as several areas in need of improvement. It includes a variety of features concerning the text dimensions as described by Becker-Mrotzek (2014, p. 508) and presented in Chapter 7.3.3.1 that can be addressed in the feedback conversation. There are features on the surface-level considering

language correctness, like spelling, punctuation and lexico-grammar. For example, the word *parents* is spelled as **parants*. Frequently, as an example for language appropriateness, the present simple/bare infinitive is used in the text although a past tense, i.e., past simple, is required in the specific context. With regard to deep-level features, the text is, for example, missing a headline, distinct paragraphs on a structural level, and excludes specific pieces of required information in terms of content. Also, no revision marks are visible in the text concerning the process level. Additionally, the text includes several error patterns. For instance, the past simple is used inconsistently throughout the text, often the present simple/bare infinitive is used instead, and if used, the past simple is formed incorrectly for irregular verbs. Also, several false cognates, i.e., L1-inferences, are found in the text which is typical for learners that have German as their L1. There are erroneous text features that can be regarded as relevant for achieving the learning objective. This relates most importantly to genre-specific features like the use of past tense when reporting on a past event, the objective and neutral tone of a news report, and the structural elements such as a headline. However, there are also errors that are not directly related to the genre, like spelling mistakes or L1-inferential false cognates. A list of features that could potentially be diagnosed in the text and addressed in the feedback conversation is depicted in Table 2.

Table 2: Text features to diagnose and address in the feedback conversation.

Aspect	Successful aspects	Errors / Discrepancies between current and desired state	Possible corrections
Language correctness		Missing preposition before <i>Tuesday</i>	On Tuesday
		Missing agreement (<i>*win</i>)	wins
		False cognate: <i>action</i>	project
		Missing quotation marks (<i>Plant trees for bees</i>)	"Plant trees for bees"
		<i>*envirenmetal minister</i>	Minister for the Environment
		Syntax: <i>*too was there</i>	... was there, too.
		Missing colon and full stop in the direct quotation	He said: "... "
		False cognate: <i>become</i>	received, got
		Syntax: <i>even</i>	We even received...
		<i>*for plant</i>	for planting
		<i>*tress</i>	trees

Aspect	Successful aspects	Errors / Discrepancies between current and desired state	Possible corrections
Language correctness		Exclamation mark not appropriate for the genre	
		False cognate: <i>*It give</i>	There are
		Preposition: <i>*We raise money with</i>	We raised money by
		Wrong conjugation: <i>*bayed</i>	bought
		Missing comma: <i>*Also</i>	Also,
		Wrong punctuation (colon): <i>*the bank paid 5€ per collected kilo garbage: 207kg garbage were collected and 1035€ donated</i>	
		Missing agreement: <i>207kg of garbage was collected*</i>	207kg of garbage were collected
		<i>*parants, *grandparants</i>	parents, grandparent
		<i>*speeck</i>	Speak / speech
	Tuesday; gymnasium; was there		
Language appropriateness		Wrong Tense (<i>win, is, meet, says, become, raise, *speeck</i>)	Won, was, met, said, received, raised, said
		Copying of 'NRW Umweltpreis' without quotation marks or paraphrasing	
		Use of subjective language (we)	they
		<i>We raise money with selling waffles, selling art...</i>	We raised money by selling waffles and art.
		<i>Also*</i>	So,
	Neutral beginning of the news report		
Structure		Missing headline	
		Early positioning of quotes	Position them later in the text
		<i>*It give different projects.</i>	Missing audience guidance
		Missing coherence: <i>*Also the bank paid 5€ per collected kilo garbage</i>	
		Value statement: It was an amazing event.	
	Lead includes relevant information; quotations		

Aspect	Successful aspects	Errors / Discrepancies between current and desired state	Possible corrections
Content		Aim of the project is not addressed	
		Number of trees is not addressed	The students planted 500 trees.
		<i>*more tress</i>	Missing information
		Copying from the task sheet: <i>the bank paid 5€ per collected kilo garbage: 207kg garbage were collected and 1035€ donated</i>	
		Missing inclusion of the pictures	
	Most facts are included, i.e., the donations, projects		
Process	Linguistic risks	No visible revision marks	

M2 contains the task for the learner. The task is formulated as follows:

Task: Yesterday, your school received the “NRW Umweltpreis” for its project “Plant trees for bees”. As a junior journalist, it is your task to **write a short news report** on the ceremony and project for the local newspaper.

Keep in mind the **structure** and **style** of a news report.

Use the **notes** and **pictures** below. (M2; emphasis in the original)

The task can be situated within the demands of the guidelines, as indicated in Chapter 6.2.3. The report is a genre that is specified in the CEFR as relevant for B1 (Council of Europe, 2020, p. 68) and is also included in the *BISTAs* (KMK, 2004, p. 78, 2023a, p. 16). The emphases in the task formulation are intended to indicate to the learner which aspects to focus on in the writing process. The task clearly states the genre, i.e., the news report, and highlights that the learner must focus on structure and style, i.e., genre-specific features. Also, the learner is required to use the notes and the pictures that are provided beneath the task. The notes are pre-structured according to the project, the aim and the ceremony. Pictures are provided on the task sheet and could hypothetically be included in the text, as the learner text is provided as a print out of a digitally created text. Beneath the respective picture, scaffolding is provided in the form of a description of the respective project, i.e., selling waffles, selling art, and picking up garbage. In the context of the ceremony, two quotes are provided in German,

which could be included in the learner text by the means of direct or indirect speech.

M3 contains the excerpt from the series of lessons on the topic. The overarching topic for the unit is the environment, a topic which is potentially relevant to the learners as it relates to their lives. The lesson plan is primarily based on a process and genre approach to writing, as described in Chapter 6.2.2 (Hyland, 2019, pp. 10–21). First, learners are familiarized with newspapers and the articles that they include on the topic of the environment, which they have to briefly summarize and present. In the following lesson, learners work with a news report and are introduced to relevant lexical items by creating a mind-map. Those first sessions can be classified as “setting the context and building the field,” according to the teaching-learning cycle (Hyland, 2019, p. 20). Afterwards, learners are encouraged to more specifically engage with the genre by modelling and deconstructing news reports. They are instructed to identify specific structural elements such as the headline, and are made aware of specific linguistic means that are typical to news reports, like the past tense or use of direct and indirect speech. This step reflects the steps of “modelling and deconstruction” (Hyland, 2019, p. 20). The next lesson is the one which is focused on in the RobS. Learners receive the task in M2 and write a first draft of a news report themselves. As this session is specifically focused on a first draft, a clear orientation towards a process writing approach is evident. It therefore explicitly fosters the process aspect of writing competence, as indicated in the CEFR (Council of Europe, 2020, p. 68) and the original *BISTAs* (KMK, 2004, p. 17) (see also Chapter 6.2.3). In the subsequent session, learners receive feedback on their reports from their peers in a writing conference and can then use the information to revise their draft. This stage is, on the one hand, reminiscent of the step of “joint construction,” in which learners collaborate in the writing process (Hyland, 2019, p. 20). Also, it emphasizes the relevance of revision in the process of writing as it is implemented within instruction. As an outlook for the further direction of the unit, a project that aims at the creation of a class newspaper is hinted at. The excerpt is purposefully kept brief so as not to provide too much information, considering the time restraints of the RobS and, thus, lacks the necessary detail for actual plans of a series of lessons.

M4 contains an abbreviated version of the news report “Prince William reveals Earthshot Prize winners in global bid to tackle climate crisis” (Lock & agencies, 2021), published in *The Guardian*, which is covered in that series of lessons. The text is thematically similar to the task that the learners have to focus on, i.e., an event that includes an awards ceremony focusing on the environment. It shows the relevant structural elements of a news report, e.g., headline, lead up and lead development (Martin & Rose, 2008, pp. 76–77; Watanabe, 2017, p. 68;

White, 1998, p. 266), is written using past tense and includes direct citations. Therefore, the report serves as an exemplar for the deconstruction as indicated in M3, but also as a scaffolding device in the feedback conversation, which the PSETs can use. M5 is simply an empty sheet for notes.

8.2 Simulation

In the simulation, participants interact with the SL. After 15 minutes of preparation, the ten-minute simulation starts with the SL knocking at the door. In Figure 17 a still image of a simulation is displayed and the transcript of a (different) simulation is included in Appendix G.1 as a (coding) sample.



Figure 17: Still image of a simulation (AAS13).

The overall procedure of the simulations is identical to all. The SL enters the room and opens the conversation with the initial VT "Good morning." Following that, the SL sits down next to the participant and the feedback conversation unfolds. After eight to nine minutes, the SL is supposed to indicate that the break is almost over and that she needs to go to the next lesson. After about ten minutes, the SL is supposed to signal that the time is up. The SL is instructed to behave in a certain standardized manner throughout the conversation, which will be outlined in the subsequent sections.

8.2.1 The Role of Mia

The SL received training to depict the role of Mia in the most standardized way possible (see Chapter 14.1.2). A detailed German-language role description for Mia, which the SL bases their performance on, can be found in Appendix A.7. Mia is 15 years old, identifies as female, and is in the ninth grade of a

comprehensive school in North Rhine-Westphalia, Germany. She is rather average in her performance in English, but also has not put much effort into the subject recently. This has led to her failing the last class test, getting the grade of 5+. She has decided to put in more effort and has, thus, decided to ask for feedback on the first draft to find out what she can improve on and how she can practice. When writing the text, Mia simply started without planning and used the task sheet as an exemplar. Mia considers the text to be 'okay' – she realizes that there are some language issues, but she has tried to include all the facts from the task sheet. Mia is supposed to have a language level of A2, approaching B1. Her main issue can be found in the area of verb conjugation, especially past simple, as she did not really learn or memorize the irregular verb forms, relying on the regular formation by adding the *-ed* suffix. However, she understands that these forms exist and would be able to use them with appropriate scaffolding and guidance. Mia did not really comprehend the content in the last few lessons because she has difficulties understanding English. In the past, she has managed well with a mix of English and German. Thus, she also prefers to have the feedback conversation in German, but does not want the teacher to know that she has difficulties, and thus tries to participate in the conversation in English initially. In the course of the conversation, she makes it increasingly clear that she does not understand the English input and wants to have the conversation in German. Mia has several aims for the feedback conversation and she wants to find out the following:

- Is the structure appropriate for a news report?
- What are the language issues in the text? (Especially grammar)
- Did she miss any information for the report?
- What does she have to do in order to improve?

These aims relate to the dimensions of text, but also to *feed forward* as the aim of feedback, which is considered to be most important to learners (see Chapter 7.3.1.1).

8.2.2 Standardization of the Learner

In order to achieve standardization in the simulation, the SL is instructed to behave in a standardized manner. For one, this is realized by a description of Mia's competence, which the SL can use as a backdrop. The competence levels are defined in reference to the concept of the *Zone of proximal development* (ZPD) Vygotskij (1978):

On her own, Mia...

- can apply the macro structure of texts (introduction, main part, conclusion),
- can report on something with provided information,

- knows about the past simple and its regular formation,
- is able to use simple English language syntax,
- can indicate the she did not understand something in English,
- can ask simple questions in English,
- knows about the rules concerning the conjugation of verbs in the present simple, i.e., the 3rd person singular -s.

With appropriate scaffolding/assistance, Mia can ...

- understand and use the feedback, if it is provided in German,
- identify irregular past simple forms in the text,
- spell accurately,
- name characteristics of news reports,
- use quotation marks,
- use direct speech,
- add missing pieces of information to the text.

Even with help Mia cannot...

- use more complex syntax,
- explain complex grammatical rules herself,
- appropriately use commas,
- name the correct irregular verb forms,
- speak in English fluently with the teacher,
- understand the teacher talking in English.

In order to standardize the dialogue through the output of the SL, several VTs are designed to elicit specific facets of the framework in a dynamic but standardized manner. The VTs are listed in Table 3. Some VTs need to be used at specific points in the RobS. For instance, the SL is instructed to start the RobS with the VT "Good morning," so that the participant is more inclined to start the conversation in English, which is necessary for the aspect of *Adaptation of language*. Subsequently, other VTs are used to induce a language switch. The VT "Muss ich den Text nochmal neu schreiben?" (*Do have to write the text again?*) is to be used rather at the end, and relates to the aspect of *Process orientation*. This VT is supposed to be used if revision has not been addressed in the RobS by the participant. The VT "Wie kann ich das üben?" (*How can I practice that?*) can be used more flexibly, but it should occur after an aspect of the text has been discussed which can sensibly be practiced, for example, the use and formulation of past simple. The VT can also refer to the aspect of *Self-regulation* as it provides an opportunity to provide the learner with information on learning strategies, but also to *feed forward*, as it directly relates to actions the learner can take to improve in certain aspects. Other VTs can be used flexibly, if the situation allows for them. The VTs for *Adaptation of content* are to be used to

generally indicate that the learner has issues with the content of the previous lessons, nudging participants to check their understanding more often. The VTs for *Adaptive error correction* are to be used after an opportunity for self-correction has been given, triggering adaptive behavior. The VTs that relate to diagnosis each refer directly to a text dimension. If one of the dimensions has been discussed by the participant, the SL is not supposed to use a VT that focuses on that aspect of the text.

Table 3: Verbal triggers used in the RobS.

Aspect of the framework		Verbal trigger (Original)	Verbal trigger (English)
Aims	Feed Up	-	-
	Feed Back	-	-
	Feed Forward	Wie kann ich das üben?	How can I practice that?
	Positive feedback	-	-
	Self-evaluation	-	-
Level	Task	-	-
	Process	-	-
	Self-regulation	Wie kann ich das üben?	How can I practice that?
	Self	-	-
Communication	Dialogue orientation	-	-
	Adaptation of content	Das habe ich nicht verstanden.	I did not understand that.
		Was habe ich hier falsch gemacht?	What did I do wrong here?
		Ist das hier richtig?	Is this correct here?
	Adaptation of language	Good morning.	Good morning.
		Can you say that in German?	Can you say that in German?
		Können Sie das (nochmal) auf Deutsch sagen?	Can you say that in German (again), please?
	Descriptiveness	-	-
	Specificity and Clarity	-	-

Aspect of the framework		Verbal trigger (Original)	Verbal trigger (English)
L2/FL writing pedagogy	Adaptive error correction	Nein, das weiß ich nicht. (As a reaction to a question.)	No, I don't know. (As a reaction to a question.)
		Können Sie das nochmal erklären?	Could you explain it again?
	Focused error selection	-	-
	Target-orientation	-	-
	Process orientation	Muss ich den Text nochmal schreiben/machen?	Do I have to write the text again?
	Prioritization of deep-level features	-	-
Diagnosis	Language appropriateness	Ist das so wie ein Zeitungsbericht sein soll?	Is this like a real news report should be?
	Language correctness	Ist das richtig so? Ich bin mir unsicher.	Is this correct? I am not sure.
	Structure	Ist das vom Aufbau her in Ordnung?	Is the structure okay?
	Content	Ist vom Inhalt her alles drin?	Is everything included with regard to content?
	Process	-	-
Information	Detailedness	-	-
	Professional accuracy	-	-
	Language accuracy	-	-

8.3 Assessment

The participant's performance in the simulation is assessed using a scoring instrument. For assessment, only the behavior of the participant is accounted for. The behavior of the SL is not considered, as they are instructed to behave in a standardized manner and their actions do not indicate the success of the feedback conversation. Assessment could occur either concurrently or subsequently to the performance. If assessment is done concurrently, two assessors are advised in order to achieve a reliable item score (see Chapter 15.4). If assessment is done subsequently, a video recording of the RobS is advised to ensure reliability. In the following, the instrument will first be presented as a whole, and more details on the individual items presented thereafter.

The scoring instrument, as presented in Table 4, is used to assess the performance⁹⁰. It consists of 24 items in total, which leads to the maximum score of 48 points. Four different types of items are used in the scoring instrument. Dichotomous items are scored with either two or zero points. Of these 13 items, twelve are event-based and scored with two points if a described event occurs (or, in the case of *Self*, does not occur) at least once. One dichotomous item (*Body language*) considers the whole duration of the RobS and is technically based on relative-frequency of behavior, but is grouped with dichotomous items due to the scoring similarities. The items *Revision* and *Adaptive error correction* are tiered event-based, i.e., the observed event that relates to the highest item score for these criteria is scored. The item scores of zero, one or two points are differentiated. There are two conditional relative frequency-based items, *Quality of Feed Back* and *Dialogue orientation*. For the item score to be included in the test score, specific events have to occur at least once. If the event has occurred, the relative frequency is considered. Overall, seven items are either based on relative or absolute frequency and are scored with either zero points, one point, or two points. Relative frequency considers the share of observed behavior compared to the whole performance. The qualifiers that are used to describe this relative frequency are *never* (zero points), *sometimes* (one point) and *often* (two points). *Never* clearly indicates that behavior is not evident in the RobS. *Sometimes*, as a qualifier, refers to behavior that is evident, but less frequent than *half of the time* (Bocklisch et al., 2012, p. 148). *Often* refers to behavior that is more frequent in relation to a qualifier that describes behavior as occurring *half of the time* (Bocklisch et al., 2012, p. 148). These three qualifiers are perceived as nearly equidistant in their relation to frequency, as reported by Bocklisch et al. (2012, p. 147). Absolute frequency considers the distinct instances of behavior that occur and is specific to each item. A detailed rating manual which provides further specifications and examples can be found in Appendix B. Each item is derived from the framework and can be viewed as a corresponding operationalization.

⁹⁰ The presented scoring instrument represents the final version after validation. Some items from previous versions have been excluded after reliability analyses (see Chapter 15.4.1.3.1).

Table 4: *Scoring instrument for the RobS.*

Dichotomous Criteria			
	Explanation	0	2
Self-evaluation	TP ⁹¹ asks the learner for their own evaluation of the performance (task, process or self-regulation), e.g., by asking the learner for specific challenges or specific aspects the learner succeeded in.		
Positive feedback	TP mentions specific aspects that the learner has achieved successfully concerning the task, process or self-regulation (not self) at least once.		
Learning objectives (Feed Up)	TP clearly states at least one context-relevant and specific learning objective.		
Process	TP addresses aspects of the process of solving the task (feed up/back/forward) at least once.		
Self-regulation	TP addresses the use of specific strategies (feed up/back/forward) on a metacognitive level at least once.		
Self	TP does not address the learner's personal level, e.g., no generic praise or criticism of the learner is given.		
Prioritization of deep-level features	TP first provides feedback on deep-level features that need improvement, e.g., content or structure.		
Error patterns	TP explicitly addresses error patterns (e.g. tense, irregular verbs, L1- inferential errors like false friends or cognates) and provides feedback on them (KR, KCR or further elaboration) at least once.		
Surface: Language correctness	TP refers to aspects of language correctness, e.g., forming of past simple at least once.		
Surface: Language appropriateness	TP refers to aspects of language appropriateness, e.g., inappropriate use of tenses at least once.		
Deep: Structure	TP refers to aspects of structure, e.g., cohesion, headline or audience-orientation at least once.		
Deep: Content	TP refers to aspects of content, e.g., missing information at least once.		
Body language	TP consistently conveys attentiveness through body language, appropriate eye-contact and facial expressions.		

⁹¹ TP is the abbreviation of test person.

Conditional Relative Frequency-based Criteria							
	Condition			Available points			Achieved points
		Yes	No	0	1	2	
Dialogue orientation	The triggered engagement is at least once cognitively demanding for the learner and relevant for the feedback conversation.			TP does not involve the learner in the feedback conversation, e.g., through questions or prompts which are at least once cognitively demanding for the learner.	TP sometimes involves the learner in the feedback conversation, e.g., through questions or prompts.	TP often involves the learner in the feedback conversation, e.g., through questions or prompts.	
Quality of Feed Back	The feedback is at least once explicitly provided against specific learning objectives.			TP never provides feedback that is accurate, clear and on a descriptive level.	TP provides feedback that is sometimes accurate, clear and on a descriptive level.	TP provides feedback that is often accurate, clear and on a descriptive level.	

Tiered Event-based Criteria				
	Available points			Achieved points
	0	1	2	
Revision	TP does not state that the feedback needs to be used for revision or for future tasks or TP is not clear/contradictive in their instruction.	TP advises (but does not instruct) the learner to use the feedback, e.g., for revision or for future tasks. This includes revision in class or at home.	TP clearly instructs that the text needs to be revised using the feedback. This includes revision in class or at home.	
Adaptive error correction	TP does not provide an opportunity for self-correction and only provides explicit and direct feedback	TP at least once provides opportunities for learner self-correction but explicit feedback is provided subsequently without an intermediate step. Opportunity for self-correction can be a question or implicit/indirect feedback.	TP at least once provides opportunities for learner self-correction through graded feedback that begins with a question and includes at least one intermediate step before providing explicit feedback.	

Relative and Absolute Frequency-based Criteria				
	Available points			Achieved points
	0	1	2	
Quality of Feed Forward	TP provides calls to action that are never positive, unambiguous and specific.	TP provides calls to action that are sometimes positive, unambiguous and specific.	TP provides calls to action that are often positive, unambiguous and specific.	

Relative and Absolute Frequency-based Criteria				
	0	1	2	Achieved points
Target-oriented error selection	TP does not provide feedback on selected errors, error patterns or mistakes that are relevant for achieving the learning objectives. Errors, error patterns or mistakes are discussed chronologically without selecting specific errors.	TP sometimes provides feedback on selected errors, error patterns or mistakes that are relevant for achieving the learning objectives. Other errors, error patterns or mistakes that are not directly related to the learning objectives are discussed equally.	TP often provides feedback on selected errors, error patterns or mistakes that are relevant for achieving the learning objectives. Other errors, error patterns or mistakes that are not directly related to the learning objectives can be discussed marginally.	
Detailed-ness	TP only provides KR or KCR without specific elaboration, i.e., only the correct solution is given or it is said whether something is (in)correct.	TP sometimes provides specific elaborated feedback, e.g., explains rules, gives reasons or provides examples. KCR and KR dominate, i.e., only the correct solution is given or it is said whether something is (in)correct.	TP often provides specific elaborated feedback, e.g., explains rules, gives reasons or provides examples. In some instances, only KCR is provided, i.e., only the correct solution is given.	
Amount and balance of information	TP provides an inappropriate amount of information (too little, too much), regardless of balance between surface- and deep-level features.	TP provides an appropriate amount information on different aspects of the text and the writing process (about 3-5 aspects). However, surface- and deep-level features are not addressed in a balanced proportion.	TP provides an appropriate amount information on different aspects of the text and the writing process (about 3-5 aspects). Surface- and deep-level features are addressed in a balanced proportion.	
Checking understanding	TP never assures themselves of learner's understanding and does not adapt to their needs, e.g., by rephrasing or	TP assures themselves a few times, but at least once, of learner's understanding and adapts to their needs, e.g., by rephrasing or giving alternative explanations.	TP assures themselves several times of learner's understanding and adapts to their needs, e.g., by rephrasing or giving alternative explanations.	

Relative and Absolute Frequency-based Criteria				
	giving alternative explanations.			
	0	1	2	Achieved points
Language accuracy	TP shows consistent inaccuracies and errors in the English language that either in sum or individually potentially hinder learning.	TP predominantly uses the foreign language correctly. However, the TP also shows rare instances of inaccuracies and errors that potentially hinder learning, e.g., providing incorrect corrections.	TP uses the foreign language correctly almost always with only a few minor inaccuracies that do not potentially hinder learning, e.g., providing incorrect corrections.	
	This includes speaking English as well as when talking about the English language, e.g., providing corrections or rules on surface-level features.	This includes speaking English as well as when talking about the English language, e.g., providing corrections or rules on surface-level features.	This includes speaking English as well as when talking about the English language, e.g., providing corrections or rules on surface-level features.	
Professional accuracy	TP shows consistent inaccuracies when addressing professional non-language related aspects and providing information on them, e.g., deep-level features and learning strategies. TP does not address non-language related, e.g., deep-level features and learning strategies	TP is predominantly correct when addressing professional non-language related aspects and providing information on them, e.g., deep-level features and learning strategies. However, TP also shows rare instances of inaccuracies and errors that do potentially hinder learning, e.g., providing incorrect corrections.	TP is always correct, e.g., does not potentially hinder learning when addressing professional non-language related aspects and providing information on them, e.g., deep-level features and learning strategies.	
Sum				
				Max: 48

Most of the event-based items are a direct result of aspects in the theoretical framework and named equally (i.e., *Self-evaluation*, *Positive feedback*, *Learning objectives (Feed Up)*, *Process*, *Self-regulation*, *Self*, *Surface: Language correctness*, *Surface: Language appropriateness*, *Deep: Structure*, and *Deep: Content*). They will not be discussed in more detail here as the explanation in Table 4 suffices and additional information is also provided in the rating manual (see Appendix B). The operationalization of the item *Prioritization of deep-level features* deviates from its origins, which argue that the prioritization should occur in relation to multiple drafts (see Chapter 7.3.2.2). For the context of the RobS, it was changed, so that the initial focus of the conversation is on deep-level features, as such a temporal prioritization within the RobS might serve as an indication of how teachers value different features of the text. The item *Error patterns* is derived from the aspect of *Focused error selection* in the framework (see Chapter 7.3.2.2). This aspect, among other features, describes that it is more beneficial to provide feedback on error patterns. Thus, the item evaluates whether the participant provides feedback on an error pattern which they have identified in the text. It is operationalized as such, so that the extent and quality in which the feedback is provided is not to be considered for this item. The item *Body language* is addressed in Chapter 7.3.2.1 as part of *Dialogue orientation*. It is specifically operationalized as attentiveness in behavior that can be considered to be supportive of *Dialogue orientation*, i.e., the overall posture, eye-contact and facial expressions.

Two items are described as tiered event-based. The item *Revision* describes three tiers of quality, in which participants can address the issue of whether the text needs to be revised based on the feedback, and is in reference to the aspect *Process orientation* (Chapter 7.3.2.2). Zero points are scored if the participant indicates that no revision has to occur, as this clearly contradicts the process approach to writing, which is also indicated in the material (M4). One point is scored if the revision is advised but not mandatory. Such an event indicates that the participant is aware of the benefits of the writing process, but it is not ensured that the feedback is actually used by the learner. Two points are then scored if there is a clear instruction by the participant that revision has to occur, as this aligns most clearly with a process approach. The item *Adaptive error correction* is related to the aspect of the same name (see Chapter 7.3.2.2). It is operationalized as such that zero points are scored if no opportunity for self-correction is provided. One point is scored if at least one opportunity for self-correction is provided, but it is directly followed by explicit feedback, while two points are scored if there is an intermediary instance of implicit or indirect feedback as a sort of scaffolding between self-correction opportunity and explicit feedback.

Two items are described as conditional frequency-based. The conditions are events that are related to the frequency-based item and which have to occur within the performance in order to include the item score in the aggregated test score. For the item *Dialogue orientation*, the conditional event is an instance in which the engagement that the participant initiates is cognitively demanding at least once. This condition is added to ensure that no points are scored if dialogue occurs, but it is only on a personal level, without relating to the context. Engagement is cognitively demanding if it involves the learner more deeply with subject-relevant issues and is outlined in more detail in the rating manual. Points are then scored for the extent to which the participant engages the SL in the conversation, regardless of the focus of that engagement. For *Quality of Feed Back*, the condition requires an instance in which the feedback information is directly related to a specific learning objective. Such an event is potentially quite impactful for learning, as it clearly indicates the gap between the current and desired state. The points are then scored for the frequency with which qualitative feed back occurs. The qualifiers are accuracy, clarity and descriptiveness. Accuracy refers indirectly to the aspect of information (see Chapter 7.3.3.2). The other qualifiers, clarity and descriptiveness, directly refer to the communicative aspect of the framework, emphasizing the necessity to articulate the feed back information precisely and in a nonevaluative manner (see Chapter 7.3.2.1). Similarly, *Quality of Feed Forward* also provides points with regard to the frequency with which qualitative feed forward is provided. Quality is operationalized as positive, unambiguous and specific, which all refer back to Chapter 7.3.2.1. The first qualifier refers back to Rosenberg's concept of positive action language (Rosenberg & Holler, 2016, p. 75). Furthermore, the actions that the learner is supposed to take should be unambiguous, e.g., it should be clearly communicated what they have to do, and specific in that the calls to action directly relate to an issue in the text and how the learner can improve on that distinct issue.

Besides the dichotomously operationalized item *Body language* and the conditional relative-frequency-based items, there are three further relative frequency-based items. They consider the relative share of which the described behavior occurs compared to the whole duration of the conversation. *Quality of Feed Forward* has already been addressed with *Quality of Feed Back*. The item *Target-oriented error selection* combines two aspects of Chapter 7.3.2.2, i.e., *Target orientation* and *Focused Error selection*, and reflects the share of feedback that is relevant for the learning objectives. The item *Detailedness* reflects the share of feedback that is provided with elaboration, e.g., with additional information beyond a correct solution (KCR) or statements of (in-)correctness (KR).

The last four items are absolute frequency-based, as they consider a distinct number of instances within the performance. *Amount and balance of information*

is reflective of the aspect of *Focused error selection* in the framework (Chapter 7.3.2.2). The item combines the notion of consciously reducing the amount of feedback, so as not to overload the learner, with the findings that an overall balance with regard to text-level features is more beneficial (Chapter 7.3.2.2). Zero points are provided if the amount of information is either too little or too high, relating especially to the notion of selective feedback. One point is scored if the amount is appropriate, which is defined as about three to five aspects, but there is an imbalance between surface- and deep-level features. The specific number of aspects is kept consciously vague in order to allow for performance-specific considerations, especially if errors are grouped and clustered in the feedback conversation. Two points are then scored if the amount is appropriate and there is a balance of text-level features. The item *Checking understanding* is the operationalization of the aspect *Adaptation of content* (see Chapter 7.3.2.1), in that, it accounts for the distinct number of times in which the participant assures themselves of the learner's understanding. This step is a prerequisite for adaptation, and also indicates the extent to which the participant assures themselves that the learner is actually able to use the feedback later on, which is essential considering feedback literacies (see Chapter 6.1.2). The item *Language accuracy* considers the instances in which the participant uses erroneous language, either when talking about features of the English language or speaking in English, that can be considered harmful for the learning process. Zero points are scored if inaccuracies that could potentially be harmful to learning occur consistently. One point is scored if there are only rare instances of inaccuracies that bear potential to hinder learning, while two points are scored if only few and minor inaccuracies occur that do not potentially impede learning. The item *Professional accuracy* is operationalized similarly, but as errors in professionalism are considered to necessarily lead to learning impediments, two points are only scored if no errors are made. One point is scored if the participant is predominantly correct, while zero points are scored if inaccuracies occur consistently. The accuracy items thus slightly deviate with regard to their operationalization of frequency to account for the importance of providing correct information to the learner.

The scoring instrument, as presented in its final version here, does not directly address several aspects from the framework. The aspect of *Task* is not distinctly addressed as an event-based item like *Process*, *Self-regulation* or *Self*, because it is addressed in several other items, especially in the items that focus on the dimensions of the text as a product of the task. Similarly, *Feed Back* and *Feed Forward* are also not operationalized as event-based items like *Feed Up*, as the focus is put on the frequency of qualitative events. With regard to the former, the decision was made because, similar to the task-level, the distinct way in which feed back is provided on the text is differentiated in other items like that

of *Diagnosis*. For *Feed Forward*, this decision was made in order to account for the frequency of high-quality events, as this aim is considered most important by learners. Other aspects that are part of the framework but not included in the instrument are *Adaptation to language* and *Deep: Process*. Both items can be found in the rating manual (see Appendix B). They are not included in the final scoring instrument due to reasons of internal consistency, as outlined in detail in Chapter 15.4. They can, nonetheless, be relevant for discussing feedback on writing in the EFL context and are thus still included in the framework and addressed as aspects of the simulation.

8.4 Discussion and Limitations of the RobS

The RobS, as presented, can be discussed against the theoretical background of outlined in Chapter 4. From a content perspective, the simulation focuses on a core practice for EFL teachers, i.e., the provision of feedback, while still being connected to the statutory demands stated in the guidelines by the KMK. The simulation itself can be categorized along the spectra outlined in Chapter 4.3.2 as fully analogue, with exclusively human face-to-face interaction, and a relatively high extent of fidelity. Fidelity is mostly reduced due to the fact that a) the physical context is not identical to a school setting and b) the actor is not actually the age of a ninth grader. However, it is still an interactive situation in which the actor is supposed to react like a learner and the task, i.e., providing feedback, occurs similarly in the school context. The RobS can also be discussed against the characteristics of simulations as outlined in Chapter 4.3.1: simplification, authenticity, flexibility, and structure (Dieker et al., 2014, p. 22; Grotegut, 2023, p. 17). The simulation is an “approximation of practice” (Grossman, Compton, et al., 2009, p. 2076), i.e., a simplification, as it only includes a conversation with a single learner, not a larger group. Participants also receive an extended amount of time to prepare with a predefined set of material, similar to other simulations in the field of teacher education (e.g., Codreanu et al., 2022). Several design choices are considered to create an authentic situation for the participants, as described previously, such as the interaction with an actual human. The simulation is flexible as the actor reacts to the behavior and statements of the participants, but still structured, as the reactions are standardized, e.g., in the form of VTs. The assessment is also designed criterion-oriented, specifically focusing on performance (Biggs, 1996, p. 353).

Like any other test, the RobS presented here, comes with its own limitations. Regarding the preparation for the simulation, the amount of information is confined to the material provided. Participants do not know much about the learner’s previous performance or behavior in or outside of the classroom and less about their personality. This complicates participants’ ability to adapt more

specifically to the learner in the preparation phase. Moreover, there is no personal history alluded to in the material, restricting participants' ability to build on an established foundation. While a personal relationship might be beneficial to the overall learning atmosphere and conducive to learning (Lotz, 2016, p. 84), the RobS is designed to focus on the feedback on writing in the EFL context. Thus, a deliberate choice was made to restrict personal information to a minimum, so that participants focus on the text from the beginning.

For the simulation, the issue of stereotyping learners is a potential limitation. As the RobS is intended to be used as an assessment format, standardization is an essential aim, to provide equal opportunities for all students. Consequently, an attempt was made to standardize learner behavior as the key element of the RobS, especially with the instrument of VTs. Hence, a specific learner type is presented in the RobS, which bears the risk that PSETs refer to specific stereotypes regarding, for instance, gender. In the material of the RobS, the learner is presented as female. Participants, also possibly induced by the minimal background information on the learner, might thus be inclined to refer back to stereotypes that refer to female learners, such as that they are generally better in modern foreign languages than male learners in the conversation (K. Field, 2004, p. 130). Consequently, transparency is required so that participants are made aware of that risk before participating in the RobS.

The scoring instrument is necessarily limited as it provides a specific and normative perspective on the RobS. As a criterion-oriented instrument, it is restricted to assess behavior that is represented by the different items and is less flexible to account for individual behavior in the RobS. Also, it focuses on observable and measurable behavior evident in the RobS. It does not account for the potential decision-making process that leads to that performance, i.e., situation-specific skills (Blömeke et al., 2015). To account for that, participants could be asked to reflect on their performance. As the criteria and items are mostly based on empirical findings, these norms can only be considered valid to the extent to which the findings are not refuted. Moreover, new relevant criteria might arise in the future, which might require modifications of the scoring instrument and the test. The presented version should thus always be seen in the context of its development. Lastly, the scoring instrument includes a relatively large number of items. This might complicate assessment, as assessors need to keep them in mind and account for several different facets while observing the RobS.

9 Summary

The aim of this part is threefold: Reasoning the choice of the test, i.e., focusing on feedback on writing in the EFL context, providing the framework for the development of the RobS, and presenting the RobS as an operationalization of that framework.

In order to explain the relevance of the test focus, the terms feedback and writing are defined first. Feedback is traditionally defined as a transmission of information, used to reduce discrepancies between a current and desired level of performance (Hattie & Timperley, 2007, p. 81), but is more recently considered from a socio-constructive perspective, viewing it as an “interactive process” (Schluer, 2022, p. 17). Conceptualizations of feedback literacy focus on the learner’s role in the feedback process and emphasize the importance of learners understanding and using the information (Carless & Boud, 2018, pp. 1316–1318). Teachers should thus foster learners’ feedback literacy and assist them in that process (Carless & Winstone, 2020, p. 154). In order to emphasize the relevance of teachers being able to provide appropriate feedback, the effectiveness of feedback is described based on several meta-analyses (Hattie & Timperley, 2007; Wisniewski et al., 2020), the statutory guidelines (KMK, 2019a, 2019b) and the perspective of core practices (Grossman, 2018; Pang, 2019). Writing is modelled as a process, which includes planning, translating/transcribing and revision (Bereiter & Scardamalia, 1987; Hayes & Flowers, 1980) for L1, and similarly for the L2 or FL context (Börner, 1987; Hyland, 2019; Krings, 1989). When writing is the focus of instruction, Hyland argues that it is mostly seen from the genre and process perspective, but that other dimensions of language structure, text function, creativity and content all factor in (Hyland, 2019). The statutory guidelines for the ninth grade in Germany (KMK, 2004, 2023a) describe that the targeted level is B1, according to the CEFR (Council of Europe, 2001, 2020) for the *Mittlerer Schulabschluss*. The descriptors indicate that writing is seen as a process that includes planning, writing and revising. Also, specific genres that learners need to be able to produce are provided, such as the report (Council of Europe, 2020, p. 67). Through the synthesis of these two facets, it is evident that the writing process requires feedback in order for revision to occur. Findings of meta-studies indicate that teacher-generated feedback on writing may be effective for L1 and L2/FL learner’s development, if provided appropriately (Biber et al., 2011; Scherer et al., 2024). However, there is an indication that learners rarely use feedback to revise texts or drafts (Porsch, 2010, p. 64). Hence, as it is important for teachers to be able to provide appropriate feedback on learner texts, but as it is rarely done appropriately, RobS could be implemented to assess this important competence. However, this requires a more explicit focus on this action-oriented competence in the first phase of teacher education.

Methodologically, the theoretical approach to designing a framework is discussed, and the inevitable limitations that appear with it, outlined. An emphasis is given to the potential bias that arises in the choice and distinction of *relevant* theories. A brief overview on existing research which seeks to examine feedback performances is provided and the frameworks, upon which the research instruments are based are reflected on as a starting point for the present framework. The framework consists of three basic dimensions: function, strategy, and topic. The functions of feedback include the subdimensions of aim and level. Considering the aim, this includes *Feed Up* (*Where am I going?*), *Feed Back* (*How am I going?*), and *Feed Forward* (*Where to next?*), based on the model of Hattie and Timperley (2007), and adds the notions of *Self-evaluation* and *Positive feedback*, in order to account for aspects of feedback literacy and motivation (Carless & Boud, 2018; Mory, 2004, p. 766; Narciss, 2006, 2008, 2017). With regard to the level of feedback, the respective levels described in the model of Hattie and Timperley (2007) are used, i.e., the levels of *Task*, *Process*, and *Self-regulation* level, on which feedback should occur, and *Self*, on which feedback should not occur. The strategies of feedback include the subdimensions of communication and L2/FL writing pedagogy. Concerning communication, a *Dialogue orientation* is advocated for (Aljaafreh & Lantolf, 1994; Carless, 2020; Carless & Boud, 2018; Carless & Winstone, 2020) and *Adaptivity* is considered with regard to content level (Hattie & Clarke, 2019, p. 51; Price et al., 2010, p. 279; Schluer, 2022, p. 42) and language complexity (Bleichenbacher et al., 2017, p. 17; Rütli-Joy, p. 28). The language that teachers use should be descriptive and nonevaluative (Kopp & Mandl, 2014a, p. 158; Rosenberg & Holler, 2016, p. 38; Shute, 2008, p. 153), as well as specific and clear (Ferris, 2013, p. 8; Hattie & Timperley, 2007, pp. 85–86; Kluger & DeNisi, 1996, p. 260; Rosenberg & Holler, 2016, pp. 75–77; Shute, 2008, p. 167, p. 174). Concerning L2/FL writing pedagogy, teachers should provide error correction that is adaptive (Aljaafreh & Lantolf, 1994, p. 468), focused and selective (Ellis, 2009, pp. 6–7; Ferris, 2013, p. 8; Hyland, 2019, p. 199; Lee, 2017, p. 75, 2019; Nicol & Macfarlane-Dick, 2006, p. 210; Shute, 2008, p. 159), and target-oriented (Hattie & Timperley, 2007, p. 104; Lee, 2017, p. 76). Writing should be understood as a process that includes planning and revision (Bereiter & Scardamalia, 1987; Börner, 1987; Ferris, 2013, p. 8; Hayes & Flowers, 1980; Hyland, 2019, p. 11; Lee, 2017, pp. 14–15; Porsch, 2010, p. 57) and deep-level features should be prioritized in the conversation first (Ferris, 2013, p. 8; Lee, 2017, pp. 76–77; Zamel, 1985, p. 96). Concerning the topics of feedback, a wide range of text-related issues should be addressed (Ferris, 2013, p. 8), e.g., the correctness and appropriateness of language, as well as the structure and content of the text and the process of text creation (Becker-Mrotzek, 2014, p. 508; Becker-Mrotzek & Böttcher, 2020, p. 128). The information that is provided should be

elaborated (Kürzinger & Pohlmann-Rother, 2019, p. 399; Lotz, 2016, pp. 133–134; Wisniewski et al., 2020, p. 12) and accurate in terms of profession- and language-related aspects.

The RobS has three distinct phases: preparation, simulation, and assessment. In the 15-minutes of preparation, participants receive a material package, which includes a vignette with their task for the RobS and information on the context (M0), the learner text, on which they have to provide feedback (M1), the task for which the text was written (M2), an excerpt of the series of lessons to situate the RobS (M3), a sample news report (M4) and an empty sheet for notes (M5). In the ten-minute simulation, the participants engage in a conversation with a standardized learner, portrayed by an actor. The SL is instructed to behave in a standardized manner and to use VTs to elicit reactions for specific aspects of the framework. The scoring instrument is used to assess the performance of the participant, using items that are operationalized, based on the framework. The RobS is discussed against the theoretical backdrop and situated within the previous work on simulations. Limitations of the RobS are outlined, referring to potential issues in all three phases. The next part of this thesis will lay out a roadmap that could be followed in order to evaluate the validity of the test, focusing on the second aim of this thesis.

Part III - Methodology: Investigating the Validity of the RobS

Validation is simple in principle, but difficult in practice.

Michael T. Kane (2012, p. 15)

After making the case for the necessity of performance-based assessment formats in Part I and presenting the test that was developed in Part II of this thesis, the remaining parts will discuss the merits of the test in detail, trying to address the second half of this thesis' aim and give answers to the overarching research question of this thesis (RQT):

RQT: To what extent can the RobS be considered valid?

This part focuses on the methodological background. For that purpose, the relevant criteria for discussing a test's merit are introduced. The focus is placed on validity as the primary concern for test developers (AERA, APA, NCME, 2014, p. 11). Using the argument-based approach to validation proposed by Kane (1992, 2006, 2013a), the specific interpretation-use argument that must be evaluated in the validity argument are outlined, using the framework for validation proposed by Schreiber and Gut (2022, p. 129).

10 Criteria for Test Evaluation

This chapter will outline the criteria that should be considered when the quality of a test is evaluated. Firstly, the traditional criteria that are often referred to will be described, referring to both so-called secondary and primary criteria. Secondly, the special position of validity will be outlined based on its history and recent conceptualization in literature. Thirdly, implications for the present validation process will be drawn.

10.1 Overview of Test Criteria

In literature that focuses on the (psychometric) quality of tests and assessment, a distinction is made between three primary criteria (objectivity, reliability, and validity) and several secondary criteria, such as practicality, fairness, and authenticity (Döring, 2023, pp. 437–440; Jürgens & Lissmann, 2015, pp. 127–133; W. Stadler & Kremmel, 2018, pp. 39–55).

10.1.1 Secondary Criteria: Practicality, Fairness and Authenticity

Several relevant criteria for discussing a test's merits are considered to be of less urgent concern than primary criteria, albeit still contributing substantially to a test's perception.

Practicality / Economy

Conducting tests usually requires a plethora of different resources. This includes time for preparing, organizing, executing, evaluating and providing feedback, the availability of space for people to take the tests in, and financial resources, i.e., material or personnel (W. Stadler & Kremmel, 2018, p. 52). The criterion of practicality, sometimes also referred to as economy, takes this into account as test organizers need to take these resources into consideration when designing and planning tests (Jürgens & Lissmann, 2015, p. 133; W. Stadler & Kremmel, 2018, p. 52). Especially in educational contexts, resources are limited and educators constantly weigh their resources with potential benefits of tests. Consequently, some assessment tasks are disregarded due to them being impractical or uneconomic, even though they may be considered valid for a certain intended use.

Fairness

Different conceptualizations of fairness exist in the literature. On the one hand, it is closely connected to the notion of objectivity (see Chapter 10.1.2), when fairness is described as the scoring of a test independent of gender, race or ethnicity (Jürgens & Lissmann, 2015, p. 133). The issue of comparability is also of concern (Jürgens & Lissmann, 2015, p. 133). Variations of tests that are administered to the same sample, such as parallel tests used to prevent cheating, or RobS with different actors, should still enable participants to perform comparably – differences in the test design should not influence the test takers potential performance. On the other hand, test takers should perceive the test format as allowing them to show their actual competence and what they have learned (Sambell et al., 1997, p. 360). In that sense, students in higher education often also perceive assessment as fair if it "seemed pertinent to the everyday world" (Sambell et al., 1997, p. 361), echoing the notion of authenticity (see below). This idea of fairness, thus, more closely resembles validity, in that students want their competences to be measured in contexts that actually require them to show their capabilities (Sambell et al., 1997, p. 362).

Authenticity

As a criterion to evaluate a test, authenticity looks at the degree to which the test replicates real-life situations that would actually demand the tested competences, especially when exercising the later profession (Gulikers, Kester, et al.,

2008, p. 172; W. Stadler & Kremmel, 2018, p. 49). As an individual criterion, authenticity has strong connections to fairness, as perceived by students (see above) and thus also validity (see Chapter 10.2). It should be noted that authenticity as a concept is closely related to fidelity, a relation that is discussed in more detail in Chapter 11.2.2. Additionally, authenticity may also influence learning behavior, as discussed in Chapter 4.4.2.

While these secondary criteria offer an interesting perspective into specific aspects of test evaluation, they individually do not necessarily allow any substantial inference regarding a test's quality, rather, they amend or serve as subsets of validity as the primary focus of test evaluation, as evident in the overlap in criteria descriptions. The primary criteria are thus of more direct relevance for the test.

10.1.2 Primary Criteria: Objectivity and Reliability

A trifecta of criteria is referred to when the focus is on primary criteria: objectivity, reliability, and validity (W. Stadler & Kremmel, 2018, p. 39). These three terms will be explored in more detail, with a distinct focus on validity and its conceptualizations.

Objectivity describes the degree to which processing and scoring of a test is dependent on the person who is conducting and scoring the test (W. Stadler & Kremmel, 2018, p. 40). The main aim of a test is, thus, "value-free measurement" (Wilkinson et al., 2003, p. 219), i.e., no inference from the person making the assessment. Objectivity can refer to the processing, evaluation and interpretation of a test, depending which facet of testing is in focus (Jürgens & Lissmann, 2015, p. 128; Lienert & Raatz, 1998, pp. 7–8).

On a basic level, the idea of reliability (and to a certain degree also that of objectivity) is that there is a *true value* that a person has achieved in a test and an *observed value* that is assigned to the test (Yan & Fan, 2021, p. 479). Ideally, both would align perfectly, but realistically, there usually is an error of measurement that leads to either over- or underestimation of the true value. As reliability seeks to describe the degree to which a test (re-) produces identical or similar results, it tries to examine the degree of measurement error within the test, testing or scoring process (Yan & Fan, 2021, p. 479). In order to distinguish reliability from other constructs, like objectivity, terms such as "consistency" (Yan & Fan, 2021, p. 479) or "precision" (AERA, APA, NCME, 2014, p. 33) are used to describe it. While objectivity, however, focuses more on the individual who is responsible for the test procedure and scoring, the concept of reliability is more expansive and also refers to the stability of answers of test takers over a period of time (Yan & Fan, 2021, p. 479), as well as the test's internal consistency (AERA, APA, NCME, 2014, p. 33; W. Stadler & Kremmel, 2018, p. 43). To measure that notion

of reliability, test-retest, parallel-form or split-half reliability procedures are administered⁹² (Wirtz & Caspar, 2002, pp. 17–18; Yan & Fan, 2021, pp. 479–481). Coefficients like Cronbach's α are calculated to determine the internal consistency of a test as an aspect of reliability, (Wirtz & Caspar, 2002, p. 17; see also Chapter 15.4.1.1.2). In tests that require human judgements, reliability is often considered from the perspective of raters: Either as *intra-rater reliability*, i.e., the consistency of scores of the same rater on the same performance on different occasions, or as *inter-rater reliability*, i.e., the agreement or consistency of different raters' scores of the same performance (Yan & Fan, 2021, p. 482). The last aspect closely resembles objectivity, as both criteria try to estimate the influence that the rater has on scoring, highlighting the blurriness of the criteria overall. The individual requirements for reliability thus depend on the nature of the test. Reliability measured through test-retest, parallel-form or split-half reliability procedures is mostly relevant for multiple-choice or other similar close tests as they allow the assumption that the scoring itself is "perfect" (Hughes, 2003, p. 43). The potential error of measurement therefore lies within the test itself. The evaluation of more complex, open tasks, such as providing appropriate feedback to a learner, requires a different approach, as the given answers also require human interpretation and judgement. Thus, the potential measurement error lies within the rater(s).

Recently reliability has, though undisputedly relevant, been described as possibly only a "necessary but insufficient condition of validity" (Yan & Fan, 2021, p. 477). Validity has also been discussed as the overarching criteria for arguing test quality (AERA, APA, NCME, 2014, p. 11), subsuming the criteria of objectivity and reliability. Consequently, the next chapter will provide a deeper discussion on the concept of validity and its history.

10.2 From Validity to Validation

In its simplest form, validity has traditionally been described as the degree to which a test measures what it is supposed to measure (Chapelle & Lee, 2021, p. 17; Kline, 2015, p. 4; Moosbrugger & Kelava, 2020, p. 30). Consequently, validity could be regarded as a property of a test which persists through time and context, regardless of the purpose for which the test is used. However, this notion has subsequently been called into question. In 2014, the Joint Committee on the *Standards for Educational and Psychological Testing*, which consists of members of the *American Educational Research Association* (AERA), the *American Psychological Association* (APA), and the *National Council on Measurement*

⁹² See also Döring and Bortz (2016, p. 444).

in Education (NCME) proposed an understanding of validity that reflects that changing sentiment, focusing more on the interpretation of a test score for a specific use. In their *Standards for Educational and Psychological Testing* they state the following: “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests” (AERA, APA, NCME, 2014, p. 11). Following the approach laid out by the Joint Committee, validity is focused on the test score and the inferences made based on that score with respect to the intended use of the test. Furthermore, due to its complex nature, validity can hardly be achieved by a single indication; rather, a sufficient amount of sound and diverse kinds of evidence must be gathered so that the proposed interpretation of test scores is appropriately supported for the intended use (AERA, APA, NCME, 2014, p. 11). Validation is, therefore, a process of finding theoretical and empirical evidence and evaluating such in light of the proposed interpretation and use. The coherent formulation of that evaluative discussion is the validity argument (M. T. Kane, 2006, 2013a). This also includes critically reflecting and contextualizing potential limitations of the argument (Döring, 2023, p. 96). In order for a validity argument to have a robust foundation it should “[integrate] various strands of evidence into a coherent account” (AERA, APA, NCME, 2014, p. 21). In the following, these different strands of evidence will be elaborated on individually by looking at the diachronic development of the concept of validity⁹³. This approach is warranted as the aspects in discussion are still relevant to, and infused in most recent understandings of validation, including the proposed argument-based approach to validation.

10.2.1 Traditional Approaches to Validity

Traditionally, three types of validity are distinguished⁹⁴: criterion validity, content validity, and construct validity (Cronbach & Meehl, 1955, p. 281; Messick, 1989, p. 16, 1995, p. 741). Criterion or criterion-related validity (Hughes, 2003, p. 27), used to be the “gold standard” (M. T. Kane, 2006, p. 18) for validity. In order to estimate criterion validity, an external, separate measurement was taken by the same individual and correlations calculated (Messick, 1989, p. 17). The score achieved on the additional test was thought to represent the *true score* and validity was achieved if correlation between the two tests was high (M. T. Kane, 2006, p. 18). If the test were to be taken at about the same time, the term

⁹³ For a comprehensive historical overview on the development of validity, see M. T. Kane (2006).

⁹⁴ Cronbach and Meehl do mention four types, as they explicitly distinguish between concurrent and predictive validity (1955, p. 281).

concurrent validity could be used; if the two tests were administered at substantially different times, and the purpose of the first test was to predict the result of the second test, one could use the term *predictive validity* (Hughes, 2003, pp. 27–29; M. T. Kane, 2006, p. 18). For instance, learners' grades could be used as an external criterion to validate achievement tests in school (concurrent validity) or to predict learners' success in higher education (predictive validity). While relatively simple to calculate, the main problem of a limited approach that only considers criterion validity is that the strength of the argument is solely dependent on the separate criterion that the test's score is measured against; especially the question of how the separate criterion or the measurement of it could be validated is critical (M. T. Kane, 2006, pp. 18–19). For instance, the validity of grades has long been questioned, putting into doubt the suitability of that criterion (Ingenkamp, 1977). Also, as in the case of predictive validity, a substantial amount of time might be necessary between test and external criterion measurement to actually allow for any inference to be made, so that an evaluation cannot be made swiftly.

Content validity is more concerned with the actual substance of the test, and whether the items of a test are actually suited to depict and consequently measure the facets of the underlying concept that the test is supposed to measure. While traditionally this referred to content in the narrower sense of subject knowledge (M. T. Kane, 2006, p. 19), the term has also been used to refer to latent constructs (Meinhardt, 2018, p. 14). In order for a test to be considered valid in terms of content, all its items should be relevant to the construct, and no irrelevant items included in the test. This does not necessarily mean that the items in a test need to cover all the facets of the construct; however, the items should represent the construct to an adequate extent (W. Stadler & Kremmel, 2018, p. 44).

In order to achieve this aim, a primarily theoretical point of view needs to be taken initially. First, the construct itself needs to be argued thoroughly and in detail (Meinhardt, 2018, p. 14). This includes a theory-oriented framework and discussion about in- or exclusion of adjacent aspects, sharpening the edges of the construct. In order to evaluate content validity, the items are then critically evaluated against the construct. Additionally, expert judgements can be used to either evaluate the construct itself or the item-construct relation (Bartels, 2018, pp. 16–17). This is a necessary step in order to ensure that potential biases by the authors and creators of the test are reduced (M. T. Kane, 2006, p. 19). A type of content validity is curricular validity, which is especially relevant for tests developed in educational contexts. It describes the specific instance in which a previously outlined syllabus, framework or legal guideline spells out the learning

objectives and the test's items are developed to test these (W. Stadler & Kremmel, 2018, p. 44).

The importance of the (latent) construct in validation has already been addressed as a backdrop for discussing content validity. While content validity puts an emphasis on whether the content of the items refers to the described construct, *construct validity* emphasizes the overall test and its score and the degree to which it allows inferences to be made about the targeted construct (Meinhardt, 2018, p. 16). It was developed as a third validity aspect to address the limitations of criterion validity, i.e., lack of external criteria, and content validity, i.e., no available content area (Cronbach, 1971, pp. 462–463; M. T. Kane, 2006, p. 20). In order to investigate construct validity, determining the constructs that potentially cause variance in a test is necessary (Cronbach & Meehl, 1955, p. 282). Then, a set of hypothetical relations between constructs should be described and theoretically reasoned *a priori*, the so-called “nomological net” (Cronbach & Meehl, 1955, p. 290), and empirically investigated afterwards (Kline, 2015, pp. 6–7). This includes, on the one hand, the possible relation of facets within the targeted constructs, as well as, on the other hand, the possible relation of the construct with other constructs. Depending on the theoretically delineated relation and the strength of the correlation of the variables, the findings need to be discussed with regard to their robustness and their implications for construct validity. This initial conceptualization was disregarded as “the weak sister” (Shepard, 1993, p. 416) of criterion validity because it was seen only as an inadequate substitute for instances in which a proper criterion was lacking (M. T. Kane, 2006, p. 20). However, it quickly evolved and has, for a substantial period of time, been the focus for validation studies as it was seen as an overarching validation concept that not just supplemented criterion and content validity, but much rather “embraces all forms of validity evidence” (Messick, 1989, p. 17). Messick's comprehensive (re-)conceptualization of construct validity (1989, 1995) popularized it and thus warrants a distinguished look (Newton & Shaw, 2014, pp. 22–23).

10.2.2 Messick's Approach to Construct Validity

Based on the criticism of focusing on only criterion and content validity to argue test quality, Messick proposed a comprehensive reconceptualization of validity under the approach to construct validity. He states the following: “In essence, construct validity comprises the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables” (Messick, 1995, p. 743). With this notion, validity no longer is a property of a test, but an accumulation of evidence that indicates that a proposed interpretation of test

score is warranted. Validity is to be understood as a “unified concept” (Messick, 1995, p. 744), in that, all the evidence needs to be taken into account. Nevertheless, different nuances of that concept can still be classified which, in turn, provide a different type of evidence to potentially support the test score interpretation. Messick distinguishes six different aspects that, to a certain degree, build upon the already established concepts.

Content validity

While the previous comments on content validity are included in Messick’s broad discussions, he also introduces the idea that the format of the test, not only the items itself, should have relevance for the domain (Messick, 1995, p. 745). The domain and the actions that need to be mastered for that domain should be reflected in the construction of the test and the items. It therefore also extends to the outlined secondary criteria of fairness and authenticity (see Chapter 10.1.1). If a test were, for example, to test teachers’ competence on feedback, it should be considered how feedback is potentially provided in the profession, e.g., in written comments, in oral discussions, or in brief corrective forms.

Substantive validity

The substantive component of validity aims at the cognitive processes that lead to a test performance and score (Leuders, 2014, p. 10). This facet thus expands on content validity, in that it looks at the processes that are representative of a certain domain, not only the content (Messick, 1995, p. 745). For example, if a test that was to measure teachers’ feedback competence only has participants correct errors in a text, participants would not necessarily use the cognitive structures necessary for feedback, i.e., providing information to the learners on how to improve on these errors. Thus, such a test does not necessarily reflect the cognitive processes that the target domain requires and are thus potentially lacking of substantive validity. In order to examine test takers’ cognitive processes, introspective methods such as think-aloud procedures, or, as a retrospective approach to introspection, stimulated recall interviews (SRIs) are suitable (Aguado et al., 2018; Messick, 1995, p. 745).

Structural validity

The structural facet of construct validity focuses on the scoring instrument. The scoring model should reflect the construct which is to be measured, i.e., the development of scoring rubrics and criteria should take the construct structure into account (Messick, 1989, p. 43, 1995, pp. 745–746). A test on lexico-grammar that is, for instance, conceptualized according to the facets of the mental lexicon, i.e., form, meaning, and use (Hutz, 2017), should not only have items that measure each facet, which would be necessary for content validity, but the scoring model

should also reflect those facets, potentially providing scores for each facet. Also, items from different facets might have a different degree of difficulty and the scoring model should also represent these differences (Leuders, 2014, pp. 10–12; Messick, 1989, p. 43). Structural validity can, for example, be examined by comparing the dimensions of a proposed construct and the empirically-obtained test scores (Leuders, 2014, p. 18) or looking at the internal structure using inter-item correlations and factor analysis (Shepard, 1993, p. 418).

Generalizability

The aspect of construct validity that refers to generalizability is concerned with the question of “the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks” (Messick, 1996, p. 7). It is conceptually related to aspects of external validity (see below). This notion addresses the concern that the score obtained in a test can only really be interpreted in terms of the items used and not to make broader inferences (Messick, 1995, p. 746). Usually, the generalizability of a test score is researched by means of correlation to other constructs that either represent the target domain in full or partially so that the boundaries of generalization can be mapped out (Messick, 1995, p. 746). For instance, one could look at the hypothetical correlations between tests on listening, writing and general language skills. While the two skills could correlate with the general language skill score, listening and writing might correlate to a much lesser extent with each other. Thus, while both subdomains allow for inferences to be made towards the broader notion of general language skills, they do not necessarily allow the generalization that a high score on the listening test means that the test person is equally well-skilled in writing. An important aspect that must be reflected on is that an experimental or laboratory setting in which tests are administered potentially influences the generalizability for a natural setting of the domain (Messick, 1989, p. 58). Generalizability is also dependent on the generation of the test score, as only reliable test scores can be used to make broader arguments (Messick, 1995, p. 746). It further requires the tasks of the test to be representative of the domain, which can also be framed as “task generalizability” (Messick, 1989, p. 57).

External validity

External validity, in its essence, looks at the empirically found relationship that the test score has to other variables and compares it to theoretically delineated relationships (Messick, 1989, p. 45, 1995, p. 746). In that sense, external validity can either be convergent, i.e., a (substantial) correlation between variables is assumed and empirically found, or discriminant, i.e., no, or only a low degree of, correlation is expected and reflected in the data (Messick, 1995, p. 746).

Conceptually, this facet is closely connected to the notion of generalizability if the external criterion is meant to represent a broader frame of reference (see above). This aspect, however, is also more focused on “nonassessment [sic!]” aspects like experience or types of degree (Messick, 1995, p. 746). Furthermore, this facet is similar to the previously outlined criterion-validity, which sometimes leads to a synonymous use of the two phrases (Meinhardt, 2018, p. 15) and the subsumed subsets (discriminant vs. convergent; concurrent vs. predictive) (Bartels, 2018, p. 18). What might also distinguish the terms is the use of the collected information. Criterion validity was regarded as the singular source of validity for a long time, while in Messick’s perspective, information that contributes to the external validity is only one part of the overall evaluation of a test’s construct validity. It is thus the purpose of the comparison that might offer a way to differentiate the relatively closely-connected aspects of validity. This approach is also taken in the *Standards for Educational and Psychological Testing*, in which these similar aspects of validity are all discussed under the banner of “Evidence Based on Relations to Other Variables” (AERA, APA, NCME, 2014, pp. 16–19).

Consequential validity

The last aspect refers to the extent to which the interpretation and use of the test score has (un)desired consequences (Messick, 1995, p. 746). Messick thus goes beyond the test construction and scoring, and includes the potential effects that testing and scoring has on the individual that has been tested. Thus, the potential consequences of the test should be considered in the construction of a test and the interpretation of the score should always be reasoned against the backdrop of the intended use. With that in mind, it is important to prohibit test results from being harmful to individuals, not only due to unintended score interpretation, but also due to potential test invalidity, highlighting the necessity of careful deliberations when it comes to the test’s construct validity (Messick, 1995, p. 746). Consequential validity, however, also extends beyond the individual and also includes the impact that test results from large-scale assessment studies, such as like PISA, have on educational governance (Leuders, 2014, p. 14). Moreover, consequential validity would be in question if a test is designed to be used for diagnostic purposes, but is actually used to assign a summative grade. Conversely, consequential validity would be higher if, based on the findings in formative assessment, appropriate measures are taken to support the learners.

Messick’s approach, in contrast to the previously outlined views on validity, aligns more with the cited introductory statement made by the Joint Committee (2014). Messick advocates for a multi-perspective evaluation of the test’s inner structure and content, the generation of test score and the consequences that a test score might have, which is appropriate considering most test’s complexity.

The last point, consequential validity, however, has been discussed controversially in research, not because the consequences of a test are disregarded, but rather because of the complexity of obtaining data and evidence to support this aspect (Meinhardt, 2018, pp. 20–21). While well-intentioned, researchers have found it difficult to meet Messick's high standards and question the argument that the consequence of a test needs to be considered for validity. In a special issue, which is fittingly titled "The Great Validity Debate," Mehrens pleads "that the authors of the revision to the Standards [for Education and Psychological Testing] recognize the debate and relegate any discussion of consequences to a context separate from the validity chapter" (1997, p. 18). This plea is based on the concern that by including the consequences that the use of a test might have, the concept of validity becomes itself less clear. Mehrens, among others, like Popham (1997), therefore argues for a narrower definition of the term validity that excludes the consequence, without necessarily disregarding the discussion of them, just not as part of validity (1997, p. 18). Based on the two contrary positions, the discussion is also framed as a "meaning vs impact" debate (Lederman, 2023, p. 244). As far as the most recent edition of the Standards for Education and Psychological Testing (2014) is concerned, they do not take a stance in either direction, as their definition of validity seems to favor impact (i.e., consequence), while the closer operationalization seems to favor meaning (i.e., content or construct) (Lederman, 2023, p. 248).

Another related issue of Messick's conceptualization is that of practicality. Messick himself concedes that the variety of potential evidence and data to support construct validation cannot feasibly be obtained by researchers, stating that *relevant* data for interpreting test scores need be obtained (Messick, 1989, p. 50). He argues that, at the end, the amount of evidence should suffice to argue for or against a specific interpretation, and the selection of evidence should be reasoned plausibly (Messick, 1989, p. 50). Kane, who is generally in favor of Messick's inclusive approach to validation, questions the approach that the sheer amount of data without a consistent framework of argumentation concerning its selection is sufficient, asking: "[How] much evidence is needed to adequately support a proposed interpretation or use?" (M. T. Kane, 2013b, p. 8). Trying to consolidate the discussion on consequential validity and practicality, Kane has proposed and continuously developed an "argument-based approach to validation" (1992, 2006, 2013a), which tries to reconcile the ongoing debate in validation. This approach seeks to make the case for a well-reasoned decision on which aspects to include or exclude when considering a test's validity. It most closely reflects the stance taken by the Joint Committee and thus serves as the foundation for this thesis and will be explored in the subsequent section.

10.2.3 The Argument-based Approach to Validation

Kane's argument-based approach to validation draws upon the previously outlined aspects and concepts of validity and adds to that the demand for providing a clear framework in which the data is discussed (M. T. Kane, 2013a, pp. 450–451). The approach is two-tiered: First, test developers need to state a clear, theoretical line of argumentation that makes claims aligning test development with test use, the so-called *interpretation/use argument* (IUA). Secondly, evidence is obtained and discussed against the backdrop of these claims, the so called *validity argument* (VA) (M. T. Kane, 2013a, p. 448). The VA seeks to discuss the eligibility of the claims made in the IUA. The focus of the validation process is, thus, the evaluation of these claims.

The Interpretation/Use Argument – Making the Claims

The purpose of the IUA, which Kane referred to as an *interpretative argument* in previous work (M. T. Kane, 1992, p. 527, 2001, p. 329, 2006, p. 23), is to “state what is being claimed” (M. T. Kane, 2013a, p. 451). The claims lay out the necessary conditions for considering a test's validity. The IUA should include those claims that are relevant to appropriately interpret the score of a test for a certain use. Within the argument-based approach, not all possible aspects of validity need to be considered, but it has to be made transparent which claims are made. The relevance of the individual claim for the overall interpretation also indicates the required strength of evidence or backing that the warrant needs (M. T. Kane, 2013b, p. 12). Thus, not all claims need extensive backing, but the reasons have to be stated in the IUA. The IUA should start with a formulation of the intended purpose of test use and test score interpretation, as should any validation process (AERA, APA, NCME, 2014, p. 11). The consequences of test use are thus embedded in the argument-based approach and must be taken into account from the beginning (M. T. Kane, 2013a, p. 452). The IUA can also be viewed as a “Mini-Theory” (M. T. Kane, 2006, p. 25) as it describes a framework for interpreting obtained data. It should, consequently, be formulated in a way that it is “clear, coherent and complete” (M. T. Kane, 2013a, p. 451). The IUA has multiple claims, each being a requirement for the succeeding claim. A useful model to conceptualize the connection between evidence and claim is Toulmin's layout of arguments (Toulmin, 2003, p. 94), or as Kane put it, the model of inference⁹⁵ (M. T. Kane, 2006, p. 27; 2003, p. 97) (see Figure 18).

⁹⁵ For coherence, Kane's terminology will be referred to in this thesis.

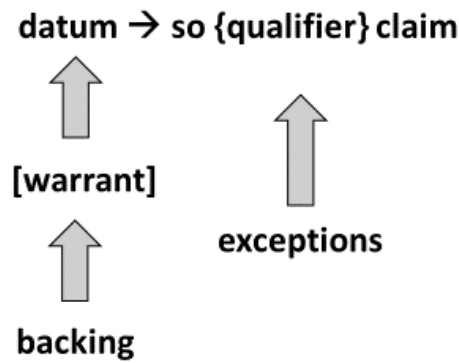


Figure 18: Toulmin's model of inference (M. T. Kane, 2006, p. 28).

An inference is structured by stating a datum, or fact, which leads to a (qualified) claim. The claim is based on a warrant, i.e., a rule, which logically explains why the fact can be used to make the claim. This warrant needs to be backed by evidence, which supports the rule to be made. While warrant and backing might logically explain the causality between datum and claim, there often are exceptions to that rule, which should also be addressed in order to properly contextualize the whole inference. The likelihood with which the claim can be made based on the datum, is stated by a qualifying operator (M. T. Kane, 2006, pp. 27–28; Toulmin, 2003). Consequently, an inference is the result of the evaluation of the backing with regard to the claim. One example of an inference using the model is as follows: The sidewalk is wet (datum) so you claim that it must have rained. Based on your personal experience and observations (backing) you know that usually, after it has rained, the sidewalk is wet (warrant). However, you know that your neighbor sometimes likes to use lawn sprinklers, which also leads to a wet sidewalk (exception; rivaling hypothesis), thus you come to the preliminary conclusion that it has likely (qualifier) rained. The purpose of the IUA is then to state the claims, weigh their relevance for the argument and state the kind of evidence that is necessary to support the warrant. The degree to which the evidence then supports the claims is discussed separately in the validity argument.

The Validity Argument – Evaluating the Claims

The IUA itself “cannot be proven” (M. T. Kane, 2013b, p. 14), only evaluated. This evaluation is done in the validity argument:

The validity argument evaluates the plausibility of the proposed interpretations and uses of the test scores. The proposed interpretations and uses can be considered valid to the extent that the IUA is complete in the sense that it specifies the proposed interpretation and use of the test scores, that it is coherent in the sense it hangs together, and that its inferences are either inherently plausible or are supported by appropriate evidence. (M. T. Kane, 2013a, p. 452)

Consequently, the IUA should ideally stand up to scrutiny of the outlined criteria, which are usually referred to as clarity, completeness, coherence and plausibility (M. T. Kane, 1992, p. 528, 2013b, p. 10; Meinhardt, 2018, p. 24):

- 1) Clarity: The inferences that are made should be formulated precisely and unambiguously.
- 2) Completeness: The claims should include all the relevant aspects that are necessary to make interpretations for the intended use.
- 3) Coherence: The IUA should be coherent to the extent that the inferences made are plausible and logically linked to each other (M. T. Kane, 2013b, p. 14).
- 4) Plausibility: The individual warrant for the claims within the IUA should be corroborated with appropriate backing that supports the conclusions that are drawn, i.e., the inference made must be plausible. This includes theoretical support as well as empirical evidence.

Simply put, the IUA should be logically constructed, the single inferences articulated unambiguously, and the connections between the inferences logical. Importantly, inferences should be based on evidence, are reasonable and cover a variety of perspectives. The type of inference indicates what kind of evidence is required. While Kane has changed terminology with regard to the kinds of inferences, there are four consistent different kinds which make up the core aspects⁹⁶: Scoring, generalization, extrapolation and implication (M. T. Kane, 2001, p. 330, 2006, p. 34).

Scoring inferences⁹⁷ describe the portions of the IUA that are concerned with the process of turning an observed performance into a score. This includes, but is not limited to, the adequacy of the criteria used to score and the consistency with which scorers use the scoring rules (M. T. Kane, 2006, p. 34) and are thus reminiscent of reliability and objectivity aspects, as the consistency and bias of scoring is addressed (M. T. Kane, 2013a, p. 454). Scoring inferences can be evaluated by using experts to judge the criteria, which is also similar to approaches for evaluation of content validity, or by having different raters use the scoring rules and examine the results with regard to the consistency and agreement.

Generalization inferences focus on the extent to which the score of a test can be used to make assumptions about the trait that is aimed at being measured. For example, developers of general language skills tests might argue that their score can be generalized to make assumptions on a person's language skills and

⁹⁶ Occasionally, Kane also added "theory-based inferences" (2013a, p. 455).

⁹⁷ Kane (1992, p. 529) uses the phrase "observation"; Kane (2001, p. 330) uses the phrase "evaluation" (see also Meinhardt, 2018, p. 22)

that the test score is indicative of that. Thus, the items of the test should be representative of the universe of possible items for the content, alluding to notions of content validity. Furthermore, the sample size should suffice so that sampling errors can be excluded, and the condition of the test procedure should reflect the actual domain (M. T. Kane, 2013a, p. 454).

Extrapolation inferences are made if the test score is meant to predict the performance in other, more or less adjacent, domains (M. T. Kane, 2013b, p. 11). Such inferences can be evaluated analytically by comparing the relationship of the domains, or empirically by correlating test scores to scores from the target domain (M. T. Kane, 2006, p. 35). This inference resembles criterion-related and/or external validity (M. T. Kane, 2006, p. 36), in that the comparison of the test score to a separate measure is the focus.

Implication inferences⁹⁸ are relevant when the test score is used to make specific decisions (M. T. Kane, 2001, p. 330). The rules that guide the decision are evaluated based on the consequences that they might have for the individual test takers. Generally, unintended negative consequences should be avoided, but more importantly, the task developer is not responsible for how the test is used. Rather, the potential consequences that arise from the intended use must be reflected. For instance, placement tests have immediate consequences for the test takers' futures as they are (at least partially) used to decide whether a learner may take a specific course. The allocation-result rule should thus not harm the test takers by falsely allocating the person based on invalid results. This inference type resembles what Messick understood as consequential validity.

As the validity of the test relies on evidence and the interpretations made based on it, evidence that refutes these claims naturally has an impact on the validity of the test. Considering contradicting evidence and rivaling hypotheses that contextualize and qualify the claims is an important aspect of the validation process. It potentially strengthens the overall argument of validity as it indicates a more objective evaluation of the supporting evidence and provides clearer guidelines for the interpretation of the test score (AERA, APA, NCME, 2014, p. 12). Moreover, if new evidence is generated, which does not support, or even contradicts the previously made claims, the validity argument has to be reevaluated. Evidently, this reflects a statement made by Cronbach: "[Validation] is never finished" (1988, p. 5). Validation is thus an open-ended process of evaluation and always preliminary.

The supposed advantage of Kane's approach compared to the previously outlined concepts of validity is that researchers must not necessarily consider *all*

⁹⁸ M. T. Kane (2001, p. 330) refers to this inference as "decision making", M. T. Kane (2013a, p. 455) as "score uses".

aspects of validity as indicated in Messick's framework. Rather, a plausible argument in the form of the IUA has to be made, which argues why a certain inference is necessary to argue a test's validity and which kind of evidence is required to support the respective claim (M. T. Kane, 2006, p. 23). Thus, the selection of evidence becomes more transparent. Furthermore, it makes the argument for a test's validity simpler, as the claims are evaluated, not the test directly. As a test is only considered valid to the extent that the provided evidence supports the claims that are made in that IUA, claims that are not made, do not factor into that discussion as long as the IUA is still plausible, complete and coherent (M. T. Kane, 2013a, pp. 448–449). This gives test developers some freedom as they, for example, do not have to consider aspects of, or provide evidence for, consequential validity in their claims, as long as sufficient evidence that supports the claims that have been made is provided. Thus, Kane's approach is more pragmatic and takes into account the limitations that test developers encounter. On the one hand, it limits the extent of a test's validity to the claims that have been made, but on the other, it also lessens the burden of considering all potential aspects of validity.

The juxtaposition of Kane and Messick's notions to validity has shown that at the core, both approaches are almost identical in which arguments they understand to be important for validity. For example, the notion of content validity or the comparison of the test score in question and other relevant test scores are found in both. The difference is that Kane focuses on the reasons why each element is relevant in the overall argument for validation, while Messick does not necessarily require the framework of reasoning behind the choices of inferences. In order to evaluate the IUA and come to a validity argument, one should consider its clarity, completeness, coherence, and plausibility. Evidence must be obtained that can be used to back the specific warrants of claims. The kind of evidence required depends on the relevance and type of inference. Contradicting evidence and rivaling hypotheses should be included as they strengthen the validity argument overall.

10.3 Implications for the Validation Process

Based on the presented approaches to validation, the dominant tendencies that are indicative of the understanding of validation in this thesis will be summarized briefly in bullet points:

- Validity can be understood to describe overall test quality, subsuming other aspects like reliability and objectivity.
- Validity is not a property of a test. Rather, validation refers to the extent to which there is evidence that indicates that a test can be used for an intended

use. This use must be made explicit at the beginning, as it is the foundation for the validation process. Validation is thus an evaluative process.

- Validation requires making transparent the inferences that would allow assumptions that the test is appropriate for the intended use. An inference consists of a claim, which is based on a warrant, which, in turn, is supported by evidence. Each inference is a predisposition for the subsequent inference. This results in the chain of inferences, the IUA.
- The validity argument, then, is the evaluation of the IUA. Evidence is provided that purports to support the warrant and thus substantiates the claim. The strength of the evidence as well as the discussion of possible limitations, contribute to the overall robustness of the VA.
- Validation processes need to consider many types of evidence. The relationship of the evidence strands is not of an alternating nature, but more of a supplementary one (Cronbach, 1971, p. 445).
- The argument for validity cannot and is not required to consider all potential sources of validity and is thus necessarily limited. The transparency of the IUA and the claims that are made is thus made even more important.
- The validation process is open-ended. Consequently, it requires the test developers to include and discuss counter hypotheses and balance evidence that supports, and does not support, the claim. Overall, this strengthens the VA as this provides a clearer picture for the potential use-cases of the test. Furthermore, the validity of a test can only be argued against the evidence that is available. If new evidence is available, the VA needs to be reevaluated against this backdrop. Results of a validation process are thus always preliminary.

In order to evaluate the IUA and come to a validity argument, evidence must be obtained that can be used to back the specific claims. In the following, the IUA for the RobS will be laid out by stating the individual claims, as well as the warrants and the kind of backing required for support.

11 IUA for the RobS

The IUA for the RobS is based on the understanding of the argument-based approach to validation by Kane (1992, 2006, 2013a). As a structuring device for the IUA, the classification system for validation by Schreiber and Gut (2022) will be presented and adapted to the validation of the RobS.

11.1 A Classification System for Validation by Schreiber & Gut (2022)

A useful and hands-on classification system for validation processes, which is situated in the argument-based approach to validation, is proposed by Schreiber

& Gut (2022). Based on the works of prominent researchers in the field of validation, e.g., Cronbach, Messick and Kane, the authors propose a thorough framework, which classifies the different types of claims that could possibly be made in an IUA (Schreiber & Gut, 2022, p. 134). The authors preface that their system is not necessarily exhaustive and state that it is, thus, not intended as a strict roadmap (Schreiber & Gut, 2022, p. 134). Still, it allows researchers who follow a similar understanding of validation to plan their research and to organize their claims for an IUA and VA (Schreiber & Gut, 2022, p. 135). Depending on the needs of the test and its developers, some aspects in the classification system may be excluded, others more emphasized or some criteria of validity rearranged within the system (Schreiber & Gut, 2022, p. 135). In their system, as depicted in Figure 19, the authors list warrants for the claims as causal subordinate clauses, which must be supported by appropriate evidence in the VA (Schreiber & Gut, 2022, pp. 4–6). For reasons of brevity, the focus will be on the structure of the system as the warrants are further explicated in the subsequent chapter, in which the IUA for the RobS is presented.

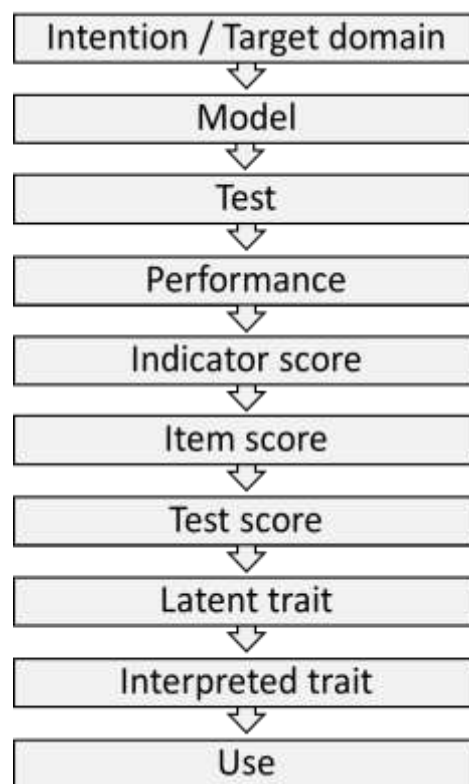


Figure 19: Framework for an argument-based approach to validation according to Schreiber and Gut (2022).

In the first step, the **intention** of the test and the **target domain** in which the test is situated must be stated. This is the necessary foundation for all the following steps, as they are evaluated against the target domain and intention described.

The second step is the **modeling of the construct** that one intends to measure within that target domain. Modeling is necessary in order to explicate the different facets of the target domain that are being measured. It also allows for a later evaluation of the model with the empirical evidence. In order to create a model, researchers should take into account existing norms as well as previous theoretical and empirical research (Schreiber & Gut, 2022, p. 142).

The third step looks at the construction of the **test** (materials, items, etc.), which is based on the previously outlined model. The items that constitute the test, as well as the general design of the test, should align with the demands of the target domain and be appropriate for the sample (Schreiber & Gut, 2022, p. 142).

The fourth step is concerned with the **performance** of the participants using the test. The performance of each participant should be conducted in a standardized manner and it must be ensured that the performance is representative of the target domain (Schreiber & Gut, 2022, p. 143).

The fifth step focuses on the translation of performance into **indicator scores**. Indicator scores represent the smallest unit of a performance evaluation and should thus comprehensively and holistically reflect the facets of the performance. Indicator scores are later aggregated into item scores, as described below. If performance or a facet of performance is evaluated singularly as an item, there are no indicators and this step must be considered for item scores (Schreiber & Gut, 2022, p. 143).

The sixth step is the aggregation of indicator scores into **item scores**. The reduction of indicators is necessary for an analysis of meso-level construct aspects but should not diminish the representation of global performances. If there are no indicator scores, this step is not necessary to consider (Schreiber & Gut, 2022, p. 143).

The seventh step focuses on the summation of item scores into a **test score**. While the previous steps apply to individual items, this step looks at the extent to which the summations of item scores for different individuals result in test scores which are representative of the sample (Schreiber & Gut, 2022, pp. 143–144).

The eighth step is the extrapolation of a test score to describe a **latent trait**. The final test score is usually used to describe the extent to which a participant fulfills aspects of certain latent traits. Thus, the situation in which the test score is generated should be representative of the target domain (Schreiber & Gut, 2022, p. 144).

The ninth step is the **delineation of an interpreted trait**, which is made based on the described state of the latent trait. In order to use the test score to make interpretations about a latent trait, the empirically derived data must

comply with the theoretical model; a mismatch would indicate that the test score and the respective latent trait description cannot be used for further interpretation (Schreiber & Gut, 2022, p. 144).

The tenth step focuses on the implications that the **use** of and participation in the test has. As validation depends on the specific use and interpretation of a test score, the actual use should align with the previously stated intended use by the stakeholders and not be harmful for the test taker (Schreiber & Gut, 2022, p. 144).

11.2 Making the Claims: The IUA and Sources of Evidence for the Test

Based on that classification system, the proposed IUA for the RobS is presented as follows (for the evaluation of the IUA, i.e., the VA see Chapter 17).

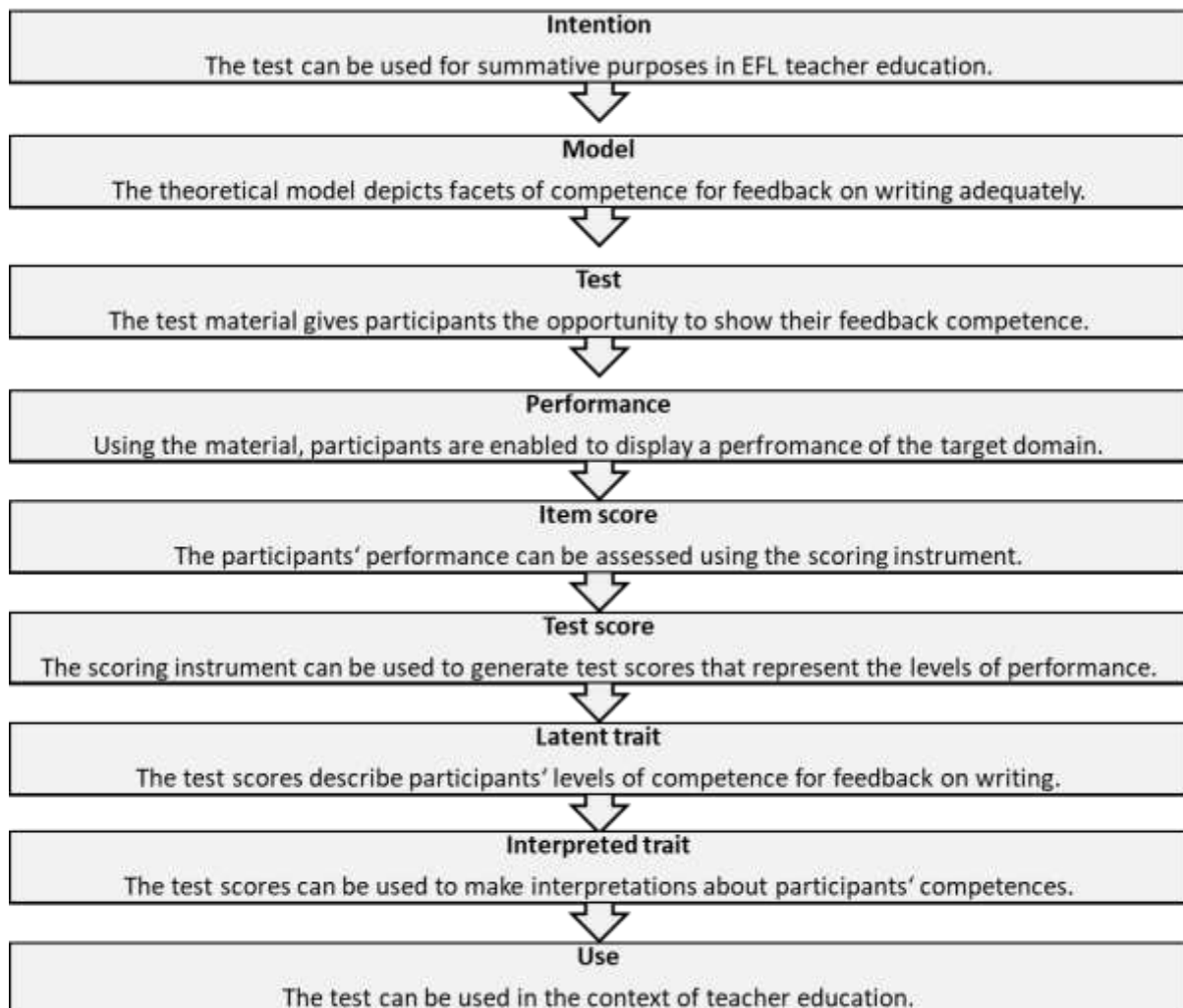


Figure 20: Overview of the IUA.

Intention

The test, as presented in Chapter 8, is intended to be used as a summative assessment format for EFL teacher education in Germany, especially the first phase. Importantly, development focuses on a prototype, thus actual use for assessment purposes is not aimed at in this thesis. Its test score, which is generated using a scoring instrument based on a theoretically constructed model, is meant to represent the test takers' competence in providing feedback on a learner text in a one-on-one situation. The specific target domain in which the test is situated is that of EFL teacher competence. The test focuses on the notion of "shows how" (Miller, 1990, p. 63) and thus seeks to measure a performance-oriented aspect of the competence model as outlined by Blömeke et al. (2015).

11.2.1 Claim 1: Model

Claim 1: The theoretical model for feedback on learner texts depicts the required competence facets adequately.

Warrant: ... because it is based on relevant theoretical literature and sound empirical evidence. The model includes the relevant facets that are required for providing feedback on a learner text in a one-on-one situation and takes into account the specifications of the intention to evaluate performance-oriented aspects of competence. It refers to given normative settings, current empirical findings and sound theoretical guidelines in order to include multiple perspectives that are necessary to holistically describe the model (Schreiber & Gut, 2022, p. 131).

Evidence that supports this claim is the transparent description of the model. The model that is used is legitimized by the means of a detailed discussion, in which the selection of dimensions and facets is displayed differentially and critically reflected upon. This kind of evidence also refers to Messick and the content aspect of validity (1995, p. 745). Another approach to examine the extent to which the facets of the model are sufficient is the perspective of experts, whose statements can be used to judge the model's merits based on their exhaustive knowledge in the field (M. T. Kane, 2013a, p. 449). This claim is of more relevance for the test's validity, as it is the foundation for items, material design and scoring rules. Thus, it requires more extensive backing and a proper contextualization of evidence and reasons for inclusion and exclusion of certain facets.

11.2.2 Claim 2: Test

Claim 2: The test material gives participants the opportunity to show their competence of providing feedback on a learner text in a one-on-one situation. The general format of the test focuses on the notion of "shows how" (Miller, 1990,

p. 63) and the intention of measuring the performance-oriented aspect of competence (Blömeke et al., 2015).

Warrant: ... because the material, i.e., the set of documents, including the learner text and the verbal triggers aligns with the model and is appropriate for the sample (Schreiber & Gut, 2022, p. 131).

This notion is also reflected in Messick's conceptualization of content validity (Messick, 1995, p. 745), the secondary criterion of authenticity and, to a lesser extent, perceived fairness. The test includes two types of material:

- 1) the material designed for preparation and as a resource in the simulation (*a priori*) and
- 2) the material designed for use in the dynamic situation (*in situ*).

The first set of the material is provided to the participants in the fifteen-minute preparation time (M0-M5). This material is described in detail in Chapter 8.1 and is included in Appendix A. The second set of material refers to the situation, specifically the actors, whose depiction of the role of Mia can be viewed as test material. Similar to items in a written test, they act as a source in the test which the participants have to react to. Of primary concern are the following aspects:

- actors' behavior (verbal and non-verbal)
- verbal triggers made by the actors.

The actor training is detailed in Chapter 14.1.2 and the VTs are described in Chapter 8.2. The material of the test must withstand critical evaluation from different and potentially antinomic perspectives, which will be further elaborated on in.

Alignment of model and material

The material should allow for the participants to display the competencies that have been outlined in the model. The learner text especially, as it is the focus of the feedback conversation, should feature characteristics that enable the participant to show the facets described in the model. This claim can be supported by a transparent description of the construction and design of the learner text as well as an analysis of the text, pointing out the characteristics. In order to ensure that the characteristics are reliably found, inter-coding procedures of the learner text can be conducted. Another relevant perspective to consider is that of the participants. As subjects of the assessment, they experience the demands of the test first-hand. The requirements they perceive to have been tested on can thus be compared to the intended aspects outlined in the framework and scoring instrument. This is also relevant as to anticipate potential false expectation that PSETs might have when assessed via RobS.

Within the conversation, VTS are used by actors in order to elicit facet-specific reactions from the participants that show the extent to which a participant is able to fulfill that specific facet of the model (Dotger et al., 2010, p. 133). In that sense, VTs are comparable to items in a written test. One way to examine the extent to which the VTs actually meet the targeted facet of the model is to conduct think-aloud studies, for example, as stimulated recall interviews (Barras, 2018).

Authenticity of material

The material should be representative of the required demands of the target domain. Two approaches could be taken in order to examine this:

- 1) The test format and the material actually resemble real-life situations;
- 2) The test format and the material are perceived to resemble real-life situations by the test takers, i.e., they are perceived to be authentic.

The first approach would require a high degree of fidelity and an assumed objective reality with which the test is compared, e.g., a close replication of the real-life situation (Bland et al., 2014, p. 1112). The second approach requires that the situation is perceived to be authentic by those who participate in the test (Gulikers, Kester, et al., 2008, p. 173). For validation of this test, the second claim is more relevant. First, as a simulation-based assessment, the test construction is built on the paradigm of approximating practice (Grossman, Compton, et al., 2009). Therefore, it does not seek to resemble real-life and fidelity is not a necessary requirement for arguing the test's validity. Second, it is important that participants perceive the test to be authentic (Schreiber & Gut, 2022, p. 142). Participants' perception of the test is relevant for the extent to which it potentially influences their learning behavior (Gulikers, Kester, et al., 2008), not necessarily the fidelity (Bland et al., 2014; Chernikova et al., 2024; Dahlstrom et al., 2009) (see also Chapter 4.4.2). Hence, it suffices to provide evidence to satisfy the claim that the test takers perceive the test and the items as authentic.

Authenticity needs to be considered from multiple perspectives: The situation in general, the material provided in the preparation and the actors' behavior. The overall demand, providing feedback on a learner text in a one-on-one discussion should be a situation which the participants think is authentic, in that it is a situation that happens at school. This extends to the context of the situation, in which a learner actively seeks feedback and to the time constraints of ten minutes, with 15 minutes to prepare. Regarding the material, different perspectives of authenticity are relevant. Together, the task for the text, the excerpt of the series of lessons and the model text, should approximate material and practice from school. The task should be relevant for the specified learner group and preferably, also be part of the guidelines. The learning objectives stated in

the excerpt of the series of lessons should be constructively aligned and the task that is presented should fit into that plan. The model text should include features that are essential for a news report and easily identifiable for the learners, be appropriately adapted for the learner group and also fit to the overall topic and series of lessons. The learner text should be authentic, in that it features linguistic characteristics, like typical errors, error patterns and/or mistakes or other remarks that learners of that age and from Germany make when doing a task as proposed⁹⁹. Similarly, the actors also need to portray learners authentically. Most importantly, the statements they make should be phrased in a way that they are perceived to resemble learners from that age group. This is especially relevant for the VTs, as they are designed for the test and are not necessarily part of the natural conversation that occurs. This extends to the mode of language, specifically the choice of whether the learner speaks English or German, but also their tone, body language and facial expressions, as these components should all contribute to an overall natural depiction of the learner. In order to find evidence supporting the claim that the test material and situation are perceived as authentic, participants should be asked to state the extent to which they perceived the test and its different components to be authentic.

Balancing the demands

A challenge for the material design, and the discussion of validity, is that the demands of alignment of the model and material and an authentic perception of the material might be in conflict. For example, a VT might be used that elicits a facet from the model, while it might simultaneously be perceived as an atypical statement for a learner by many participants and reduce authenticity. While both notions individually are necessary to strengthen the validity argument, this potential antinomic relation must be considered in test construction and when discussing the evidence, as the aim is to find a satisficing balance (Ajjawi et al., 2024, p. 506).

This claim, as indicated by the more comprehensive description, requires more extensive backing. Considering the novelty of the test format, it is essential to provide backing for each of the aforementioned aspects so that the validity can be argued plausibly. The test format should allow participants to show their competence in a context that approximates practice. The material that is provided for that purpose must, on the one hand be perceived as authentic by the

⁹⁹ It must be noted, and will be in the limitations, that this approach necessarily reduces the heterogeneity and diversity of a classroom to certain, potentially stereotypical German EFL learners. It must be argued that this standardized learner is only one of a plethora of potential learners that could be depicted in such a scenario. The main focus is on the teachers and how they cope with these relatively frequently-occurring situations.

participants in a satisficing manner, but, on the other hand, also allow for the relevant facets to be elicited. Therefore, the participants' perspective must be included and discussed in order to make substantiated claims.

11.2.3 Claim 3: Performance

Claim 3: The participants are enabled to display a performance in the test that represents the target domain.

Warrant: ... because the performances are generated in a standardized manner, participants are able to access and comprehend the material, show their actual capabilities and the performance resembles aspects of the target domain (Schreiber & Gut, 2022, p. 132).

Evidence to support this claim consists of a description of the process of data collection. Furthermore, the performances of the participants should not depend on the individual actor that they interact with in the RobS and be able to show their actual capabilities. This also reflects the participants' perception of fairness (see Chapter 10.1.1). This claim requires backing, in that the material should be piloted with a group comparable to the sample so that the material's accessibility is ensured before a larger sample is confronted with it, the impact that the individual actor has on the test scores is explored and participants are asked about their perception of fairness.

11.2.4 Claim 4: Item score¹⁰⁰

Claim 4: The performance of the participants during test can be assessed using the scoring instrument.

Warrant: ... because the items in the scoring instrument, which are based on the framework, reflect the relevant facets to measure the construct comprehensively. The individual items are operationalized appropriately to the test format and can be generated in a replicable manner (Schreiber & Gut, 2022, p. 132).

The claim can be supported by several kinds of evidence. Qualitatively, expert judgements are a potential source, as stated by Kane (2006, p. 28). Experts in the target domain may evaluate the relevance of the criteria for the construct, the operationalization of the specific criteria as items. This is in addition to expert judgements that focus solely on the model (see Chapter 11.2.1) and is at the very least, part of Messick's concept of structural validity (1995, p. 745). Moreover, a description of item frequencies also contributes to an understanding of their appropriateness. Another relevant aspects to consider is the extent to

¹⁰⁰ The scoring instrument does not include indicators that are aggregated to items, thus the individual criteria are referred to as items. This aligns with the comments made by Schreiber and Gut (2022, p. 143).

which the item scores are generated in a replicable and consistent manner, reflecting parts of Messick's notion of generalizability (1995, p. 746). Evidence to support this aspect can be generated in inter-rating procedures, which examine the applicability and transparency of scoring rules for individual items. The calculation of rater agreement and reliability coefficients offers a quantitative measure of replicability. This claim, too, requires extensive backing as it is the foundation for the generation of summative grades. It is thus essential that the items of the test are appropriately operationalized and relatively consistently replicated by different raters.

11.2.5 Claim 5: Test score

Claim 5: The sum of the item scores generates a test score that represents the participants' levels of performance.

Warrant: ... because the test scores are generated reliably by several raters and the range of generated test scores depicts a distribution across the spectrum, indicating that it differentiates sufficiently. The individual items that contribute to the test score are all relevant for the construct and are internally consistent (Schreiber & Gut, 2022, p. 132).

Evidence that supports these claims is first and foremost of a quantitative nature, including a description of the test score data, validity coefficients of the individual items that are used to aggregate the test score, and measures of the inter-rater reliability of the test scores. Similarly to the previous discussion of item score, this argument contributes to Messick's notions of structural validity and generalizability (1995, pp. 745–746). This claim requires substantial, but less extensive backing than the previous claim, as most of the aspects that are relevant for the test score is part of the item score discussion, e.g., item-specific reliability, which is a prerequisite for the test score discussion.

11.2.6 Claim 6: Latent trait

Claim 6: The individual test score can be used to describe the participants' level of competence of providing feedback on learner texts in one-on-one settings.

Warrant: ... because the situation in which the performance takes place closely resembles the target domain (Schreiber & Gut, 2022, p. 132).

Evidence to support this claim is usually generated to support prior claims (model, test). It is however extended to include evidence that describes the relationship of the test score to other criteria that are related to the target domain (M. T. Kane, 2006, p. 36). This could include the scores of other tests, such as in this case, knowledge about feedback, professional knowledge in the area of EFL

or lexical knowledge/general language competence and is, thus, compatible to Messick's overall understanding of generalizability (1995, p. 746).

Due to the novel nature of the RobS, which seeks to extend competence measurement of existing tests that focus on the levels of "knows" and "knows how" (Miller, 1990, p. 63), it is relevant to investigate the nature of the relation of test results which, according to the competence model by Blömeke et al. (2015) are all facets of competence. Substantial backing is thus required that looks at several different facets of the target domain. However, the nature of the relationship between knowledge and performance is still open to discussion (Krauss, Lindl, Schilcher, & Tepner, 2017, p. 33). Thus, the collected evidence and delineated assumptions should be viewed reflectively in this context.

11.2.7 Claim 7: Interpreted trait

Claim 7: The test score can be used to make interpretations about the participants' competence.

Warrant: ... because the collected empirical data aligns with assumptions that have been made *a priori* (Schreiber & Gut, 2022, p. 133).

These assumptions should be derived from the theory and be logically construed. In this context, for example, assumptions can be made that the test score is higher for participants who are further along in their professionalization process, have a higher graduation grade, and have more experience in the target domain. Moreover, the test score should align with experts' global evaluation of the performance, as they are supposed to be able to distinguish between competence levels. Hence, measures that describe the relationship between these external variables can be used as evidence. The evidence can also be used to discuss the external validity as described by Messick, as it focuses on the "nonassessment [sic!]" aspects (1995, p. 746). Although backing for this claim is relevant, there is scarce prior research which would indicate that specific relations with external variables would strengthen certain interpretations. Similar to the previous claim, results should be seen in the context of (scarce) prior research.

11.2.8 Claim 8: Use

Claim 8: The test can be used for the intended summative purpose.

Warrant: ... because the generated evidence supports the intended use of the test and stakeholders view the test positively (Schreiber & Gut, 2022, p. 133).

Participants' perspective on the test is relevant as it indicates how the RobS could be seen. As the test is not yet used as intended in practice, no data can be generated that would allow inferences to be made as if it were in actual usage. This also extends to the impact that the score interpretation has. The use of the

test aligns with Messick's proposal of consequential validity (Messick, 1995, p. 746). Backing for this claim is of lesser relevance for this thesis. As the developed RobS can be thought of as a prototype which can be used for summative assessment purposes, the use and consequences are not of primary concern. Subsequent implementations should, however, investigate the use and how it impacts students in their professionalization.

12 Summary

This chapter's aim was twofold: 1) Outline the understanding of test criteria and validity that is the basis for this thesis and 2) present the framework for the validation of the test. First, secondary and primary criteria were outlined to evaluate a test's merits, with validity being the main focus of interest. The concepts of validity are then briefly outlined diachronically, looking at the traditional three-pronged approach (Newton & Shaw, 2014), Messick's concept of construct validity (1989, 1995) and Kane's argument-based approach to validation (2006, 2013a). The latter approach, which is the approach taken in this thesis, views validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, NCME, 2014, p. 11) and describes two distinct, but connected arguments: the claims that are being made, i.e., the *interpretation/use argument* (IUA), and the evaluation of these claims, i.e., the *validity argument* (VA). Based on the framework for validation proposed by Schreiber and Gut (2022), who have a similar approach to validation, the IUA that is to be evaluated in this thesis presented. The following part of the dissertation will present the collected evidence so that an evaluation of the IUA is possible in the subsequent part, which focuses on the VA.

Part IV - Empirical Findings: Results of the Investigations

To validate is to investigate.

Lee Cronbach (1971, p. 443)

In this part, the empirical data that is collected in order to provide backing for the warrants and claims outlined in the IUA (see Chapter 11.2) is presented. As the test development and the investigations into the validity are strongly intertwined, this part will start rather chronologically by exploring the initial test development phase, including pre-pilot and pilot phase. After modifications to the test, the main survey was conducted. Drawing on the data collected in the main survey, four studies are presented that provide further evidence for the discussion of the claims made in the IUA. Before each study is presented, the corresponding claims and type of evidence that the study is meant to generate are described.

13 Research Design and Overview

In this chapter, the general research design that is used in order to develop the test and collect evidence to discuss the validity will be presented. Generally, two research cycles can be identified: The 1st research cycle focuses on the initial test development and (pre-)pilot phase, while the 2nd cycle includes the main survey, i.e., the evaluation of the collected video-recorded performance data and the design of the final version of the test and scoring instrument (see Figure 21). Parallel to the different steps in the research process, validity investigations are conducted, which, as the arrow in Figure 21, indicates, may continue after the 'final' version has been settled on. Consequently, the final version of the RobS relates only to the last version of the test, as presented in Chapter 8.

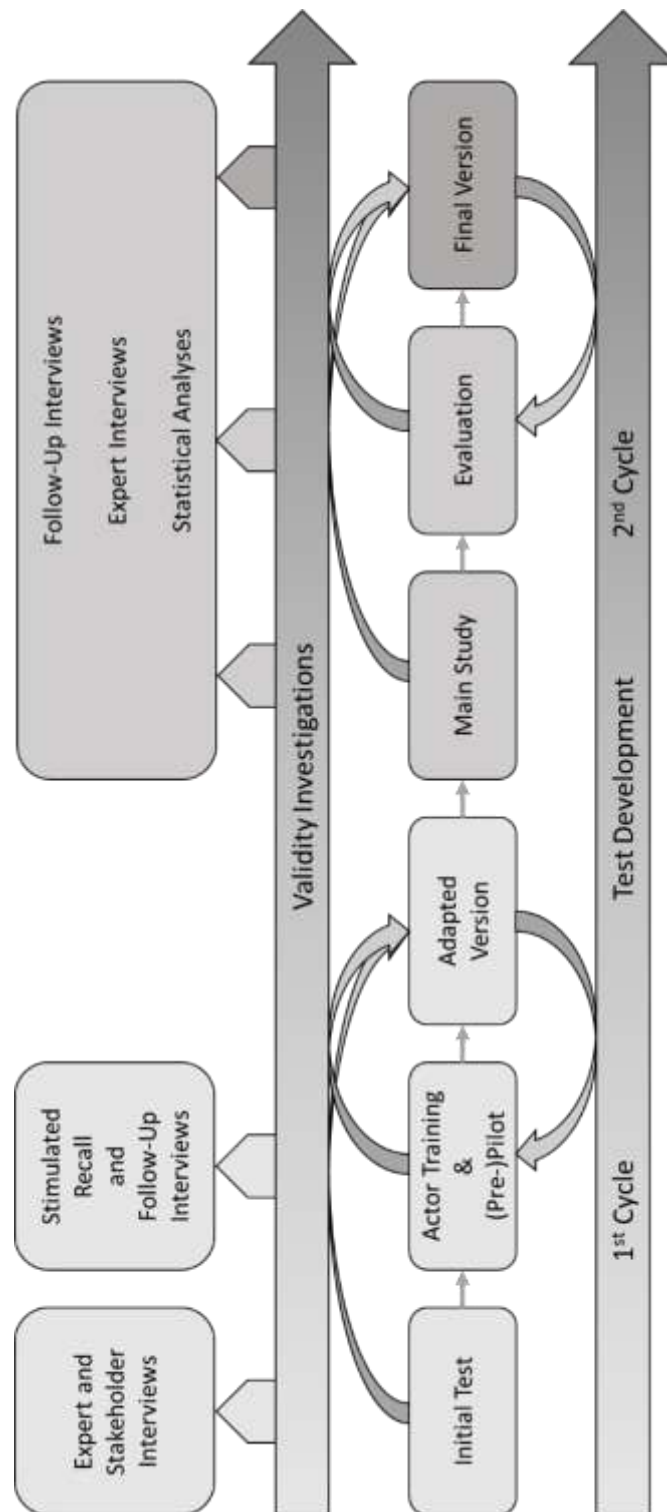


Figure 21: Research design of the survey.

The 1st cycle focuses on the development of the material for the test procedure, i.e., the material for the preparation and simulation. First, the framework for the assessment is developed, as described in Chapter 7. Based on the framework, the material that the participants are given to prepare for and use in the

simulation is designed, as described in Chapter 8. In the development process, the learner text is given specific attention as it is the focus of the feedback conversation. Several steps are taken, including expert and stakeholder interviews, to ensure that the text allows participants to show the facets of the framework and is still perceived as authentic with regard to Claim 2. Furthermore, the role of Mia, including the description of the role, as well as the VTs, which are designed to trigger facets of the framework, are developed. During the training of the actors and the pre-pilot phase in which colleagues took on the role of the teacher, different approaches are tested and a set of VTs selected for the pilot study.

The main purpose of the pilot survey is to identify issues in the simulation and research procedure, and adapt the material and process accordingly before the main survey. Concerning the simulation, the pilot serves as a first exploration of the target group's perception of authenticity concerning the material, task scenario, and the actors' behavior. Thus, two different learner texts and different material designs are tested. Another focus in this context is on the VTs: Stimulated recall interviews (SRIs) are conducted to explore their appropriateness, in order to discuss Claim 2. Concerning the research procedure, the pilot is used to identify issues and reduce questions in the guideline for the follow-up interviews. This also applies to the test instruments, with the aim of reducing the overall duration of participation. Lastly, the pilot also serves as an opportunity for the actors to gain more experience in the simulation, as these are the first instances in which they are confronted with the target group. Based on the results of the pilot survey, modifications are made and described for the main survey.

In the second cycle, the main survey, the simulation is tested with a larger number of participants, using the identical set of material. A larger sample of scored performances allows for statistical analysis, which is necessary for the discussion of Claims 2 to 8. Participants are interviewed with regard to their perspective on authenticity and tested competences (Claim 2), fairness (Claim 3), and potential use of the RobS as a summative assessment format (Claim 8). Expert interviews are conducted to obtain data on the scoring instrument from a content (Claim 1 and Claim 4) and rating perspective (Claim 7). With the data of the main survey, the scoring instrument is also refined by means of a quantitative perspective on the generated item and test scores. Based on the final test score results, further analyses are conducted to discuss Claims 6 and 7. Table 5 provides an evidence organizer, which can be used to navigate the empirical part of this thesis when evidence for specific claims is of interest.

Table 5: *Evidence organizer.*

Claim	Evidence	Chapter
1 - Model	Presentation of the Theoretical Framework	7
	Expert Judgements of Items	15.3.3.2
2 - Test	Participants' Perception of VTs	14.2.2
	Participants' Perception of Authenticity	15.2.2.1
	Participants' Perception of Tested Competences	15.2.2.2
3 - Performance	Presentation of the Data Collection Process	Pilot survey: 14.2.1 Main survey: 15.1
	Participants' Perception of Fairness	15.2.2.3
	Test Score Independence from Actors	0
4 – Item Score	Expert Judgement of the Items	15.3.3.2
	Frequency of the Item Scores	15.4.1.1.2
	Inter-rater Reliability of the Item Scores	15.4.1.2
5 – Test Score	Description of generated Test Scores	15.4.2.1
	Internal Consistency of the Item Scores	15.4.1.1.2 15.4.1.4 (Revised)
	Inter-rater Reliability of the Test Scores	15.4.2.2
6 – Latent Trait	Correlations regarding Facets of Professional Knowledge	15.5.4.3
7 – Interpreted Trait	Comparisons and Correlations regarding School-related experience	15.5.4.1
	Comparisons and Correlations regarding Academic Progress and Success	15.5.4.2
	Correlations regarding Expert Ratings	15.5.4.4
8 - Use	Participants' Perception on the Use of RobS as Summative Assessment	15.2.2.4

14 (Pre-)Pilot Phase and Survey: Actor Training and Material Development

In this chapter the preparatory steps of the RobS are outlined and focused on. The pre-pilot phase was predominantly concerned with developing the material and training the actors. The development of the material included several smaller steps with individual small studies. As not to inflate this thesis, the results are not analyzed and presented systematically. Rather, a narrative approach is taken to describe the material development. In the pilot survey, the aim was twofold: On the one hand, the appropriateness of the material was explored, and on the other hand, the research design was refined for the main survey. Consequently, a smaller sample of $N = 10$ PSETs were recruited to examine these aspects. Similar to the pre-pilot phase, findings are presented in a condensed manner, focusing on the results, as well as indications and modifications for the main survey.

14.1 Pre-Pilot Phase

The primary concern during the pre-pilot phase is creating the initial test material that is used in the RobS and the training of the actors. The material design was accompanied by several evaluative steps that included empirical findings, which will be described in the following chapter.

14.1.1 Creating and Evaluating Authentic Material

In the initial test development phase, i.e., the 1st cycle, a particular emphasis was put on the learner text (M1) as it is the focus of the feedback conversation, i.e., the simulation. The text has to allow the participants to show their capabilities as outlined in the framework while still being perceived as authentic in the context of the RobS, which requires a specific balance. Thus, hybrid learner texts, i.e., texts that have authentic and artificial elements, were developed in several steps.

In a first step, six texts were created that imitated possible solutions of the task, outlined in M2, for varying learner levels. To enhance authenticity, typical features of interlanguage were included, such as overgeneralization or false cognates (Lightbown & Spada, 2013, pp. 44–63). For instance, overgeneralization could be found in learners applying the tense marker *-ed* for regular forms of past simple to irregular ones, e.g., **buyed* instead of *bought*. False cognates, also referred to as false friends, occur regularly and are a form of L1 inferential error. For L1 German learners of English, for instance, the English word *become* is a false cognate to the German *bekommen*, which means *to get* or *to receive*. Additionally, the *Creative Writing in Language Teaching* corpus (CWILT) is drawn upon to get insights about typical lengths of texts and for specific features that

are found in texts by real learners (D. Gerlach & Götz, 2021a). For example, the final version of the text, presented in M1 (see Appendix A.2), includes overgeneralization in the form of **bued*, which is found in the corpus [#095, GM, m, 9]¹⁰¹, as is the repetitive misspelling of **parants* [#118, GM, m, 10].

Authenticity of learner texts and material – Experts' perspective

In an explorative approach, these hybrid texts were discussed with $N = 4$ experts, including three researchers in EFL teacher education with experience at school and one in-service teacher, in November/December 2021. Two interviews were conducted in person, two via Zoom due to the then evolving nature of the COVID-19 pandemic. The guideline used for these interviews can be found in Appendix D.1. Before the interview, experts agreed to the data collection and processing. At the beginning, experts were informed about the aim to decide on a text for the RobS but they did not know that the presented texts were created artificially. First, experts were asked to name typical features of learner texts in general. Then, they were asked to describe how they would approach teaching the genre of a newspaper report, and evaluate the appropriateness of the task, which is the fictional basis for the hybrid learner texts (M2), so as to collect information for the appropriateness of the contextualizing material. Afterwards, experts were asked to read the hybrid learner texts and evaluate which grade they may have been written by. Subsequently, experts were asked to discuss whether these texts can be described as realistic for the age group of year ten students¹⁰², and which aspects in the text can be seen as authentic. At the end, they were informed about the hybrid nature of the texts. The interviews are transcribed based on the guidelines outlined in Appendix E.2.

With regard to the evaluation of the task complexity, experts agree that the task is suitable for lower secondary, but vary in their evaluation of the appropriate grade, describing it as dependent on the specific type of school and providing a range from year seven to year nine. When asked for the approach to teaching the genre of the news report, the experts provided similar answers. As an example, expert IV¹⁰³ describes the approach as follows:

E_IV: Zeitungsberichte. Also, ehm, würde zunächst einmal müsste man ausgehen von einem, ehm, Genrebeispiel, damit die Schülerinnen und Schüler wissen, was wird eigentlich von mir erwartet. Ehm, und dann kommt es ja noch, also Zeitungsbericht ist dann ja doch ein relativ

¹⁰¹ Citation adapted from D. Gerlach and Götz (2021b, p. 199): [#textcode, schooltype (GM = comprehensive school), gender, grade] (see also D. Gerlach & Götz, 2021a, p. 5).

¹⁰² The learner group was later changed to year nine to better align with the findings and the curricular guidelines, as discussed in Chapter 6.2.3.

¹⁰³ In order to distinguish experts here from those in Chapter 15.3, Roman numerals are used.

weites Feld. Ehm, aber tendenziell grammatikalische F Formen, die relativ häufig vorkommen, ist eh Aktiv/Passiv so, sowas würde man da zum Beispiel machen. Und müsste auf jeden Fall, eh, berichten können, darüber was passiert ist, also Vergangenheitsformen beherrschen. Ehm, ja wir müssen die Struktur kennen, ne erstmal die ganzen W-Fragen und so beantworten, dass die präzise sind und dann elaboriertere Informationen später kommen.

(E_IV: Newspaper reports. So, first a genre example is used as the starting point so that learners know what is expected. Ehm and then the news report is a relatively broad field. Ehm but tendentially, grammatical forms that are frequent, active/passive voice for example. And certainly, eh, the ability to report about what has happened, so past tense forms. Ehm, we have to know the structure, answer the *wh*-questions, that they are precise and more elaborate information is included later.)

Evidently, experts describe an approach that resembles the approach to genre writing as described by Hyland (2019, p. 20; see also Chapter 6.2.2). Consequently, the approach adopted for the excerpt of the series of lessons in M3 is supported (see Chapter 8.1, Appendix A.4).

The primary aim was, however, to explore the extent of authenticity of the artificially created hybrid learner texts. Overall, all four experts state that the texts are sufficiently authentic and representative for learner texts. For instance, expert IV states the following:

E_IV: Ehm, also was was alle Texte haben ist so dieses naja ich kann mir vorstellen, dass es ein Schüler geschrieben hat, ne, weil halt typische Fehler, Fehler drin sind. Ehm.

(E_IV: Ehm, so what is common in all texts is this, well, I can imagine that a learner has written it because typical errors, errors are included.)

Another expert states that, while reading the text, they were immediately reminded of specific learners that might have written these texts:

E_III: [...] man hat irgendwie direkt eine Art Schüler im Kopf, wenn man sich die Texte durchliest, wie das klingt. Ehm, ja. Und auch man hat direkt auch individuelle Beispiel im Kopf wer, oh das hätte doch von dem und dem oder von der und der sein können.

(E_III: [...] somehow, a type of learner directly comes to mind while reading the texts, how it sounds. Yeah. And also, individual examples come to mind who, oh, this person and this person or this person could have written it.)

The experts also highlighted individual features in the texts which they considered particularly authentic. Expert I, for instance, references a segment from a text sample, which is later used as the basis for M1:

E_I: [...] So „we become 1000 euro for this“ ist glaube ich sehr authentisch.¹⁰⁴

(E_I: [...] I believe “we become 1000 euro for this” is very authentic.)

Occasionally, the experts also address aspects they consider unrealistic or inauthentic. For example, expert I states:

E_I: [...] sowas wie „The price is for 500 pflanted trees.“ Also „pflanted“ würde ich, habe ich noch nie gesehen bei Schülern, würde ich glaube ich auch nicht erwarten, dass das kommt, aber vielleicht vertue ich mich da auch.

(E_I: [...] something like “The price is for 500 pflanted trees.” So, “pflanted” is not something I have seen before with learners, and not what I would expect, but I might be wrong here.)

The experts also mention different aspects as typical features of learner texts in general, e.g., inferences from the L1, lack of planning, lack of objectivity, a personal writing style, word order, spelling, lexis, grammar, 3rd person singular -s, punctuation, repetitive syntax. Based on these findings, the texts were modified to ensure a higher extent of authenticity. The modified texts were used for the subsequent study with PSETs.

Authenticity of learner texts – PSETs' perspective

As authenticity is subjective, it is essential that the target group also perceives the learner text as authentic. Thus, an interview study was conducted with $N = 5$ PSETs to investigate their perception of authenticity. The research questions for this study are as follows:

RQ1: To what extent are PSETs able to distinguish between authentic and artificial learner texts?

RQ2: Which text features do PSETs address with regard to authenticity?

To explore RQ1, authentic learner texts were used. These were collected in December 2021 in a ninth grade at a *Gymnasium* in North Rhine-Westphalia. Due to the pandemic, the data collection was done remotely by the teacher as no external persons were permitted. The teacher was given a slightly modified version of the task (M2) to account for the class's lack of preparation. Learners were informed about the project and the General Data Protection Regulation (GDPR) orally by the teacher. After writing the text, learners had the option to voluntarily hand the text in for the purpose of this study. Texts were anonymized by the teacher and transcribed into a digital format for the research process.

¹⁰⁴ For consistency, and transparencies, all quotations taken from transcripts of interviews that are generated in this study are formatted as bloc quotations.

The study with the PSETs was conducted via Zoom in February 2022. Potential PSETs were contacted via e-mail and participated on a voluntary basis, without reimbursement. The guideline for the interviews can be found in Appendix D.2. Interviews are transcribed based on the guideline outlined in Appendix E.2. Before the interview, the PSETs agreed to the terms of participation and recording. In the study, the PSETs were shown seven text couplets (A-G) in two separate parts (see Table 6). In the first part, four contrastive couplets were shown that each consisted of an authentic and a hybrid text (A-D). This approach is a modification of a comparative judgement procedure, as students had to decide which of two texts they considered more authentic (Pollitt, 2012). As the PSETs were not expected to draw upon any experience as teachers for comparison in order to evaluate the authenticity, this first part allowed the PSETs to compare two texts and construct what they perceived to be authentic. In the second part, three couplets were shown that were not necessarily contrastive, i.e., one couplet could include two hybrid texts (E-G). This way, the participants could apply their generated construct of authenticity from the first part to the texts in the second part. Participants were informed about the difference in the two parts beforehand. For each couplet, participants were asked if they believed that the texts were written by actual learners and which of the two texts they considered to be more typical of a learner in the ninth grade. In the first part, participants were then asked to decide which of the texts was hybrid and which was authentic and to provide reasons. In the second part, participants simply had to decide for each text individually if they considered the text to be hybrid or authentic. The order of the text couplets rotated for the first part but was identical for the second part, in which the second couplet (F) consisted of two hybrid texts. The remaining couplets are all contrastive. Text samples B2 and F2 as well as F1 and G2 are identical, as to explore whether the perception changed throughout the study. Text samples of hybrid and authentic texts were matched according to text length and perceived similarity of complexity.

Table 6: Overview of text samples and correct identification¹⁰⁵.

Interview part	Couplet	Text		Instances of correct identification
		Hybrid	Authentic	
First Part	A	A2	A1	60 %
	B	B2/F2	B1	80 %
	C	C1	C2	100 %
	D	D1	D2	40 %

¹⁰⁵ Text samples B2/F2 and F1/G2 were identical texts.

Interview part	Couplet	Text		Instances of correct identification
		Hybrid	Authentic	
Second Part	E	E1		100 %
			E2	40 %
	F	F1/G2	-	40 %
		B2/F2	-	40 %
	G	G1		60 %
			F1/G2	80 %

First, results with regard to RQ1 are presented and discussed. From a quantitative perspective, participants correctly identified 70 % of the texts, on average, in the first part in which there was a contrastive choice and the chance of identifying the text correctly was 50 %. In the second part, in which the choice had to be made for each text individually, 60 % of the texts were identified correctly on average. Interestingly, there was a substantial difference with regard to the individual texts. For instance, text sample D1 in the study was only correctly identified by 40 % as the hybrid text, while C1 was identified as the hybrid text by all participants (100 %). With regard to the text sample B2/F2, the text was correctly identified as hybrid by 80 % in the first part, while only 40 % stated so in the second part. Similarly, F1 was identified correctly by 40 % as hybrid, while 80 % correctly identified the same text as hybrid later as G2. Results indicate that students identify texts correctly as authentic only to a slightly larger extent than chance, especially if the couplets are not explicitly contrastive, as was the case in the second part. Moreover, perception of a text varies during the course of a study, as the changing identification of text samples B2/F2 and F1/G2 indicate. Arguably, with the exception of text C1, all texts are perceived to resemble learner texts to a certain extent and can thus, as at least partially, be considered authentic.

With regard to RQ2, the qualitative data will be summarized and a focus placed on aspects that led to modifications of the texts. Generally speaking, it can be stated that the authentic perception of text features differs between individuals. As an example, there is a discrepancy between participants with regard to whether a higher number of errors or mistakes is authentic or inauthentic. S2¹⁰⁶, on the one hand, argues:

S2: Ich sage D1 ist der authentischere Text, weil je mehr ich darüber, eh, den drüberlese, desto mehr Fehler sehe ich auch.

¹⁰⁶ For differentiation, PSETs in this interview study are abbreviated with S, i.e., student.

(S2: I would say that D1 is the more authentic text, because, the more I read it, the more mistakes I see.)

On the other hand, S3 argues for the same text that it is the hybrid text, but reasons this with the perception that there are too many errors or mistakes:

S3: Ich sage der Erste [D1] ist konstruiert, weil es vielleicht zu viele Fehler und quasi zu leicht gemacht, also, dass die Schüler schlechter dargestellt wurden als sie vielleicht eigentlich sein könnten.

(S3: I would say the first one [D1] is constructed, because maybe too many mistakes and it is too easy, so that the learners are depicted worse than they actually could be.)

Moreover, texts are generally perceived as less authentic if they represent the target genre to a high extent, especially with regard to structure or multimodal elements, like an image:

S5: Mh, C1 wäre dann von Ihnen auf jeden Fall. Und das ist allein schon wegen dem Bild wird das, glaube ich kein Schüler oder keine Schülerin machen, dass man da so eine Bildunterschrift hat.

(S5: Mh, C1 would be yours, for sure. Just because of the picture, this, I believe no learner, will do, with the subtitle for the image.)

However, if the performance depicted in the texts is too far from the objectives and the texts are thus perceived as too purposefully constructed, this also decreases authenticity. Hence, for a text to be perceived as authentic, it has to be situated between the two poles – it should neither be too far away from the objective, nor should it be too close, but rather in-between.

Some distinct language features are also mentioned by participants in that context. One aspect that is mentioned as inauthentic by four of the five participants is the use of quotation marks:

S1: Ja, aber beim B2 wäre ich mir unsicherer, weil ich es untypischer in der achten Klasse so viele Anführungszeichen zu benutzen. Ich würde die direkten deutschen Zitate erwarten.

(S1: Yes, but with B2 I would be less sure, because I [think] it is atypical to use that many quotation marks in the eighth grade. I would expect direct German quotations.)

Another aspect that contributes to a text being perceived to be authentic is the inclusion of language features that are potentially influenced by German, i.e., false cognates. Also, a frequently addressed topic is the issue of spelling mistakes, which seem to be quite influential in creating authenticity in the minds of the students:

S4: [...] „selebrated“ also auch mit <s>, das passt glaube ich ganz gut für einen Schüler in, ehm, ja.

(S4: [...] „selebrated“ with the <s> that is fitting for a learner, yes.)

Similarly, challenges regarding the correct use of verb forms are frequently mentioned as an authentic notion in the text by the learners:

S2: D1 [ist typischer], weil ich würde jetzt so aus reinem Bauchgefühl sagen, dass mehr Leute Probleme mit dem Partiz, den, eh, Partizipien haben als mit der Rechtschreibung.

(S2: D1 [is more typical], because I have a gut feeling that more people have problems with the participle than with spelling.)

Other, more deep-level related features also contribute to authenticity. For instance, the chronological inclusion of the information and the final sentence in D1 is considered to be particularly authentic:

S4: die die vielen Angaben da als hätte man da halt so eine, so eine Liste irgendwie abgearbeitet. Ehm, und vor allem dieser kurze Endsatz „It was an amazing event“ könnte ich mir vorstellen, dass das so ein Schüler so als Abschluss reinschreibt.

(S4: The many facts, as if one had a list that was worked through. Especially that brief sentence at the end "It was an amazing event" I could imagine that a learner writes that as a conclusion.)

Based on the quantitative and qualitative findings, two texts were chosen to be tested in the pilot survey after modifications: D1 and B2/F2. The texts were chosen as they were perceived as authentic more often. Additionally, particular text features that were already implemented in D1 were frequently mentioned to contribute to an authentic perception, such as its final sentence. Modifications were made to the texts, based on the findings. Regarding B2/F2, these modifications relate to the headline, which was adapted from the authentic learner texts in which "Plant trees for bees" is frequently used. Moreover, quotation marks were reduced, and a syntax error, i.e., **make we*, and the false cognate from A1, **its give a festival where*, were added. Modifications in D1 relate to the verb errors, in that fewer verbs depict the overgeneralization of the past simple tense marker *-ed*, instead, the infinitive form is used, e.g., **wined* was exchanged with **win*. Furthermore, the comma after 'also', as well as all quotation marks were removed as the results indicate that this would increase authenticity. Additionally, false cognates, i.e., **become* and **it gives* were added, as well as the spelling mistake **tress*. In order to facilitate feedback on structural elements in D1, the headline was removed¹⁰⁷.

After final versions of the texts were created, both texts were coded by different people, using the dimensions of text quality outlined by Becker-Mrotzek (2014; see also Chapters 7.3.3.1 and 8.3) as categories. The purpose of this inter-coding process is to ensure that the texts include the respective dimensions so

¹⁰⁷ Text sample A of the pilot study, i.e., after modification, can be found in the Appendix A.8; the other text sample is used in the main study and is found in Appendix A.2.

that the participants can identify them and refer to them in their feedback conversation. $N = 3$ people, two researchers in EFL and a student assistant who studies English, coded the texts. After the initial coding, results were discussed and agreement found between the three coders with regard to aspects that could be addressed in the feedback conversation. Both texts were found to include feedback possibilities for all dimensions. Results of this exploration for the final text (M1) are described in Table 2 (Chapter 8.1).

The remaining documents, i.e., M0, M2-M4 (see Chapter 8.1 and Appendix A) were evaluated with regard to their authenticity by $N = 2$ experts who have experience as EFL teachers. Both were sent the material and a brief written survey via e-mail. The survey focused on the authenticity and appropriateness of the material and included the following questions¹⁰⁸ for each M0, M2 and M3:

- 1) Is the situation / task sheet / series of lessons authentic/realistic?
- 2) What needs to be changed to create a more authentic situation / task sheet / series of lessons?
- 3) Could the task sheet / series of lessons be used / implemented like this?
- 4) If not, what needs to be changed?

For M4, the questions were slightly different:

- 1) Could the text be used as presented in a year 9?
- 2) If not, why?
- 3) What needs to be changed?
- 4) Is it a realistic/authentic text sample for a news report?

Generally, both experts consider the material to be authentic and appropriate for the age group as they do not negate any question. One expert proposes that the task sheet in M2 could be designed “more attractively” (translated by TJ) for actual use with learners. Regarding the series of lessons (M3), the other expert notes that the rewriting phase, which is planned for the lessons after the one focused on in the RobS, does not necessarily have to occur during instruction. No modifications were made, as the propositions were not considered relevant for the purpose of the material in the RobS.

14.1.2 Actor Training¹⁰⁹

Before the pilot survey, actors needed to be prepared to depict their role in a standardized manner. Due to the novelty of the format, the training of the actors

¹⁰⁸ Experts received the questions in German, the presented questions here are translated.

¹⁰⁹ The training of the actors was done in cooperation with the other project of PERFORM-LA, in which the same actors portrayed different roles for a simulated learner counseling session and parent-teacher conference (see Wotschel & Janzen, in preparation).

was simultaneously a first exploration into the RobS itself and modifications were made dynamically in that process. Furthermore, other aspects were tested parallel to the training of the actors, e.g., seating arrangements, camera positions, or the use of an external microphone. In the project, three student assistants functioned as actors, none of whom had substantial experience as an actor. All three actors identify as female, two have English as a subject of study and all three were in their early twenties during the course of data collection. As the actors started at different times in the project, the training for each actor differed slightly. The training of the actors took place in the spring of 2022 and can be distinguished in two major phases: In the first phase, actors and researchers collaborated internally, while in the second phase, external participants were included. Overall, the training lasted eight weeks, with meetings of about two hours twice a week.

In the initial stage of the actor training, the actors were put into the position of the participant. This was done so that the actors could get a feeling for the situation and can empathize with their counterpart. For the first actor, the researcher took on the role of the learner; subsequently, the other actors took on the role when new actors joined. The role descriptions and the feedback framework were introduced to the actors as part of their training so that they had the necessary theoretical background knowledge. Actors were instructed to fill in blanks of the role by falling back to their personal biography, if unforeseen questions come up, so as to maintain continuity in the simulation. Nonetheless, the actors are supposed to keep it to the general personality of Mia, as described in the role description. The main focus of the first phase was the use of VTs by the actors during the simulation. An initial set of VTs was created and tested in training scenarios, with slight modifications being made to their exact formulations. The researcher also took on the role of the participant several times, testing different versions of potential PSETs, e.g., a participant who does not let the actor speak. This way, actors could also train to interrupt participants in order to use the VTs. Some performances were video-recorded and reflected upon in the group to discuss VTs and their use in the RobS.

In the second phase, external collaborators participated in the training. This stage allowed the actors to train with less familiar people, increasing the complexity and the stakes of the situation slightly. Collaborators were colleagues who have a varying extent of experience in the role of the teacher. After each session, there was a debriefing with the participant and the actors, discussing issues like the notion of language adaptivity. In this context, the starting VT "Good morning" was developed and tested as an approach to nudge participants to start the conversation in English. Additionally, actors met for separate

sessions and provided each other with peer feedback in training to continuously practice their roles before the pilot survey.

14.2 Pilot Survey

After the material, i.e., the documents and VTs, was prepared and actors trained, a pilot survey was conducted. Due to the explorative nature of pilot surveys, the collected performance data of the participants cannot be used to discuss the claims in the IUA, as learner texts and material were neither completely identical within the pilot survey, nor the same as in the main survey. Besides the actors gaining more experience in the role by being confronted with the target group, the main purpose was to identify potential issues that were previously undetected or could not be assessed without directly asking the target group. This relates to the participants' perception of the test format, including the documents and VTs, as well as the research procedure, i.e., the guideline questions, as modifications for the main survey were still possible at that point. The following aims are pursued during the pilot survey.

Investigation of the VTs

An important focus of the pilot survey is to determine whether the developed VTs actually trigger the desired competence facets. VTs are conceptually similar to test items in pen-and-paper tests and should thus be of particular interest in the validation process before using the test with a broader sample. For that purpose, stimulated recall interviews (SRIs) (Barras, 2018; Finkbeiner & Schluer, 2018) were conducted with the participants, in which they were asked to describe their reaction to the triggers in the simulation. This study will be described and discussed in more detail in Chapter 14.2.2.

Determining the final version of test documents

While the material was previously investigated concerning its authenticity, the question remained whether actual participants are able to adequately use it. Furthermore, two learner texts emerged from the pre-pilot and a choice needed to be made for the main survey to ensure comparability of the performances. Another question is that of language, i.e., English or German, for the material, specifically for M0, the task description for the learner.

Determining the research procedure

The written survey in the pilot survey includes several test instruments for the relevant constructs, as previously outlined. One important consideration for the pilot survey is the question of how to reduce the overall time spent on the survey as the estimated time for participation in the pilot survey is about three hours.

Moreover, the final set of questions for the interview in the main survey is determined.

14.2.1 Data Collection and Sample Characteristics

Before the three aims are discussed, the process of data collection and the characteristics of the pilot survey sample are described.

14.2.1.1 Data Collection

An overview of data collection process for the pilot is outlined in Table 7. In the following, the data collection process will be described and reasoned in detail as the structure of pilot survey was adopted for the main survey.

Table 7: Data collection process in the pilot survey.

Step	Focus	Minutes
Greeting & Organization	Information on the project & GDPR Agreement on Participation	5-10
Simulation Procedure		
Preparation	Use the material to prepare for the feedback conversation Opportunity for questions	15
Simulation	Simulate a feedback conversation with an SL	10
Follow-Up		
Written Survey	Demographic Data Study Progress	5
	Feedback Knowledge	15
	PCK English	60
	Lexical knowledge	5
	C-Test	25
Stimulated Recall	Recollection of thoughts when confronted with a VT	25-30

Step	Focus	Minutes
Guideline Interview	Authenticity of • Material • Situation • Actor Fairness Validity	20-25
Conclusion		
Debriefing	Options for further questions Opportunity for reflection	5-10

The pilot survey was conducted from June 2022 - October 2022 at Paderborn University. Recruitment of participants was done by promotion in specific TEFL seminars. Participants were informed about the duration of the procedure (approx. three hours) and the 30€ reimbursement. For scheduling, participants registered in a moodle-based course and were able to select a time slot of their choice. Participation consisted of three different components: Greeting & organization, the simulation procedure and a follow-up. The follow-up itself consisted of a written survey, a stimulated-recall interview and a guideline interview (see Table 7).

Firstly, participants were greeted and information on the current COVID-19 regulations provided. All participants were given the opportunity to wear a mask and could also ask for the research staff to wear a mask. Participants were informed of the aim of the project, which is to develop and validate a novel test format for EFL teacher education. The different parts of the research procedure, e.g., simulation and follow-up, including written survey and interview, were outlined. Participants were informed orally about the GDPR, their respective rights, and about the essential conditions that had to be agreed upon to participate in the data collection. They received information on the types of data that were collected (audiovisual data and notes from the simulation, digital and analogue data from the surveys and audio files from the interviews) and the respective purpose of the data collection (evaluation of the simulation, participants' perception of the simulation and investigation of correlations between different constructs). They were informed about the pseudonymization process of the data as a means to relate individual data points to each participant and the anonymization process of transcribed data. They received further information on data storage, e.g., that the data will be stored for ten years after the end of

the project (i.e., 2035) and that the anonymized transcripts will be archived. Additionally, participants could agree to the two following uses of their data:

- a) Use of excerpts of the raw material (audiovisual, audio) for scientific purposes (e.g. use in publications, presentations etc.);
- b) Archiving of the video and audio data.

Participants then had time to individually read the written consent document, ask for clarification and sign the documents. Lastly, participants were asked to fill out the code for pseudonymization. This first part amounted to about ten minutes.

In the main part of the research process, the simulation was carried out similar to the detailed description of the final test in Chapter 8. Figure 22 shows a graphic illustration of the constellation of objects and interactors during the simulation. Participants sat at a table, which was equipped with disinfectant, a fresh bottle of water, paper cups and fruit lozenges (in case of coughing fits or similar incidents for both participants and actors) as well as a pen and a digital clock. They were given the material and 15 minutes to prepare for the feedback conversation. In the pilot, participants received different versions of the material. Variation related to the language of M0, which included the instruction, and the learner text. Both learner texts included similar features, e.g., overgeneralization of the past simple tense marker *-ed*, but also differences, for instance, text A had a title, while text B did not (see Appendix A.2 and A.8). After 15 minutes, participants had the chance to ask the administrator questions of clarification. Then, video-recording was started and the participant's code was shown to one camera and read aloud to ensure that both recordings can definitely be identified in the future. Afterwards, the administrator left the simulation room and entered the staff room. The simulation always started with the actors knocking on the door that separated simulation room and staff room. The simulation ended when both parties agreed to end the conversation and the actor re-entered the staff room. The test administrator then entered the simulation room and stopped recording. All simulations of the pilot study are transcribed. Transcription is done on two levels: The audio level is transcribed according to the GAT-2 minimal transcript conventions (Couper-Kuhle et al., 2011, pp. 7–17), while the visual level is transcribed descriptively, inspired by the same GAT-2 conventions.

After the simulation, participants had the opportunity to take a break. Then, they took part in a written survey which was provided online, except for the C-test, which was conducted as a pen-and-paper test for technical reasons. The survey consisted of six parts:

- A. Door to the simulation room
- B. Simulation room
- C. Door connecting staff and simulation room
- D. Staff room
- E. Test person
- F. Standardized Learner
- G. Test administrator
- H. Table for simulation, including equipment
- I. Table
- J. Camera 1 (J_P for Pilot, J_M for Main Survey)
- K. Camera 2
- L. Windows
- M. Whiteboard
- N. Display wall 1
- O. Display wall 2

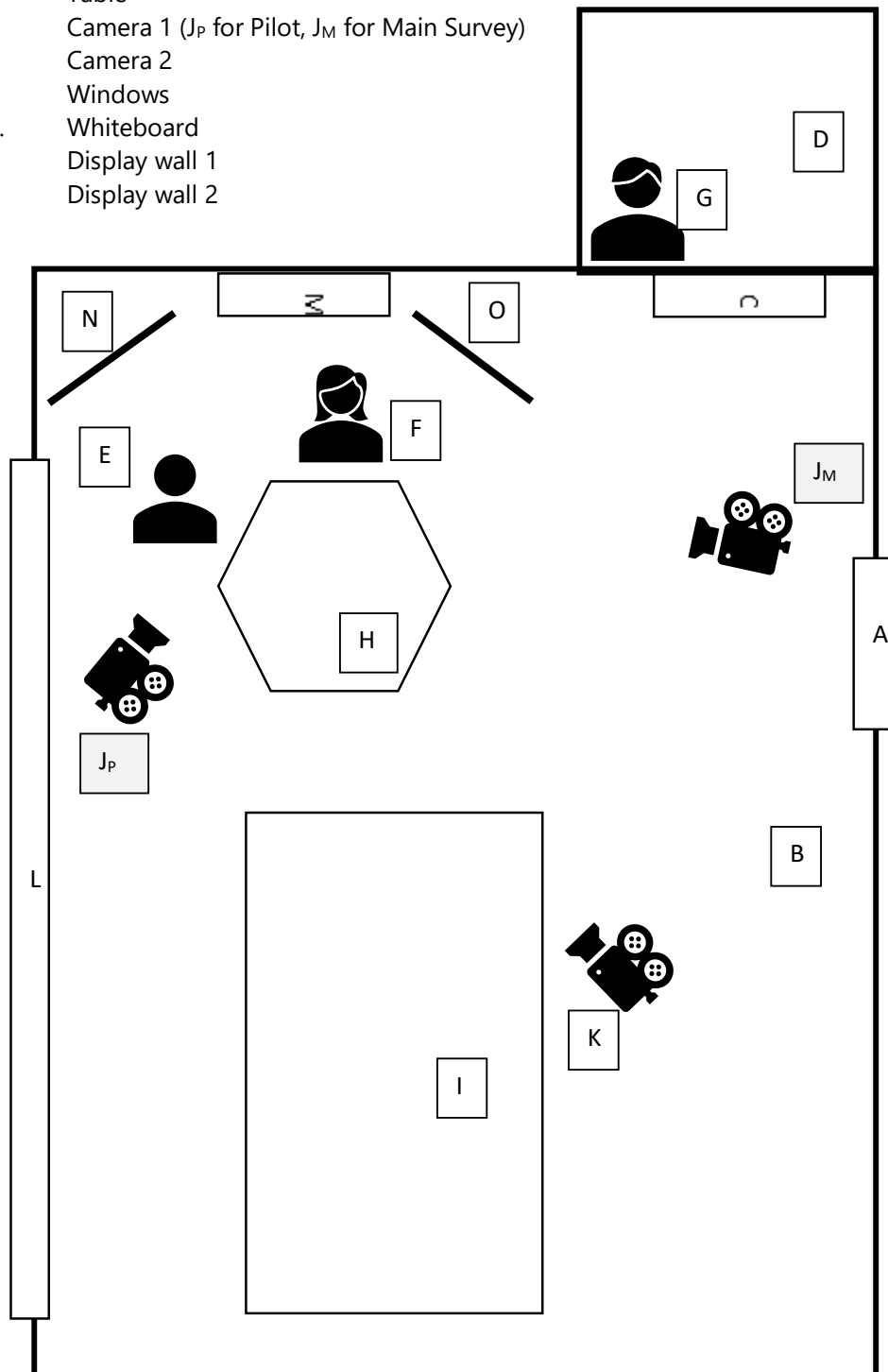


Figure 22: Constellation during the simulation.

- 1) demographic data, including information on academic success and progress,
- 2) a test on feedback knowledge (Schütze et al., 2017),
- 3) a test on EFL-specific PCK (Kirchhoff, 2017),
- 4) a C-test (Keijzer, 2007),
- 5) a test on lexical knowledge / general language competences (LexTALE) (Lemhöfer & Broersma, 2012).

The purpose of the written survey is twofold. First, demographic data is needed to contextualize and describe the sample. Second, the information on academic success and progress, as well as results of the tests are needed to discuss Claims 6 and 7^{110,111}. The C-test and the LexTALE test both aim at evaluating the participants' language competence. The C-test is a cloze test, described as an economic and reliable instrument to measure global language competence (Grotjahn, 2002a, p. 211). The specific C-test that was used in the pilot was designed by Keijzer for adult Dutch emigrants in Canada and consisted of five texts (2007). For the test, there was a strict time limit of five minutes per text (Keijzer, 2007, p. 166). Before use in the pilot, the C-test was pre-piloted with a range of differently skilled English speakers to test if it is appropriately sensitive to differentiate the different target group, which it was considered to be. As not to prime participants in their performance in the RobS by activating prior knowledge, these tests had to be conducted after the simulation. Due to technical and organizational reasons, demographic information had to be collected digitally as well, thus, this data also had to be collected after the simulation. Participants had no time limit for the three other tests, but, 90 minutes were estimated as a time frame, based on the information provided by the test developers.

Following the written survey, and an optional break, two interviews were conducted and recorded using Microsoft Sound Recorder, (1) stimulated recall interviews (SRIs) and (2) a guideline interview. The SRIs (1) in which the participants watched their recorded simulations, were conducted exclusively in the pilot survey. For the SRIs, a specific camera angle was used that only filmed the SL so that the participants are not looking at themselves but rather experience the situation from a similar perspective again (see camera J_p in Figure 22 and Figure 23). The purpose is to explore the extent to which the VTs elicit the intended reaction in the participants, allowing for a discussion on Claim 2. A more detailed description of this process is described in Chapter 14.2.2.1. Thereafter, the guideline interviews (2) were conducted. The purpose is to collect data on the participants' perception of the simulation with regard to relevant criteria for the

¹¹⁰ The reasoning behind the choice of specific tests can be found in Chapter 15.5.3.2.

¹¹¹ An overview of the specific items is found in Appendix C.

validation as outlined in the IUA, specifically authenticity, fairness and tested competences¹¹². The pilot survey is then tasked, on the one hand, with testing the questions and the answers they elicited and, on the other hand, getting first insights into how the participants perceive the test, potentially allowing for some adjustments to be made for the main survey. A particular focus of the pilot survey is the authenticity of the material, as this is the last opportunity to modify the material before the main survey.

Lastly, participants were offered a voluntary debriefing of the simulation. In the debriefing, participants were first asked to reflect on and self-evaluate their performance. Subsequently, the test administrator described three different criteria as aims, against which a feedback conversation can be evaluated. These were, according to the framework:

- 1) The model of feedback according to Hattie and Timperley (2007);
- 2) A focus on error patterns instead of comprehensive feedback;
- 3) Addressing aspects which the learner has achieved successfully.

It was refrained from providing explicit feedback to the participants because the scoring instrument was, at that point, still in development. By providing these criteria, the students were however still given the option to self-evaluate their performance and use the participation in the study as a learning opportunity.

14.2.1.2 Sample Characteristics

An overview of the sample in the pilot survey is provided in Table 8. As the data from the pilot survey (with the exception of the SRIs) is not evaluated in detail, the description focuses on the necessary aspects. The total sample consisted of $N = 10$ participants, of which $n = 9$ identify as female and $n = 1$ as male. Age was not included in the pilot questionnaire but added for the main survey. Most of the participants, $n = 9$, are from Paderborn University and $n = 1$ from the University of Münster. The sample is distributed relatively equally among the phases of study, with $n = 4$ in the Bachelor program and $n = 6$ in the Master program. Of the Bachelor students, $n = 1$ participant studies in the program for lower secondary and $n = 3$ for higher secondary. For the Master students, $n = 4$ participants study in the program for lower secondary and $n = 2$ for higher secondary. The average number of semesters is $M = 14$ ($SD = 2.83$), the median is 14.

¹¹² See Chapter 15.2, for a comprehensive description and analysis of the data collected in the main survey.

Table 8: Sample characteristics of the pilot survey.

		<i>N</i>	<i>n</i>	%	<i>M</i>	<i>SD</i>	Median
Gender		10					
male			1	10			
female			9	90			
University		10					
Paderborn			9	90			
Münster			1	10			
Phase of study		10					
Bachelor of Education / Polyvalent Bachelor ¹¹³			4	40			
Master of Education			6	60			
Study program		10					
Bachelor of Education / Polyvalent Bachelor	Lower Secondary (HRSGe)		1	10			
	Upper secondary (Gym/Ges)		3	30			
Master of Education	Lower Secondary (HRSGe)		4	20			
	Upper secondary (Gym/Ges)		2	20			
Number of semesters		10			13	2.83	14.00

14.2.2 Investigating the Verbal Triggers

Within the pilot survey, a particular focus is put on the validation of VTs in SRIs. This specific portion of the pilot survey will be focused on in more detail, as the VTs are an essential aspect of the RobS. VTs are addressed as a control element for simulations in Chapter 4.3.1. They can be defined as “specific questions, information, and/or comments” (Dotger et al., 2010, p. 133) that the SL uses to elicit certain reactions. Traditional tests usually rely on written test items, which are often validated in the context of a pilot survey with regard to their construct validity to estimate whether participants understand the items as intended (Baras, 2018, pp. 72–73). VTs resemble test items in that both are used to standardize a test and elicit answers that refer to specific latent constructs or facets of a framework. Consequently, VTs should also be investigated for discussion in the VA. Arguably, investigation into how participants understand the VTs is particularly relevant in RobS, as the situations are dynamic, socially constructed

¹¹³ At the time, participants from the University of Münster only had the option of a Polyvalent Bachelor if planning on becoming a teacher. Thus, the two different strands are combined here.

interactions. The meaning of a VT is thus negotiated within the context of the conversation, reflecting the indexicality of language (Bergmann & Meyer, 2021, p. 46; Nunberg, 1993, p. 2). Consequently, VTs can be understood in several different ways by the participants and it should be explored to which extent the intended meaning is conveyed specifically in the RobS. The evidence generated in this chapter can be used to discuss the following claim:

Claim 2: Test

VTs are a constituting element of the RobS and can be considered to be test material as they are to be implemented in a standardized manner by the SL. They are designed to elicit specific facets of the framework (see Chapter 8.2.2). Consequently, the extent to which they do so should be investigated.

RQ3: To what extent do the VTs elicit the intended facet in the participants?

14.2.2.1 Method: Stimulated Recall Interviews

In order to explore the extent to which the VTs elicit the desired facets, stimulated recall interviews (SRIs) are conducted. SRIs are a method for introspective data collection which seeks to “uncover cognitive processes that are not evident through simple observation” (Gass & Mackey, 2016, p. 25). In SRIs, participants retrospectively reconstruct the cognitive process that occurred during a specific, past action in a think-aloud approach, which is initiated by a stimulus (Finkbeiner & Schluer, 2018, p. 28; Heine & Schramm, 2022, p. 174). SRIs are based on the assumption that human beings have “access at some level to their internal thought processes and can verbalize those processes” (Gass & Mackey, 2016, p. 1). Thus, SRIs seek to collect data on what the participants thought at specific moments and thus provide a different perspective on the video-recorded simulation. Arguably, the video only allows for the observation of how people reacted, but does not provide an indication on why they acted in a certain way. For instance, the VT “Good morning” is designed to nudge participants to start the conversation in English. While through video analysis, it can be observed which language the participants speak, it cannot be derived whether the VT was influential in that regard. Thus, SRIs are considered to be the appropriate method to investigate RQ3.

All participants of the pilot survey took part in the SRI. As described, the simulation was filmed from two angles, one which specifically only focused on the SL for the SRI (see Figure 23). This perspective was chosen to reduce the possibility that participants are distracted by watching their own performance and turn towards reflection rather than recollection (Knorr & Schramm, 2012, p. 193).



Figure 23: *Perspective of SRI video stimulus in which only the actor is visible.*

The SRI was conducted about 60 minutes after the RobS was filmed. This approach is in line with literature, which advises to conduct the SRI as fast as possible after the action (Finkbeiner & Schluer, 2018, p. 33). During this time, participants filled out the written survey and the video was watched by the researcher in order to identify time stamps in which a VT was uttered. Before the interview, participants received information of the SRI purpose, specifically that the purpose was not for them to reflect or explain their action, but to verbalize their thought process at that moment. The interviews were conducted in German. During the SRI, both the researcher and participant watched the performance on a large screen via a projector and the interview was recorded via Microsoft Sound Recorder and an external microphone. Every time the SL uttered a (potential) VT, the video was stopped before the participant's reaction was heard. Then, the participant was asked a question like: "What did you think at this moment?" Participants had the option to rewind the video if they did not comprehend the audio. The audio of the interview is transcribed based on the guidelines for the follow-up interviews outlined in Appendix E.1. The SRI transcripts include segments of the simulation, which are transcribed as described in Chapter 14.2.1.

Analysis of the transcripts was done with MAXQDA 2020, based on the structuring qualitative content analysis according to Kuckartz and Rädiker (2022, p. 129, 2023, p. 99). The main categories for analysis were developed deductively, based on the facets of the framework, and the corresponding VTs represent the first level of subcategories. These VT-specific subcategories are then used to code the transcripts. Based on the coded segments, inductive categories are developed which can be used to structure the different cognitive processes

of the participants. In a last cycle, all data was coded with the refined coding system (Kuckartz & Rädiker, 2022, p. 132). The coding manual is described in Appendix F.1. In the following, results of the SRIs will be presented with regard to the VTs that have been developed and presented in Chapter 8.2.2. Presenting and discussing all potential VTs of the simulation is refrained from, as the focus is on the VTs that are designed for specific facets and are used in a standardized manner. Furthermore, not all VTs are necessarily used in each simulation because, for instance, the facet was already addressed by the participants without a VT being required. Thus, only those VTs that were used several times are included for a substantiated discussion.

14.2.2.2 Results of the SRIs

In the following, results will be presented from a predominantly quantitative perspective, with selected statements from the participants provided to contextualize the thought processes. In the tables, the number of coded segments (N_{Code}) and participants (N_P) is presented.

Adaptation of language

All ten participants in the study received the VT "Good morning" at the start of the simulation (see Table 9).

Table 9: Overview VT "Good morning."

Trigger	Code		N_{Code}	$N_P = 10$
"Good morning"	Test situation		1	1
	Language	Conversation is supposed to take place in English	4	4
		Insecurity with regard to language; chooses German	4	4
		Surprise	3	3
		Total (For language)	11	9

When asked what they thought when the learner entered the room and uttered this VT, one person relates to thoughts concerning the general test scenario, while the remaining $n = 9$ participants refer to the SL's choice of language. Some participants take the VT as a call to have the conversation in English:

TP: Ja, dass sie erwartet, dass das Feedback jetzt auf Englisch ist. (IEH20)

(TP: Yes, she expects the feedback to be in English.)

Other participants are insecure or surprised about the learner speaking English, with some actively deciding to speak German nonetheless:

TP: Ich habe mich gewundert, dass sie mich auf Englisch angesprochen hat. Und dann war ich so am überlegen, soll ich jetzt auf Englisch oder

nicht. Und dann habe ich mir gedacht, ähm „Wir machen das lieber auf Deutsch.“ Das habe ich mir da gedacht. (AYW07)

(TP: I was surprised that she spoke to me in English. And then I was thinking, am I supposed to speak English or not. And then I thought, ehm, "Let's rather do it in German.")

Three participants received the VT "Can you say that in German?", either in English or in German (see Table 10).

Table 10: Overview VT "Können Sie das auf Deutsch sagen?" / "Can you say that in German?"

Trigger	Code	N _{Code}	N _P = 3
"Können Sie das auf Deutsch sagen?" / "Can you say that in German?"	Learner's language competence	6	3

All state that they thought about the use of language in that moment:

TP: Dass ich das vielleicht etwas zu kompliziert ausgedrückt habe (lachen). Ähm, ja, ich ne/ Ich glaube, das war mein Fehler. Das war sehr kompliziert vielleicht. Und dementsprechend habe ich dann auch / Da bin ich einmal zum Deutschen gewechselt. Und habe das versucht nochmal auf Deutsch etwas unkomplizierter zu sagen. (SAS05)

(TP: That maybe I was using too complex language (laughs). Ehm, yes/ I believe that was my mistake. That might have been very complex maybe. And accordingly, I have / Then I switched to German. And tried to say it less complex in German.)

Diagnosis – Deep-level: Content

Five participants receive the VT aimed at the text content (see Table 11).

Table 11: Overview VT "Ist vom Inhalt her alles drin?" ("Is everything included with regard to content?")

Trigger	Code	N _{Code}	N _P = 5
"Ist vom Inhalt her alles drin?" ("Is everything included with regard to content?")	Learner's motivation/cognition	2	2
	Participants' lack of knowledge	4	4

Two participants refer to the learner, e.g., their motivation:

TP: Hm. Ja, da habe ich mir überlegt: „Wie kann ich ihr eine hilfreiche Antwort geben, die sie halt auch nicht entmutigt. (MJZ08)

(TP: Hm. Yes, there I was thinking: How can I give an answer which helps her but does not demoralize her.)

Interestingly, four participants also state that they themselves did not know whether the relevant pieces of information are included in the text:

TP: Ich war mir nicht sicher, ob inhaltlich alles drin war, ähm, weil ich die Liste halt erst recht spät angeguckt habe (GTM21)

(TP: I was unsure whether everything was included from a content perspective, because I looked at the list relatively late.)

Diagnosis – Deep-level: Structure

Three participants receive the VT with regard to text structure (see Table 12).

Table 12: Overview VT “Ist das vom Aufbau her in Ordnung?” (“Is the structure okay?”)

Trigger	Code	<i>N_{Code}</i>	<i>N_P = 3</i>
“Ist das vom Aufbau her in Ordnung?” (“Is the structure okay?”)	Learner’s motivation/cognition	2	2
	Text Structure	1	1

While one participant simply discusses the structure of the learner text in the interview, two participants describe thoughts with regard to the learner’s motivation:

TP: Da habe ich mir gedacht, jetzt muss ich was Positives sagen. (lachen)
 [...] Weil ich / Ich hatte das Gefühl, ich habe ganz viel Negatives gesagt,
 und das ist dann ja schon immer so nicht so motivierend auch. Man
 sollte auch irgendwie was Positives noch hervorbringen nämlich.
 (AYW07)

(TP: There I thought, now I have to say something positive. (laughs) [...])
 Because I/ I had the feeling that I had said a lot of negative things and
 that is not really motivating. You should also always state something
 positive.)

Adaptation of content

Several VTs that are designed to generally nudge participants to check understanding were used in the pilot survey. The VT “What did I do wrong here?” was received by five participants (see Table 13).

Table 13: Overview VT “Was habe ich hier falsch gemacht?” (“What did I do wrong here?”)

Trigger	Code	<i>N_{Code}</i>	<i>N_P = 5</i>
“Was habe ich hier falsch gemacht?” (“What did I do wrong here?”)	Learner’s motivation/cognition	4	4
	Participants’ competence	1	1

One participant states that they were unsure about the nature of the error when that VT was uttered. Four participants describe cognitive processes that relate to the learner and their level of competence:

TP: [...] Habe ich mich dann gefragt so, ob ähm sie jetzt nicht verstanden
 hat, wenn ich sage, also wenn ich ihr einfach nur zeigen wollte, das ist
 ein unregelmäßiges Verb oder ob ihr eben zeigen wollte, das ist in dem
 Fall jetzt ein Fehler (VJD15)

(TP: [...] I was wondering whether she did not understand when I say, so
 when I just wanted to show that this is an irregular verb or in this case
 this is an error.)

The VT "Is this correct here?" was stated in five simulations (see Table 14).

Table 14: Overview of the VT "Ist das hier richtig?" ("Is this correct here?")

Trigger	Code	<i>N_{Code}</i>	<i>N_P = 3</i>
"Ist das hier richtig?" ("Is this correct here?")	Learner's motivation/cognition	3	2
	Participants' competence	3	2

For two participants, the VT initiated thought processes about their own competence:

TP: Da habe ich gedacht, ist das echt richtig oder habe ich da/ Also ich habe mich da selbst nochmal äh angezweifelt ähm weil ich mir da nicht sicher war, genau und ich habe auch nicht richtig verstanden was sie meint, weil die Schülerin so sehr vage gesprochen hat so vage, also das hier so, was genau meinst du. (HKH07)

(TP: There I thought, is this correct or did I / So I was questioning myself because I was not sure, right, and I did not understand what she meant, because the learner spoke in vague terms, so here, what exactly do you mean.)

Two participants describe that they thought about the learner's competence in that situation:

TP: Achso, genau. Äh. Da habe ich mir gedacht, dass sie es richtig erkannt hat. (AYW07)

(TP: Alright, Eh, I thought that she identified it correctly.)

The VT "I did not understand that." was used in two simulations (see Table 15).

Table 15: Overview of the VT "Das habe ich nicht verstanden." ("I did not understand that.")

Trigger	Code	<i>N_{Code}</i>	<i>N_P = 2</i>
"Das habe ich nicht verstanden." ("I did not understand that.")	Learner's motivation/cognition.	1	1
	Participants' competence	1	1

One participant, again, was focused on their competence in that situation, while the other participant considers adapting the complexity of their comments to the learner's level:

TP: Ja, ist jetzt schwierig. Ähm, sie hat wahrscheinlich gar nicht verstanden, was überhaupt eine „conjunction“ ist und dann hätte ich ihr das eigentlich nochmal erklären müssen. Aber das war jetzt aufgrund der wenigen Zeit, also, hätte es ja auch nicht ausgereicht. Deswegen bin ich da gar nicht irgendwie näher drauf eingegangen. (IEH20)

(TP: Yes, now this is difficult. Probably, she did not understand what a "conjunction" is and I should have explained it to her. But because of the limited amount of time, it would not have sufficed. This is why I did not further address this.)

Self-regulation / Feed Forward

All ten participants in the pilot survey were at some point confronted with the VT "How can I practice that?" (see Table 16).

Table 16: Overview of the VT "Wie kann ich das üben?" ("How can I practice that?")

Trigger	Code	N _{Code}	N _P = 10
"Wie kann ich das üben?" ("How can I practice that?")	Participants' competence	7	4
	Options for support	10	10
	Learner's motivation / cognition	6	6

All participants stated that, upon hearing the VT, they thought about specific options they could provide so that Mia could take action, often this related to the level of self-regulation, e.g., ways to learn vocabulary:

TP: Dass ich dann der Schülerin einige Handlungsmöglichkeiten aufzeigen kann. Also, wie sie tatsächlich diese Vokabeln dann am besten lernen kann. Vielleicht verschiedene Methoden. Einem fehlt da natürlich / also ich kannte sie ja jetzt nicht in dieser Simulation, aber teilweise kann man ja dann schon mal auf, ähm, vorheriges Verhalten oder generell die, ähm, wie sie lernt darauf zurückgreifen und dann schauen, okay, was passt am besten auf diese Schülerin. (ESG19)

(TP: That I can show the learner options to take action. Meaning, how she can actually learn the vocabulary the best possible way. Maybe different methods. One is of course missing / So, I did not know here in this simulation, but partially one could refer to prior behavior or generally how she learnt, draw on that and then look, okay, what suits the learner best.)

In addition, five participants make remarks about the learner's motivation or cognition. In two instances, this was considered negatively, but in four cases rather positively, as it was perceived as interest in learning, for instance:

TP: Äh. Ich fand es / Ich habe mich ein bisschen darüber gefreut, dass sie nachgefragt hat, so, weil das ja immer so ein bisschen auch Interesse zeigt, so dass sie es lernen möchte. (VJD15)

(TP: Eh. I thought it / I was happy that she asked about it, because this shows interest that she wants to learn.)

Moreover, four participants note difficulties in answering the question properly as they lacked the knowledge at that moment:

TP: Ähm. Ja, da geht es ja sozusagen so ein bisschen darum irgendwie Lernstrategien irgendwie aufzubauen und wenn sie das jetzt noch nicht hat in der neunten Klasse / also da wusste ich auch nicht wirklich, wie ich darauf jetzt reagieren soll, weil, ja. (IEH20)

(TP: Ehm. Yes, this is about developing learning strategies and if she does not have that in the ninth grade / I did not really know how to react, because, yeah.)

Revision

The VT aiming at revision was stated in nine simulations. In one simulation the participant did not need the VT to discuss the matter of revision, as they did so without (see Table 17).

Table 17: Overview of the VT "Muss ich den Text nochmal neu schreiben?" ("Do I have to write the text again?")

Trigger	Code		N_{Code}	$N_P = 9$
"Muss ich den Text nochmal neu schreiben?" ("Do I have to write the text again?")	Learner's motivation / cognition		6	6
	Development of writing competence	Confirming the necessity	5	5
		No necessity	1	1
		General discussion	1	1
		Total (Developing writing competence)	7	7

Seven participants state that they thought about revision in the context of developing writing competence. While one person states that revision was not necessary, and one participant does not take a stance and discusses the issue in general, five participants confirm the necessity of revision:

TP: Eine gute, authentische Frage, ähm, weil das ist immer die Angst von Schülern, dass das ganz schlecht ist und dann das alles immer neu machen muss, aber, ähm, meine Antwort war ja auch: „Ja“, ähm, aber es ist ja ein Entwurf und deswegen ist das / diese Angst muss man finde ich Schülern dann nehmen, um / weil es nicht schlecht ist etwas neu zu schreiben und es muss ja auch nicht alles neu formuliert werden, aber nur halt verbessert werden. (GTM21)

(TP: A good, authentic question, ehm, because learners are always afraid that it is really bad and that they have to do it all over again, but eh, my answer was "yes", but ,ehm, it is a draft, therefore this is / one needs to take this fear away from the learner to / because it is not bad to write something new and it is not the case that everything has to be formulated from the beginning, but it has to be improved.)

Additionally, six participants describe thought processes that relate to the learner. For instance, for one person, the VT elicited a negative reaction with regard to the learner's willingness to learn:

TP: Ähm, dass sie sich ein bisschen jetzt vor Arbeit irgendwie drücken will. Also, wenn es jetzt schon eine Herausforderung darstellt, den Text jetzt nochmal zu überarbeiten oder da nur abzuschreiben, sagt sie jetzt ja eigentlich dann, ja, habe ich mich so ein bisschen gefragt so: „Wieso hast du es dann eigentlich abgegeben, ne? Wenn du nicht bereit bist, jetzt da irgendwie noch was dran zu ändern oder so.“ (IEH20)

(TP: Ehm, a little bit that she wants to avoid working. So, if it is a challenge to revise the test or just copy it, now she actually says, well, I was

wondering: "Why did you hand it in? When you are not willing to change something about it.")

Contrarily, another participant stated that they perceived this VT as positive:

TP: Hm. Ja muss. [...] Und halt auch wieder, sie zeigt Arbeitsbereitschaft. Also sie ist willig da dran zu arbeiten. (MJZ08)

(TP: Hm. Yes, it is a must. [...] And again, she shows the willingness to work. She wants to work on it.)

14.2.2.3 Discussion and Limitations

In order to elaborate on RQ3, the described results will be discussed with regard to the extent that participants describe thought processes that align with the targeted facet of the framework. The VT "Good morning" elicits thought processes that relate to the language of the conversation in nine of the ten participants. This overwhelmingly indicates that the VT works as intended. Probably, participants were quite surprised by the learner speaking to them in English as they expected them to speak in German, as indicated by the statements. Thus, the unexpected nature of the VT might have been contributed to the fact that the aspect of language was addressed often by the participants. Moreover, the issue of language might have also been on the participants' minds, as they often asked in which language the simulation would take place before the RobS started. The question was always answered identically with: "This is your choice." Similarly, if asked by the SL in either German or English whether they could repeat an English statement in German, participants always wondered about the learner's level of language comprehension and ways to adapt appropriately. Consequently, the VTs for *Adaptation of language* evoke the desired facet of the framework satisfactorily.

The VTs that are designed for the specific aspects of diagnosis, i.e., *Deep-level: Structure* or *Deep-level: Content* were not received by all participants, partly because the participants addressed these issues without requiring a VT. The content-focused VT evokes an interesting reaction in that most of the participants that received the VT thought about the content, but questioned themselves and were not sure if the learner text is satisfactory in that regard, as they did not necessarily prepare to discuss content. This might indicate that participants focus more on language related aspects in the RobS. Similarly, the structure-specific VT led one of the three participants to actually discuss the structure of the learner text in the SRI. For both VTs, interestingly, two participants addressed the learner's motivation, trying to use the VT as an opportunity to provide positive feedback. Potentially, the VT here led the participants to believe that the SL felt slightly demoralized and tried to receive some confirmation of successful elements in the task. Nonetheless, while the VT seems to elicit reactions with regard to the learner's motivation, they generally do so in the context of the

intended facet. Arguably, VTs for the specific diagnostic dimension of texts should, thus, only be used if the text dimension is not addressed by the participant.

For the facet of *Adaptation of content*, three different VTs were used. This variety can be reasoned with the more context-dependent nature as it cannot fully be anticipated which aspects the participants address in the simulation and how they do it. The participants' statements for all three VTs can be categorized as either directed at the learner's motivation or cognition, or the participants' own competence. With regard to the latter, some participants, not unlike with the VT on *Deep-level: Content*, react with uncertainty, questioning their own competence as they are unsure why a certain element in the text is right or wrong. Mostly, however, the VTs led participants to think about the SL's abilities and how to adapt accordingly. As the VTs for this facet are generally constructed to not evoke a reaction by the participant, but to create an environment in which the participant is more likely to check whether the learner understood aspects of the feedback conversation, the VTs are deemed to be satisfactory.

The VT "How can I practice that?", which is designed to elicit a reaction for the facets of *Self-regulation* and *Feed Forward* was received by all the participants. All participants state that after the VT, they thought about specific options that the learner could take, i.e., next steps. Often, this related to the level of self-regulation, as participants thought about strategies the learner could use for their learning process. Similar to the previous VTs, this one also seems to lead to thoughts about the learner's motivation or cognition. What is more, several participants also state that they had difficulties in providing the SL with appropriate options to practice, as they did not have the knowledge to do so, indicating that this VT is generally more difficult for the participants to master. While generally more complex for the participants, the VT evoked the intended facet quite satisfactorily.

The final VT is aimed at the *Revision* of the text and was received by nine participants, as one participant addressed this topic without the trigger. Seven participants state that they thought about the benefits and drawback of rewriting the text. Interestingly, six participants discuss this VT from a mostly motivational perspective, as some regard the VT as an expression of motivation, while others regarded it as a lack thereof. Probably, as this VT usually occurred at the end of the RobS, this factored into a general perception of the SL wanting to finish the conversation. Consequently, the VT still functions as intended although the participants also perceive it as an indication of the learner's motivation.

Overall, the VTs can be considered satisfactory in that they elicit the desired thought process in most participants. Especially the VTs for *Adaptation of language*, *Self-regulation/ Feed Forward* and *Revision* show that the VTs work as

intended. As described by the indexicality of language, VTs might be understood in different ways by the participants, as the data also indicated. Notably, the SL's motivation is often addressed by the participants as well for the latter VTs of *Self-regulation/ Feed Forward* and *Revision*, i.e., those that indicate that the SL might need to act subsequently. Apparently, VTs for the specific diagnostic facets (i.e., *Deep-level: Content* and *Deep-level: Structure*) also evoke reactions with regard to the SL's motivation. While too much emphasis on motivation is not intended, as discussed in the test construction in Chapters 7 and 8, it is unlikely that participants ignore this aspect fully. This is not surprising, considering its relevance in the classroom and that motivational aspects are part of competence (Weinert, 2014, pp. 27–28). However, motivation is not necessarily the dominant aspect that is triggered by the VTs. Thus, as the intended facets prevail, the VTs can still be used as they evoke the desired facets sufficiently.

Naturally, there are also limitations that need to be addressed, especially with the method of the SRI. As with all think-aloud procedures, there is a debate about the extent to which the verbalized statements accurately represent the cognitive processes, as there might be distortions caused by the process of verbalization (Aguado, 2018, p. 16). This distortion is potentially reinforced in stimulated recalls, as the process of verbalization occurs after an action has occurred. Thus, the statements might not necessarily refer to the participant's thought in that situation, but rather, to what they believed their thoughts were. Moreover, social desirability might have factored into the participants' answers, as they might have described potential thought processes which they believed they were supposed to have had, rather than describing their actual cognitive processes. Statements from SRIs can thus only be considered an approximation of actual cognitive processes. One way to substantiate the statements is by the means of triangulation. In this study, for instance, the SRI statements could be compared to the participants' actual behavior in the simulation. However, it was not done as the potential benefit was not considered to be substantial with regard to evaluating the appropriateness of the VTs.

14.2.2.4 Conclusion

VTs are an important tool for standardization in the RobS as they can be used systematically to elicit specific reactions from the participants. In order to investigate the extent to which the VTs function as intended, SRIs are conducted as part of the pilot survey. In these, participants re-watch the simulation and are asked to reconstruct their thought processes after the VTs were uttered by the SL. Results indicate that the VTs evoke reactions that align with the intended outcome. However, there are also additional findings, such as the fact that the SL's motivation is often addressed as well, as is the participant's own inability to

adequately answer some VTs. Consequently, the described VTs are considered to fulfill their function satisfactorily and are used in the main survey.

14.2.3 Deciding on the Test Documents

An additional point of interest in the pilot survey is the suitability of the material. Three main questions need to be considered:

- 1) Which of the two learner texts is best suited for the main survey?
- 2) Is the remaining material sufficiently authentic?
- 3) In which language should the task be designed?

In order to discuss these questions, results from the guideline interviews as well as observations from the RobS will be drawn upon.

Learner text

Concerning the first question, no substantial difference of quality in the RobS could be identified concerning the texts. Both texts allowed participants to conduct the feedback conversation and the actors to use the VTs. When asked if the learner texts were authentic in the interview, seven of the ten participants argue that their text was authentic. For instance, one participant states the following:

TP: Also es ist sicherlich kein Text, der von einer Schülerin ganz oben steht, aber das war ja auch gesagt, dass das so ist, von daher denke ich schon, dass das relativ realistisch ist. (SAS05)

(TP: Of course, this is not a text of a learner on a high level, but this is also what was said, thus, I think that this is relatively realistic.)

Three participants indicate that they do not fully consider the texts to be authentic, two of which said this about text sample B:

TP: Das weiß ich gar nicht so richtig, muss ich sagen. Ähm ich kann mir vorstellen schon, obwohl ich schon das Gefühl habe, dass er äh sehr viele Fehler hat und alles, was richtig ist, eigentlich aus dem Arbeitsblatt übernommen ist. Äh ich hätte gedacht, dass das Niveau ein bisschen höher wäre. (CAH06)

(TP: I don't really know, I have to say. Eh, I can imagine yes, although I have the feeling that it has many mistakes and everything that is correct has been adapted from the task sheet. I would have thought that the level would be slightly higher.)

In the end, the choice was still made for text sample B for the main survey. While both texts can be considered suitable, text B more overtly allows for the provision of feedback on the text structure, as it does not include a headline. Therefore, this text was chosen in order to facilitate the performance of the participants regarding that aspect.

Remaining material

In terms of the second question, the remaining documents are focused on. Concerning the general situation and the formulation of the context and task (M0), a general authenticity is perceived by most participants:

TP: Hm. (...) Ich weiß nicht, ob eine Schülerin, die / Doch eigentlich schon. Dadurch, dass sie eigentlich besser ist als / Also, dass sie eigentlich zwischen drei und vier steht, jetzt aber eine fünf plus hatte, hat sie eigentlich einen großen Anreiz ähm mich zu fragen, was man besser machen könnte, weil sie eigentlich weiß, dass sie es besser kann und dass da jetzt anscheinend Fehler passiert sind, die sie noch normalerweise wahrscheinlich nicht macht. Deswegen finde ich es schon eine authentische Situation und auch die Frage ‚Können Sie bitte schauen, ob das so in Ordnung ist?‘ klingt auch authentischer als zu sagen ‚Können Sie das bitte einmal feedbacken?‘ oder irgendwie sowas. (CAH06)

(TP: Hm. (...) I don't know whether a learner who / well, actually yes. Due to the fact that she is actually better / So, because she is between grades three and four, but now she had a five plus, she is incentivized to ask me what she can do better, because she actually knows that she can do better and apparently mistakes occurred which she usually does not make. Thus, I do believe that this is an authentic situation and also the question "Could you please have a look at this? I am not sure, if all of this is okay." Is more authentic than to say "Could you please feedback this?" or something like this.)

The task itself (M2) is ambivalently perceived by the participants. While the format and the task are perceived to accurately depict the reality of actual instruction by most participants, the actual task is discussed skeptically with regard to the relevance for the learners' environment:

TP: Ja ich weiß nicht, es hat halt nicht so viel mit der Lebenswelt der Schüler zu tun, [...] aber Heutzutage vielleicht eine Mail oder, ich weiß nicht, also kommt auf die Schüler an aber vielleicht kann man es ein bisschen mehr an die Lebenswelt der Schüler anpassen.

I: [...] die Frage zielt eher so darauf ab, ob das so eine typische Aufgabe ist, die man in der Schule findet.

TP: Ah eine typische Aufgabe, ja vielleicht klar, also in Schulbüchern die noch so veraltet schon sind schon. (MJZ08)

(TP: I do not know, if this so relevant for learners' environment, [...] but nowadays maybe an e-mail or. I don't know, it depends on the learner but maybe it can be adopted more towards the learners' environment

I: [...] the question was rather focused on whether such a task is typical for schools.

TP: Ah, a typical task, yes, maybe, like in text books that are outdated.)

Other aspects that are criticized are the scope of the information, including the list of facts and images, and the lack of scaffolding. One participant explicitly states that they do not think the task was typical for instruction. Regarding the

series of lessons (M3), participants are generally positive regarding its authenticity, but weigh it against their experience of EFL instruction:

TP: Ja definitiv, obwohl an den meisten Schulen, ich war jetzt schon an mehreren Schulen tätig, da wird der Englischunterricht vermehrt am Buch tatsächlich orientiert, aber ansonsten ist das eine ziemlich schöne Reihenplanung, die man auch so definitiv so durchführen könnte, ja. (ESG19)

(TP: Yes, definitively, although I have worked at multiple schools, English lessons are usually more oriented towards the text book, but otherwise, this is a nice series of lessons which one could definitely execute as described.)

One skeptical remark relates to the notion of peer feedback as a component in the subsequent lessons, which is discussed critically. The sample text (M4), including the vocabulary aid, is also generally accepted as an adequate exemplar. However, some participants consider the text to be too long and complex for the age group:

TP: Ähm. Ich fand den ein bisschen lang, wobei es da glaube ich auch immer wieder auf die Klasse ankommt, aber ansonsten ähm finde ich schon. Ich habe mich nur gefragt, ob da nicht ein bisschen zu viel Wörter, weil da auch so wörtliche Rede, indirekte Rede, dann diese Auslassungszeichen überall, ob das nicht ein bisschen viel dann auch ist. Aber es sieht, wie gesagt es kommt auf die Klasse immer an natürlich. Also es ist / Ich würde jetzt nicht sagen, man kann den gar nicht nehmen, das. (VJD15)

(TP: Ehm. I thought the text was a bit too long, albeit I believe it always depend on the class, otherwise, I think it is okay. I was just wondering, whether there might be too many words, because also direct speech, indirect speech, then these signs of omission everywhere, whether this might be a bit too much. But it looks, as I said, it always depends on the class, of course. So it is / I would not say you cannot use the text at all.)

In the end, no further modifications were made to the material after the pilot, despite the outlined criticism by the participants. While the addressed aspects are certainly not without merit, they apparently did not prohibit or distract participants from performing in the simulation. What is more, the RobS aims at an approximation of practice, which allows for some deviation from reality, as long as it does not impede the performance. Moreover, several aspects that are mentioned do not directly question the authenticity, but the actual appropriateness of the task or text for the class, including discussions on how to improve the material according to the participants' opinion. While some of these modifications are logical, such as scaffolding in the task sheet, these choices are deliberately made for the RobS. For instance, no scaffolding is provided so that participants have to decide themselves which specific aspects to focus on, and not have this information be indirectly predetermined by the task sheet. Also, in

order to accurately evaluate the use of content by Mia, an extensive list with information has to be provided, even though this results in a less structured task sheet.

Language of the material

Considering the third aim, there is the question whether the vignette, which includes the description of context and the learner as well as the task, should be in German or in English. The underlying assumption is that a German vignette is more accessible to participants and is more authentic to the context of a German school, while a vignette in English would create a stronger incentive for the students to actually start the conversation in English, a necessary requirement for the previously outlined aspect of *Adaptation of language* in the framework (see Chapter 7). The vignette was thus handed out to the participants in either English or German and they were asked about the material in the follow-up guideline interviews. Several participants state that the language of the material influenced them in their decision with regard to language used:

I: [...] Warum haben Sie das Gespräch dann doch auf Deutsch geführt?

TP: Hm weil alle Anleitungen, alles, die Aufgabe, alle Materialien waren auf Deutsch. Das ist ein Punkt und ich war dadurch da man da nicht vorher drüber geredet hat, fiel es mir gerade schwer direkt in Englisch zu wechseln und ich war, fühlte mich sicherer dabei, jetzt gerade auf Deutsch zu sprechen. Aber ich glaube, wenn man das vorher angesprochen hätte und vielleicht auch Teile der Einleitung auf Englisch gemacht hätte oder Hinweise gegeben hätte, dass man das Gespräch auf Englisch führen sollte, hätte ich das auch so gemacht. (GTM21)

I: [...] Why did you have the conversation in German?

TP: Hm. Because the instructions, the task, the material was in German. This is a point and because it was not spoken about, it was difficult for me to switch to English immediately and I was, I felt more safe to speak in German. But I believe, if this as addresses beforehand and maybe parts of the introduction were in English or there were hints that the conversation was to be conducted in English, I would have done so.)

As the participants did not seem to have issues understanding the vignette in English and the choice of the language in the material apparently influences the participants, the English language material is chosen for the main survey, as this makes the desired outcome of an English start to the conversation more likely.

14.2.4 Deciding on the Research Procedure

After the pilot survey, several modifications were made to the follow-up. First, the number of questions is reduced in the guidelines. Primarily, questions that aim at the authenticity of material are excluded, as the development phase of the simulation material, i.e., documents and VTs, was considered to be completed for the main survey. Thus, the SRIs are not included in the main survey,

either, as they focused on the VTs. Moreover, conducting the SRIs took a lot of resources and is therefore dropped for the main survey to allow for a shorter duration of participation. Similarly, the C-test is not included as a part of the follow-up in the main survey because it extends the length of participation. This decision is supported by the fact that the results of the C-test do correlate positively with the LexTALE score in the pilot sample ($N = 10$), $\rho = .65$, $p = .046$. While these results have to be treated with caution due to the small sample size, they indicate that, in order to obtain data which can be used to make inferences about language competence, the shorter test suffices. Moreover, due to technical reasons, the C-test could not be conducted as part of the digital survey, and required a pen-and-paper format. Thus, additional analogue scoring resources would have been required, while the LexTALE test was scored automatically using a syntax in SPSS.

14.3 Summary

The pre-pilot phase includes the creation of the material, specifically the learner texts. These learner texts are hybrid, as they are artificially constructed but draw on authentic material. In several steps, several texts are designed and discussed with stakeholders to ensure that the texts are authentic but still allow participants to display the facets that are outlined in the framework. The remaining documents are also discussed with experts in the field. Three people are trained to portray the role of Mia in a standardized way using specific VTs. In the pilot survey, the RobS is explored for the first time with the target group. The aim of the pilot phase is to investigate the VTs, explore the authenticity and appropriateness of the material for the RobS, and to refine the research process. Using the method of SRIs, the extent to which the VTs elicit the desired facets of the framework is explored. Results indicate that, while VTs elicit cognitive processes that relate to the participants' lack of knowledge, or the SL's (lack of) motivation, they address the intended facets satisfactorily. Still, results need to be regarded with caution as to the limitations of SRIs, e.g., they only allow for the approximation of actual cognitive process of the participants. Regarding the remaining documents, no further modifications are made for the main survey as the material is considered to be sufficiently authentic for the RobS. Concerning the research process, the number of questions in the guideline is reduced, the SRIs not included, and the C-test removed from the written survey as the LexTALE test is considered sufficient to obtain data for language competence. In the following chapter, the four studies that are conducted within the main survey are presented.

15 Main Survey

In this chapter, the four studies that relate to the main survey are presented. The purpose of the main survey is to collect data that can be used to discuss the claims outlined in the IUA. The studies which are presented here rely, to some extent, on the data that was collected within the main survey and are thus subsumed in one chapter. Therefore, the data collection process and the sample characteristics of the main survey will be described first, before the individual studies are focused on. Study 1 focuses on the perception of the participants and relies on their statements made in the follow-up guideline interview. Study 2 investigates experts' judgements of video-recorded performances and the scoring instrument to discuss its appropriateness. Study 3 includes a detailed contemplation and analysis of the statistical data that is generated by the application of the scoring instrument, considering descriptive statistics as well as aspects of (inter-rater) reliability. Study 4 explores the relation that the generated test scores have with other external variables. All four studies generate information that can be used to make inferences about the claims outlined in the IUA (see Chapter 11.2)¹¹⁴.

15.1 Data Collection and Sample Characteristics

In this chapter, the process of data collection in the main survey and the sample characteristics are described. This relates to the PSETs who participated in the RobS and the follow-up survey and interview.

15.1.1 Data Collection¹¹⁵

The main survey was conducted from November 2022 - June 2023 at Paderborn University, with a small portion of the study also taking place at the University of Münster in March 2023. Participants were recruited by way of an extensive promotional campaign. Similar to the pilot survey, the project was presented both in the introductory TEFL lecture and in selected TEFL seminars, and was also promoted by lecturers via e-mail. Additionally, posters were distributed in the relevant buildings in the university and leaflets distributed in the university canteen. Additionally, participants in one Bachelor seminar and in one Master

¹¹⁴ For reasons of cohesion, instead of having a summary of all the empirical findings for Chapter 15, each study will have a distinct summary in this chapter. A summary of Part IV is still provided in Chapter 16.

¹¹⁵ The data collection process resembled that of the pilot study. In this chapter, only relevant modifications will be outlined. Thus, see Chapter 14.2.1.1 for a detailed description.

seminar had the option to participate in the project as part of their credit¹¹⁶. To ensure that participation remained voluntary, the participants also had the option to only do the RobS, without the follow-up and without their data being used for the project. Also, an equivalent option in the form of, e.g., a portfolio was offered so that no participant had to participate in the RobS if they did not want to. Moreover, an additional set of data was collected at the University of Münster¹¹⁷, thanks to the support of the English language education team who recruited participants and provided the rooms. Similar to the pilot survey, participants were informed about the duration of the procedure (approx. two-and-a-half hours) and the 25€ reimbursement. For scheduling, participants also registered in a moodle-based course and were able to select a time slot of their choice.

The data collection process is described in detail in the context of the pilot survey in Chapter 14.2.1.1. The process in the main survey resembles the pilot survey in most aspects and differs only in several minor points (see Table 18). Of the performances in the main survey, 27 simulations are transcribed, using the extended content-semantic transcription guidelines also used for the follow-up for the audio level (Dresing & Pehl, 2018, pp. 21–23; see Appendix E.1). Similarly to the pilot survey, the transcription of the video level is inspired by GAT-2 conventions (Couper-Kuhle et al., 2011, pp. 7–17).

Table 18: *Data collection process in the main survey.*

Step	Focus	Minutes
Greeting & Organization	Information on the project & GDPR. Agreement on Participation.	10
Simulation Procedure		
Preparation	Use the material to prepare for the feedback conversation Opportunity for questions	15
Simulation	Simulate a feedback conversation with an SL.	10

¹¹⁶ For example, participants had to write a reflective essay, using theoretical models to analyze their performance in the RobS and provide alternative options for performance.

¹¹⁷ The rooms, process and arrangements of the portion of the study that was conducted at the University of Münster closely resembled the setting in Paderborn and are thus not further described in detail.

Step	Focus	Minutes
Follow-Up		
Written Survey	Demographic Data Study Progress	5-10
	Feedback Knowledge	15
	PCK English	60
	Lexical knowledge	5
Guideline Interview	Authenticity of • Situation • Actor Fairness Validity	10-15
Conclusion		
Debriefing	Options for further questions Opportunity for reflection	5-10

15.1.2 Sample Characteristics

In the main survey a total of $N = 51$ pre-service EFL teachers participated. Based on the previously described recruitment procedures, the sample can be described as a convenience sample. Characteristics are presented in Table 19.

In the sample, $n = 22$ (43.1 %) participants identify as male and $n = 29$ (56.9 %) as female. The gender relation skews slightly more male compared to the gender relation in the population of students in teaching degree programs, i.e., 67.6 % female at Paderborn University in the winter term 2022/2023 (Universität Paderborn, 2023). The sample has a median age of 24 years, with the oldest person being 39 years old (and having just started the Bachelor program). The average of the last graduation grade is $M = 2.44$ ($SD = 0.54$). In the sample, $n = 45$ participants stem from Paderborn University, $n = 6$ from the University of Münster. With regard to the phase of study, more participants are in the Bachelor phase ($n = 30$) than in the Master phase ($n = 21$). The average number of enrolled semesters is $M = 9.40$ ($SD = 4.25$). The study programs, differentiated by type of school, show that the vast majority of students in both Bachelor and Master phase are studying to teach in upper secondary schools (*Gymnasium / Gesamtschule*).

Table 19: Sample characteristics of the main survey.

		<i>N</i> ¹¹⁸	<i>n</i>	%	<i>M</i>	<i>SD</i>	Median
Gender		51					
male			22	43.1			
female			29	56.9			
Age		51			24.35	3.27	24.00
Final graduation grade		51			2.44	.54	2.50
University		51					
Paderborn			45	88.2			
Münster			6	11.8			
Phase of study		51					
Bachelor of Education / Polyvalent Bachelor ¹¹⁹			30	60.80			
Master of Education			21	39.20			
Study program		51					
Bachelor of Education / Polyvalent Bachelor	Primary School (G)		3	5.90			
	Lower Secondary (HRSGe)		3	5.90			
	Upper secondary (Gym/Ges)		23	45.10			
	Vocational school (BK)		2	3.90			
Master of Education	Special Needs Education (SP)		1	2.00			
	Lower Secondary (HRSGe)		1	2.00			
	Upper secondary (Gym/Ges)		16	31.40			
	Vocational school (BK)		2	3.90			
Number of semesters		50			9.40	4.25	10.00
Curricular internships							
Bachelor of Education / polyvalent Bachelor	AOI	50	38	76.00			
	OI	51	36	70.59			
Master of Education	SIS	45 ¹²⁰	18	40.00			
Extracurricular school-related experience		51					
Yes			33	64.70			
	Months of experience				10.65	9.18	8.00
No			18	35.29			

¹¹⁸ Number of valid answers, excluding missing responses.¹¹⁹ At the time, participants from the University of Münster only had the option of a Polyvalent Bachelor if planning on becoming a teacher.¹²⁰ Besides four missing responses, two participants provided inconclusive answer (see Chapter 15.5.4.1).

Looking at school-related experience, most participants have completed the AOI ($n = 38$) and/or the OI ($n = 36$), while $n = 18$ participants have already completed the SIS or state they are currently in that phase. Some participants, $n = 18$, state that they have some extracurricular school-related experience, with the average number of months of experience within that group being $M = 10.65$ ($SD = 9.18$). The person with the most experience reports having 36 months of school-related experience, while two participants state that they have one month of experience as the lowest non-zero number of months. While most characteristics such as gender and age align roughly with the general population of teacher students at the respective university, it is not representative.

The sample size is comparable to that of other similar studies, e.g., in physics didactics. The sample in Bartels' study (2018) on the validation of a video-vignette test has $n = 57$ participants who are pre-service teachers, Kulgemeyer & Tomczyszyn have a sample of $N = 64$ participants in their validation of a test on explaining competence (2015), and the study by Vogelsang (2014), in which teacher trainees and pre-service teachers are video recorded in actual instructional situations, has a sample of $N = 22$. A recent pilot study for an OSCE for psychotherapists has a sample of $N = 39$, and a study using simulated role-plays as a measurement for diagnostic competence of pre-service mathematics teachers has a sample of $N = 63$ (Kron et al., 2022). Only studies which do not require the physical presence of participants but use digital simulations or formats usually have larger sample sizes ($N > 100$), for example the studies within the project COSIMA (Nickl et al., 2024; Stürmer et al., 2024).

15.2 Study 1: Pre-service EFL Teachers' Perspective on RobS as Assessment

In this chapter, PSETs' perspectives on the RobS and its potential use as assessment are explored in terms of two different aspects: The specific RobS which is developed in this thesis and the concept of RobS as an assessment format.

Regarding the first point, the participants' perception of authenticity and the tested competences is of interest. PSTs generally argue that teacher education, including assessment in its current form, does not prepare them adequately for practice (see Part I). Implementing new performance-oriented assessment formats like RobS, might be one way to address this issue because the concrete form of assessment might influence the learning behavior and have a positive backwash effect. As outlined in Chapter 4.4, especially formats that are perceived as authentic might have a positive impact on learning behavior, potentially leading to deeper learning (Gulikers, Kester, et al., 2008, p. 182). What is more, authenticity is particularly relevant to the RobS because it is designed to allow participants to demonstrate their competence in a situation that approximates practice. If the RobS is not perceived as sufficiently authentic, it cannot claim to

fulfill its purpose. As authenticity is a subjective construct, it is essential that the target group's opinion on the RobS' authenticity is taken into account. Furthermore, the RobS is designed to specifically assess the competence of providing feedback on a text, which is outlined in the framework in Chapter 7. Consequently, participants who have experienced the RobS should describe competences which they believe are being tested that align with the framework. Otherwise, the aspect of content validity might be in jeopardy.

With regard to the second point, the participants' perspective on fairness and potential implementation of RobS is relevant. Fairness is considered as the extent to which participants believe that the assessment format allows them to show what they have learned, i.e., what they believe their actual competence to be. It is relevant to explore perceived fairness as this indicates the extent to which the RobS allows participants to perform. Moreover, perception of fairness might also influence learning behavior, similar to authenticity (see Chapter 4.4.2). The developed RobS is a potentially new format of assessment, designed to address the perceived gap in assessment in teacher education. For a holistic perspective on a potential implementation, it is essential to include the perspective of those that might be subject to assessment. Thus, the participants' perspective of a potential implementation of RobS is also focused on in this chapter. To explore the participants' perception, guideline interviews are conducted. The findings described in this chapter can be used to discuss the following claims:

Claim 2: Test

For the test to fulfill its purpose of approximating practice (Grossman, Compton, et al., 2009, p. 2076) and allow participants to be assessed on the level of *shows how* (Miller, 1990, p. 63), it has to be considered sufficiently authentic by the participants, i.e., those that are being assessed (Gulikers et al., 2004, p. 69). Furthermore, participants' perspectives on the competences that are tested in the RobS are relevant to broaden the scope of content validity and discuss the extent to which the RobS reflects the facets of the framework.

RQ4: To what extent do participants perceive the RobS as authentic?

RQ5: What competences do participants believe are being tested by the RobS?

Claim 3: Performance

In order to investigate whether participants are able to appropriately perform in the RobS, their perception of fairness is relevant. In this context, fairness is to be understood as the extent to which participants believe the assessment format allows them to demonstrate what they have learned, i.e., their actual (believed) competence.

RQ6: To what extent do participants perceive the RobS as fair?

Claim 8: Use

If the RobS is to be implemented in teacher education, it is relevant to explore the perspective of the participants, as it is they who are potentially going to be the subject of assessment. This is of particular importance if the participants' learning behavior is to be modified by the inclusion of RobS as assessment, shifting from a focus on knowledge to a focus on performance.

RQ7: To what extent do participants believe that the RobS can be used for summative assessment in teacher education?

15.2.1 Methods

To explore the research questions, guideline interviews are conducted after the simulation. This way, participants can use their experience to discuss the specific aspects of the RobS. The questions for the guideline interview are tested in the pilot survey and can be found in full in the original German version in Appendix D.4. The relevant and translated questions are:

- To what extent was this an authentic situation (a situation found in school)? How did you experience the situation?
- To what extent did the actor act authentically?
 - a. Did they stick to the role continuously?
 - b. Did they seem like a natural conversation partner?
- To what extent did the role-play, in your opinion, test competences that are relevant for the profession of a teacher?
 - a. Which competences specifically for English teachers?
- To what extent can you imagine such role-plays as an assessment format in teacher education?
 - a. And for English teacher education specifically?
 - b. Is it a fair assessment in that you could demonstrate your competences?

It must be noted that the term *competence* is used in the interview in a broader sense as this allows participants to describe all the different dispositions, skills or performative aspects that they consider to be assessed by the RobS. Not all questions that are part of the guideline are analyzed for this thesis, as other questions, e.g., those that pertain to the comparison to other assessment formats, are not directly relevant for the validity argument as discussed here. The interviews are recorded via Microsoft Sound recorder and transcribed using the guidelines described in Appendix E.1.

The data is analyzed via MAXQDA 2020, using the structuring approach to content analysis as outlined by Kuckartz and Rädiker (2022, p. 129, 2023, p. 99). The comprehensive coding manual for the guideline interviews can be found in Appendix F.2. The unit for analysis is the whole interview transcript. The questions of the guideline serve as the main categories for the first coding process.

In the second coding process, different approaches are taken for the respective research questions. For RQ4, RQ6, and RQ7, deductive categories are used for the first level of subcategories which allows for a dichotomous distinction of the perception as rather positive or negative, with additional options for unclear answers or partial allocations (see Table 20 as an example). Within these first level subcategories, further inductive categories are developed that outline emphases or limitations that contribute to the perception of authenticity and fairness or the perspective on implementation as summative assessment.

Table 20: Example category system for the authenticity of actors' behavior¹²¹.

Main Category	First Level Sub-category	Second Level Subcategory	Third Level Sub-category	Fourth Level Subcategory
Actor_Behavior	Rather Authentic	With Limitations	Communicative Behavior	
		Without Limitations	Emphases	Communicative Behavior
	Rather Inauthentic	Communicative Behavior		
	Unclear			

This approach allows for each participant to be clearly assigned to a first level subcategory so that the RQs can be answered in a quantitative-evaluative approach. Moreover, the relative frequencies of perceptions within the sample can be described, while still allowing for the nuances in the statements to be analyzed, as limitations are still coded as such, even if the perception is rather positive. For RQ5, categories are developed strictly inductively that allow for the structuring of the different aspects that participants believe are being tested in the RobS. In a last coding process, the developed categories are applied to the whole data set.

The results will be presented primarily for the first level of subcategories, as this level is deemed sufficient to answer the research questions for this context. Specific findings of the second level are added for deeper context. Furthermore, a rather quantitative perspective is taken, with selected qualitative excerpts of the interviews provided to contextualize the perspective of the participants. The merit of the coding manual is investigated in an intercoding procedure. One additional person received an introduction into the coding manual and coded a subset of about 20 % ($n = 11$) of the interviews, which are chosen at random.

¹²¹ Potentially, a coded segment that describes a rather inauthentic perception could still include specific aspects that are perceived as authentic by the participant, as indicated by the subcategories for a rather authentic perception (see also Appendix F.2). For this specific example, however, no such instances were coded and, thus, not included here.

The percentage agreement of codes found in documents is proposed as an agreement indicator for qualitative content analysis (Kuckartz & Rädiker, 2023, pp. 202–205). For the aggregated subcategories in the individual main categories, there is an agreement range from 72 % to 100 % (see Table 21). Code-specific agreement is detailed in Table 81 in Appendix H.1. The agreement is good for all categories (> 80 %), except *Tested Competences_General Competences*, which is moderate (Döring, 2023, p. 559). Aggregating all sub-codes there is an agreement of 90.15 %.

Table 21: *Percentage agreement for the subcategories within the main categories of the guideline interview.*

Main Category	Percentage Agreement
Actor_Behavior	100 %
Actor_Natural Conversation Partner	100 %
Actor_Continuity	100 %
Situation_Authenticity	83.33 %
Situation_From School	100 %
Tested Competences_General Competences	72 %
Tested Competences_EFL Competences	85 %
Fairness	88.89 %
Summative Assessment_General Teacher Education	100 %
Summative Assessment_EFL Teacher Education	100 %

15.2.2 Results

Results will be presented with regard to the specific research questions. For each aspect, an overview of the code frequencies for each first level subcategory is presented and contextualized with findings from the second level of subcategories and selected statements. This is considered to provide a sufficient indication of the participants' perspectives as a whole in order to discuss the specific research questions and the respective claims in the IUA.

15.2.2.1 Authenticity of the RobS

Concerning authenticity, the perception of the actor and the situation is explored respectively. Generally, both are perceived positively by a majority of the participants (see Table 22).

Table 22: Code frequencies of participants' perception of authenticity of the RobS.

(Sub-)Category		Rather Authentic / Positive		Rather Inauthentic / Negative		Unclear		Missing		Total
		<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>
Actor	Behavior	47	92.16	2	3.92	2	3.92	-	-	51
	Natural Conversation Partner	40	78.43	0	-	1	1.96	10	19.60	51
	Continuity	48	94.11	1	1.96	2	3.92	-	-	51
Situation	Perception of the Situation	46	90.19	2	3.92	3	5.88	-	-	51
	Situation Found in School	47	92.16	1	1.96	-	-	3	5.88	51

The actor's behavior is perceived to be authentic by 92.16 % of the participants. For instance, $n = 44$ participants highlight the actors' verbal communicative behavior as a contributing factor to authenticity:

TP: Hm, ich würde sagen sehr authentisch, weil sie hat auch oft gefragt, was für Fehler sie gemacht hat und wie sie sich verbessern kann. (GJS22)

(TP: Hm, I would say very authentic, because she often asked about the mistakes she made and how she can improve.)

Nonetheless, participants also addressed limitations to authenticity, for instance, the actors' age is addressed by $n = 4$ participants:

TP: Hm schon sehr bis auf natürlich, dass die Dame gerade nicht in dem Alter war, mit dem man normalerweise dann so ein Feedback ähm - gespräch führen würde, aber ähm ich fand es schon relativ authentisch, weil sie halt auch oft so getan hat

I: Hm und ähm hat Sie das irgendwie gestört, dass das Alter nicht stimmte?

TP: Nein. (CCF27)

(TP: Hm actually a lot, except for the fact that, of course, the woman was not of that age, with which you would normally have such a feedback conversation, but I did find it relatively authentic, because she often acted like it.

I: Were you somehow bothered because the age was not right?

TP: No.)

Two participants are rather of a contrary opinion, stating that the actor's behavior is rather inauthentic to them:

TP: [...] dieses dumm stellen das kam mir zu stark rüber, dass sie wirklich stur fast alles, was ich gesagt habe, nochmal hinterfragt hat und das

kam mehr so rüber, dass es war, um mich zu ärgern, sag ich mal, ärgern in Anführungsstrichen, als dass sie wirklich interessiert daran war mehr zu lernen. (BJL19)

(TP: [...] this playing dumb was too strong for me, that she really questioned almost everything I said and that was more like, let's say, to annoy me, annoy in quotation marks, than that she was really interested in learning more.)

The answers of two participants are coded as unclear, as no tendency is identified in their statements. With regard to continuity, 94.11 % of the participants state that the actor stayed in the role continuously, for instance:

TP: Ja, also ich würde jetzt eigentlich schon sagen, ich hatte jetzt keinen Moment, wo ich mir dachte, ne ist jetzt was ganz anderes. (SSM23)

(TP: Yes, I would say so, I did not have a moment where I thought, no, now this is totally different.)

Only one participant comments that they perceived a change in the actor's behavior:

TP: Hm man hatte das Gefühl, dass sie am Ende ein bisschen erwachsener, also so mehr die Rückfragen kamen, umso mehr würde ich sagen ist sie so ein bisschen aus der Rolle gefallen. (MJO28)

(TP: Hm one had the feeling that she was more adult at the end, the more the questions came, the more I would say she fell out of her role.)

When asked about the SL acting as a natural conversation partner, 78.43 % ($n = 40$) of the participants affirm this. For instance, one participant confirms the actor's natural behavior and states that their calmness and reserved manner contributed to this perception:

TP: Hm ne ich finde schon, dass es natürlich gewirkt hat. Sie war auch so ein bisschen ruhiger, zurückhaltend. Und ich glaube so sind halt auch Schüler häufig in der Situation, wenn sie dann eben mit dem Lehrer in so einem Eins-zu-Eins-Feedbackgespräch äh eben sind. (HKV21)

(TP: Hm, no I think it seemed natural. She was a bit calmer, more reserved. And I believe learners are often like that in this situation, when they are in a one-to-one feedback conversation.)

Ten participants did not receive this question and one participant's answer is unclear.

Concerning the authenticity of the situation, the overall picture is similar as the majority of participants perceive the situation as authentic and consider it to be a situation that frequently occurs at schools (see Table 22). With regard to the former, 90.19 % of the participants state that they perceive the situation as rather authentic, even though limitations to authenticity are also mentioned by $n = 17$ participants. For instance, the volume of material in the situation is addressed:

TP: Ja schon, schon. Also mit den ganzen Zetteln, die vor einem liegen, das war schon ein bisschen unauthentisch. Ich habe immer so gedacht okay hier, hier muss ich gucken, hier muss ich gucken und man vergisst schnell mal was. (SJH28)

(TP: Yes, quite, quite. However, with all the sheets that were lying in front of one, that was a bit inauthentic. I always thought, okay, here, here I have to look and you and you quickly forget something.)

Two participants, however, state that they did not perceive the situation as authentic, citing their personal lack of experience:

TP: Ähm, ich würde sagen von, von dem Aspekt her, dass es überhaupt nicht authentisch rüberkam, aber das kann auch, das kann auch hauptsächlich daran liegen, dass das halt eine ungewohnte Situation für mich war, weil ich so eine solche Simulation noch nicht gemacht habe und ich da auch kein Vergleichswert für habe. (BJL19)

(TP: I would say that, to that aspect that it was not authentic at all but this might be mostly due to the fact that it was an unfamiliar situation for me, because I have never done such a simulation and I do not have a comparison.)

When asked if they believe that the situation as experienced might happen in schools, 92.16 % of the participants affirm this to some extent, often referring to their own experiences from school, either as a learner or teacher:

TP: Ähm also wenn ich jetzt an meine Schulzeit denke, dann war das sehr authentisch. Sowas kam öfter mal vor, dass man so einen Text abgegeben hat. Äh und die Lehrerin dann ähm ein Feedback gegeben hat. (BAH29)

(TP: Ehm thinking back to my time at school, then this is really authentic. That happened quite often, that one handed in a text. And then the teacher gave feedback.)

Contrarily, one participant states that they did not experience this situation in school:

TP: Ich habe das nicht oft erlebt. Also ähm ich habe das wirklich nicht oft erlebt leider. Ähm eigentlich kam das immer nur beim Elternsprechtag und dann auch mit den Eltern, ne. (SAD06)

(TP: I did not experience this often. Ehm I really have not experienced this often, unfortunately. Actually, this only happened at the parent-teacher conference and then always with the parents.)

15.2.2.2 Competences Tested by the RobS

Participants mention a variety of competences they believe are being tested by the RobS. An overview of the first level of subcategories and the number of participants that mention them is provided in Table 23. Illustrative quotations can be found in the coding manual in Appendix F.2. The participants' answers can be clustered into seven different subcategories, including *Other*, in which singularly addressed aspects are subsumed.

Table 23: First level of subcategories of competences (general) required in the RobS mentioned by the participants.

Competence (General)	<i>n</i> _{participants}
Feedback	39
Assessment	17
Teacher-student-interaction	38
Self-reflection	3
Situation-related	14
Professional Knowledge	7
Other	6

An overview of the second level of subcategories is provided in Table 24. With regard to *Feedback*, which is addressed by most participants ($n = 39$) six subcategories are identified. *Feedback competence (General)* includes general statements that do not specify how feedback is assessed. *Dealing with Errors / Mistakes* covers statements that are concerned with the classification and appropriate handling of errors and mistakes in the text. The subcategories *Feed Up*, *Feed Back*, and *Feed Forward* relate to the levels of feedback according to Hattie and Timperley (2007), i.e., the participant's ability to formulate learning objectives, to provide information about the current state, and propose next steps for the learner. Regarding *Positive Feedback*, this relates to the ability to provide information on achievements to the learner. With regard to *Assessment*, which is addressed by $n = 17$ participants, three subcategories are identified. *Evaluation* addresses the competence to accurately judge the performance. *Diagnosis* refers to the competence of identifying errors and mistakes and *Correction* subsumes the competence to accurately correct errors and mistakes, i.e., knowing the correct version of a diagnosed error or mistake. *Teacher-Student Interaction* is mentioned by almost as many participants as *Feedback* ($n = 38$); four subcategories are developed for that category. The category of *Dealing with Learners* addresses the competence to communicate and engage with learners. The category of *Emotional-motivational Competences* subsumes several different subcategories like *Social competence*, *Empathy*, *Coping with emotions of learners*, *Motivating learners* and *Building relationships with learners*. The category *Conversational Competences* includes categories such as *Communication* or *Structuring*. Similarly, the category of *Subject-specific Interaction* includes several categories like *Learner orientation*, *Prioritizing aspects*, *Ability to explain* and *Appropriate use of language as a medium of instruction*. *Reflection*, which is mentioned by $n = 3$ participants, addresses the competence to use the feedback situation to evaluate prior instruction and use the information to adapt for future instruction. The category *Situation-related competences*, mentioned by $n = 14$

participants, relates to the participants' *Organization / Time Management*, their *Spontaneity / Flexibility* in the situation, and their *Adaptation to the situation*. Concerning *Professional Knowledge*, the participants, $n = 7$ in total, mention *Professional Knowledge in General*, knowledge about language, i.e., *Linguistics*, and the ability to speak English, i.e., *Mastering of the English language*. Individually named aspects, which are subsumed as *Other*, include *Action-oriented competence*, *Solution-orientation*, *Efficiency*, *Provision of resources*, *Methodic competence*, *Honesty*, *Counseling competence*, and *Separation between person and product*.

Table 24: Second level of subcategories of competences (general) required in the RobS mentioned by the participants.

First Level Subcategory	Second Level Subcategory	$n_{\text{participants}}$
Feedback	Feedback competence (General)	13
	Dealing with Errors / Mistakes	13
	Feed Up	1
	Feed Back	15
	Feed Forward	17
	Positive Feedback	5
Assessment	Evaluation	5
	Correction	6
	Diagnosis	11
Teacher-Student Interaction	Dealing with Learners	5
	Emotional-motivational Competences	21
	Conversational Competences	9
	Subject-specific Interaction	22
Reflection		3
Situation-related	Organization / Time management	9
	Spontaneity / Flexibility	6
	Adaptation to the Situation	2
Professional Knowledge	Professional Knowledge in General	2
	Linguistics	3
	Mastering of the English Language	2
Other		6

When asked for competences that are specifically tested in the EFL context, the categories of *Feedback*, *Assessment*, and *Teacher-Student Interaction* are also addressed by the participants with EFL-specific examples. As they are also structured using identical subcategories, these aspects will not be repeated here. These categories are also coded to a lesser extent as indicated in Table 25 for first level subcategories and in Table 26 for the second level subcategories.

Additional, distinct subject-specific competences are mentioned by $n = 29$ participants (see Table 25). The subcategory of subject-specific competences, includes two subcategories at the second level. The category *Language* is mentioned by $n = 23$ participants and includes *General language competence*, *Mastering of the English language*, *Explicit knowledge of language*, and *Translation skills*. Concerning *Knowledge*, which is addressed by $n = 14$ participants, aspects of *Knowledge of typical errors / mistakes*, *Specific issues of German L1 learners of English*, *Genre knowledge*, and *Cultural context* are mentioned. Individually mentioned facets are subsumed under *Other*. This includes *Pedagogical implementation*, i.e., teaching the content, *Fostering oral competence*, *General knowledge*, and *Providing safety*.

Table 25: First level of subcategories of EFL-specific competences required in the RobS mentioned by the participants.

Competence (EFL-specific)	$n_{\text{participants}}$
Feedback	15
Assessment	17
Teacher-student Interaction	22
Subject-specific	29
Other	4

Table 26: Second level of subcategories of EFL-specific competences required in the RobS mentioned by the participants.

First Level Subcategory	Second Level Subcategory	$n_{\text{participants}}$
Feedback	Feedback competence (General)	1
	Dealing with errors/mistakes	3
	Feed Forward	11
	Positive Feedback	1
Assessment	Evaluation	2
	Correction	4
	Diagnosis	13
Teacher-Student Interaction	Emotional-motivational competences	6
	Conversational competences	1
	Subject-specific interaction	20
Subject-specific	Language	23
	Knowledge	14
Other		4

15.2.2.3 Fairness of the RobS as Assessment

Regarding the fairness of the test, $n = 28$ participants or 54.90 %, consider the test to allow them to demonstrate their actual competence to a certain extent (see Table 27).

Table 27: Overview of participants' perception of fairness in the RobS.

Category	Rather Positive		Rather Negative		Partially		Unclear		Missing		Total
Fairness	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>
	28	54.90	4	7.84	2	3.92	6	11.76	11	21.56	51

Two participants state that the RobS could be assessed fairly and add that it might potentially be fairer than a written exam, for instance:

TP: [...] aber ich denke, dass das doch fair bewertet werden kann. [...] Vielleicht ist es sogar am Ende des Tages fairer als wenn man eine Klausur schreibt. Da eine Note meistens nicht das widerspiegelt, was eine Person kann oder halt nicht kann. (NDA16)

(TP: [...] but I think that it can be assessed fairly. At the end of the day it is maybe fairer than writing an exam. Because a grade often does not reflect what a person is able or not able to do.)

While still considering the procedure to be rather fair, some participants also address limitations, like the general high-pressure situation that is inherent in any (high-stakes) summative assessment situation or the requirement for adequate preparation in the preceding course that is aligned with the assessment format:

TP: Ja also ich meine/ ich meine es ist und bleibt dann eine Prüfungssituation, das heißt ähm man sitzt dann jetzt als Studierender mit gewissem Druck vielleicht oder ein bisschen Nervosität dann auch da und ich denke das ist dann jetzt auch nicht dasselbe wie wenn man jetzt in der Schule also wirklich in der Schule sitzt und dann eine Schülerin nach der Stunde mal kommt ähm also das ist dann natürlich so ein bisschen schwierig aber ich denke im großen und ganzen, wenn man dann auch so darauf vorbereitet wird in den Seminaren ähm dann äh ja doch dann denke ich schon dass man dann auch das zeigen kann. (FCP15)

(TP: Yes, so I think/ I think it still remains an assessment situation, that means ehm you sit there as a student with a certain amount of pressure maybe or a bit of nervousness and I think this is not the same as really sitting there at school and a learner comes so this is of course a bit difficult but I think overall, if you are prepared in the sources ehm yeah I think then you can show what you are able to do.)

A smaller share of the participants, $n = 4$, are rather skeptical of the fairness of the RobS overall, arguing, for instance, that it is too dependent on the individual's personality:

TP: Ähm ich sage mal so fair wird halt schwierig, weil halt jeder eine eigene Persönlichkeit hat für sich, wie man mit so einer Situation umgeht. (NRS18)

(TP: I would say that fair is difficult, because everybody has their own personality, how to deal with such a situation.)

Some participants, $n = 2$, provide a more balanced opinion and cannot be decisively put in either category as they partially (dis-)agree:

TP: Teilweise. Also man ist/ man hat natürlich schon das Gefühl man wird beobachtet und versucht sich dann natürlich zu präsentieren auf eine bestimmte Art und Weise und die Aufregung spielt da dann vielleicht ein bisschen mit rein, dass man nicht alles zeigen kann oder so sagt wie man es gerne sagen möchte. (BNP09)

(TP: Partially. So, you are/ you of course have the feeling of being observed and you try to present as natural in a specific manner and the nervousness plays into this so that you cannot show everything or say how you would like to say it.)

While $n = 10$ participants did not receive this explicit question and thus did not provide an answer, some participants, $n = 6$, provide unclear answers which do not allow for an indication of their perception and tend to simply describe the implications for RobS as a format of assessment:

TP: Ähm man ist ja quasi äh in einer Simulation und ähm vor allem, wenn man einen Gesprächspartner hat, der vielleicht auch etwas sagt, was unerwartet ist, ähm man muss halt instinktiv handeln und äh selber entscheiden in dem Moment, wie man sich verhält. Und da kann man nicht viel ähm einfach auswendig lernen wie man das für die Klausuren macht und dann ne ähm äh dieses Bulimie-lernen quasi funktioniert halt nicht. Und deswegen man arbeitet da eigentlich schon äh/ Das ist was Soziales quasi. (SAD06)

(TP: Ehm you are quasi in a simulation and most importantly, if you have a conversational partner who maybe also says something that is unexpected you have to act instinctively and eh decide in the moment how to act. And you cannot memorize a lot like you do with exams and this bulimic learning quasi does not work. And that is why you actually work eh / It is something social, quasi.)

15.2.2.4 Use of the RobS for Summative Assessment

Concerning the use of the RobS for summative assessment, participants have a predominantly positive view (see Table 28).

Table 28: Overview of participants' perspective of the implementation of RobS for summative assessment.

Subcategory	Rather Positive		Rather Negative		Missing		Total
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	
Teacher Education General	48	94.11	3	5.88	-	-	51
EFL Teacher Education	44	86.27	1	1.96	6	11.76	51

When asked whether they could generally imagine the use of RobS as a format for summative assessment in teacher education, $n = 48$ participants, i.e., 94.11 %, have a rather positive view, often pointing out how the RobS addresses a gap in teacher education, and that the RobS might prepare them better for practice:

TP: Ich kann mir das sehr gut vorstellen und zwar deshalb, weil ich das Gefühl habe, dass sowas im Lehramtsstudium völlig untergeht, man es aber am Ende trotzdem irgendwie können muss. Man muss ähm mit Schülern reden können. Man muss auch mit Eltern reden können. Und keiner bereitet einen so wirklich darauf vor. Ähm deshalb finde ich, ist es sehr sinnvoll, dass mal als Prüfungsleistung anzubringen, weil sonst sind ja mündliche Prüfungen oft einfach, ne, das ist dann so ein Frage-Antwort-Spiel und ähm gefühlt wird halt einfach irgendwelches Wissen abgefragt, was man sich vorher über die Wochen irgendwie exzessiv angeeignet hat und danach ist dann irgendwie alles wieder weg. Und so ein Feedback- oder Beratungsgespräch ähm ist ja auf jeden Fall ja halt irgendwas, was man halt später auch braucht. Also finde ich das schon sinnvoll. (SNA28)

(TP: I can imagine that very well because I have the feeling that something like this completely gets lost in teacher education, but at the end of it you somehow have to be able to do it. You have to talk to learners. You have to talk to parents. And no one really prepares you for it. And that is why I think it does make sense, to include this as summative assessment because, like oral exams are often just a game of questions and answers and it feels like simply any kind of knowledge is quizzed, which you acquired over weeks somehow excessively and after wards it is all somehow gone. And such a feedback or a counseling conversation is surely something that you need later on. So, I think it is quite sensible.)

However, some participants, while having a positive attitude, name some limitations, such as the condition that there needs to be an alignment between the objectives, which are usually specified in the university-specific exam regulations, learning activities, and the assessment format:

TP: Also grundsätzlich kann ich es mir schon vorstellen und äh irgendwo würde ich es mir auch wünschen natürlich das wir dann so authentische Situationen dann auch haben, aber man muss im Studium natürlich auch dementsprechend darauf vorbereitet werden und ähm insofern finde ich das ein bisschen schwierig also jetzt ich sage mal bei der jetzigen Prüfungsordnung. (FCP15)

(TP: So generally, I can imagine that and somehow, I would of course also wish for such an authentic situation, but you have to be prepared accordingly and therefore I think it is a bit difficult with the current exam regulations.)

Three participants are rather skeptical towards an implementation, for instance, with regard to standardization and the complexity of the situation:

TP: Hm (...) jetzt gerade kann ich es mir noch nicht so gut vorstellen, ähm einfach weil ich das so als Prüfungsformat nicht kenne [...] Die Bewertungskriterien müssen ja wirklich sehr stark ausgefächert sein, dass das auch wirklich ja fair bewertet werden kann und dass es bei jedem die Situation gleich ist, dass die ähm Ansprechpartner, die Simulationspartner dann auch bei jeder Simulation gleich reagieren und dass es halt identisch bleibt. (SLG21)

(TP: Hm (...) right now, I cannot really imagine it, simply because I do not know this assessment format [...] The assessment criteria need to be really wide-ranging so that it can be assessed fairly and that everyone has the same situation, that the conversational partners, the simulation partners also react equally and that it remains identical.)

The distribution is similar when the question is specified for EFL teacher education, as $n = 44$ participants have a rather positive view. Participants frequently point to the language aspect, specifically, the appropriate choice and adaptation of language that can be addressed in such an approach to assessment:

TP: Ja, Englisch hat man ja schon noch mal diesen Punkt ähm redet man überhaupt auf Englisch oder auf Deutsch ähm gerade bei Schüler/ [...] gerade bei Schülern, die sehr schlecht sind, hat man natürlich die Gefahr, wenn man auf Englisch spricht, dass dadurch schon zu Komplikationen kommt, dass sie das Feedback gar nicht verstehen sprachlich, also macht es natürlich schon nochmal mehr Sinn das in einer Fremdsprache zu machen, weil, wenn ich da jetzt als Studierender mit meinem ganzen Englischkenntnissen Feedback gebe und nur Fachwörter verwende, ist ja komplett an der Aufgabe verfehlt, sage ich jetzt mal. Und diese Schwierigkeit gäbe es in Philosophie jetzt zum Beispiel nicht. (MAP27)

(TP: Yes, in English you have this aspect of, do you talk in English or German, especially with learners [...] especially with learners who are very bad, there is the danger, when talking in English, that this might lead to complications, that they do not understand the feedback because of the language, so it of course makes more sense to do it in a foreign language, because, when I as a student give feedback with my English abilities and only use technical term, the point of the task is missed, I would say. And you don't have this difficulty in Philosophy, for example.)

One participant, who was also skeptical with regard to the use of RobS in general, is also skeptical of its use in EFL teacher education, citing the difficulty of adequately choosing the level of language when the learner is not previously known:

TP: Ne, das ist eigentlich ähnlich dann. Jetzt so im Nachhinein betrachtet, ich habe mein Feedback auf Deutsch gemacht. Das würde ich dann auch in der Praxis dann an den Schüler anpassen äh je nachdem wie der wie das Level der Sprache bei dem ist. [...]

I: Und das konnte man hier in der Simulation nicht machen?

TP: Hm doch (lachen). [...] Also aber wenn man das im Unterricht macht oder in der Praxis macht, dann hat man ja schon eine größere Bindung zu den Schüler innen, dass man dann auch weiß, ob die dann auch auf Englisch das verstehen, was sie dann, was man grad vermitteln möchte oder ob die dann vielleicht doch eher ein bisschen Hilfe bräuchten mit dem Deutschen dann. (SLG21)

(TP: No, that is similar, actually. In hindsight, I gave my feedback in German. In practice, I would adapt that to the learner, depending on their level of language. [...])

I: And this is not something you could do in the simulation?

TP: Well, yes (laughs). [...] But if you do it in the classroom or in practice, then you have a greater bond to the learners, so that you know, if they understand English, what they then might want to convey or if they maybe need a little bit more assistance with German.)

15.2.3 Discussion and Limitations

The results will be discussed individually for each RQ, but not directly related to the validity of the RobS, as that discussion will be done separately in Chapter 17.

Regarding RQ4, the extent of the authenticity of the RobS, it can be stated that, overall, the RobS is perceived to be quite authentic by most participants. This relates to all aspects concerning the actor, as well as the general situation. While the age of the actor is addressed as a limitation, it does not seem to be a decisive factor that prohibits an overall rather authentic perception. The communicative behavior is emphasized as a contributing factor to authenticity, also providing a further indication of the appropriateness of the VTs that constitute a substantial part of SL's verbal communication. With regard to the situation, some participants indicate doubts in their ability to appropriately evaluate authenticity due to their lack of experience, thus the situation itself is not necessarily the reason for an inauthentic perception. Considering the aim of approximating practice, particularly in the perception of the target group, the RobS fulfills the goal. Therefore, four dimensions outlined in the model of authenticity by Gulikers et al. can be regarded as fulfilled from the perspective of the participants: the task, the physical and social context, and the form of assessment (2004, p. 74; see also Chapter 4.4.1). As authenticity is essential for learners to adjust their learning behavior, it can be assumed the RobS might be sufficiently authentic to influence learners accordingly.

With regard to RQ5, the competences that the participants address for both general as well as EFL-specific teacher education can be compared to the framework as outlined in Chapter 7.3. First, aspects of *Feedback* and *Assessment* are addressed in a combined manner. Several competences mentioned by the participants align with the framework, for instance, the general notion of feedback competence, the distinct levels of the feedback model of Hattie and Timperley

(2007), and the provision of positive feedback, are directly reflected in the framework. The issue of errors and mistakes is also addressed in the framework and by the participants, e.g., evaluating the text and diagnosing errors is one possible approach of addressing the different dimension of the text, as is the identification of error patterns. Naturally, the ability to provide the correct(ed) information is addressed by dimensions of accuracy of the information.

Concerning *Teacher-Student-Interaction*, participants also address several aspects of the framework. For instance, they name conversational competence as one facet, which can be compared to the notion of *Dialogue orientation* in the framework. One aspect that participants frequently mention is that of emotions and motivation. While the latter aspect is related to positive feedback, and having a dialogic approach to feedback also contributes to building a relationship, the general notion of these two aspects is consciously excluded from the framework to allow for a focus on the cognitive aspects of feedback, as discussed in Chapter 7.4. As the affect-motivational facet is, however, relatively present for the PSETs, the exclusion of this facet for summative assessment should be made transparent to avoid false expectations. The subject-specific dimension of interaction includes, for instance, *Learner orientation*, which is included in several aspects of the framework, e.g., *Adaptation of Content*, and the appropriate use of language, which is covered in *Adaptation of Language*.

Several subject-specific competences are also addressed by the participants. A prominently mentioned aspect is that of knowledge, in which mostly language-related aspects are addressed, i.e., knowledge about the English language and being able to use the English language. This aspect is accounted for from a performance perspective, primarily in the areas of *Language Accuracy* and *Professional Accuracy*, which address the correctness of the information that participants provide, but also in the aspects of *Quality of Feed Back* and *Quality of Feed Forward*, as the provided information is supposed to be correct.

Situation-related competences that the participants describe are not explicitly found in the framework. This relates specifically to the notion of *Organization / Time Management*, which is a rather overarching skill that is not specifically relevant for providing feedback and thus not included here. Arguably, this competence might still be assessed indirectly as a lack of organization and time management might affect the behavior, e.g., not all dimensions of the text are addressed if time runs out. With regard to *Spontaneity / Flexibility* and *Adaptation to the situation*, it is unclear to what extent the participants consider the actor to be part of the situation, as the adaption to the SL is specifically addressed, as previously discussed. Otherwise, these facets are similarly of an overarching nature and not specific to feedback and thus not included in the framework.

The aspect of *Reflection* is particularly interesting, as the RobS is specifically designed to focus on the performance in the situation, while reflection is generally a process that succeeds action. Possibly, participants relate the RobS to previous assessment formats they have experienced, which often have a reflective component when action or practice is included, e.g., in portfolios for different internships. The fact that several competences are also mentioned by participants individually, emphasizes the complexity and variety of requirements that participants perceive to be relevant for the RobS. While generally, there is a notable overlap between what the participants consider relevant in the RobS and what is included in the framework, it is essential to transparently describe the requirements so that the participants can appropriately focus on the required criteria.

Concerning RQ6, the fairness of the RobS, a majority of the participants believe that the RobS can be used fairly as an assessment format. However, this portion is smaller compared to the other aspects mentioned in this chapter. Interpretation of this is also complicated as a substantial share of statements are not clearly identifiable and a number of participants did not answer this question. Tendentially, it seems that participants are generally skeptical of the fairness of summative assessment, as it remains a high-pressure situation, which leads to nervousness and thus affects the ability to demonstrate what one believes they are able to do. Moreover, participants argue that individual preferences of assessment formats might also influence the performance in the RobS. Consequently, the issue of fairness is not necessarily related to the RobS, but to summative assessment in general. This assumption is supported by individual participants who state that the RobS might be fairer than traditional assessment such as written exams. It must be noted, however, that the question is hypothetical in nature – the RobS is not actually implemented as a format of assessment – and the perception of fairness is likely to be influenced by the concrete implementation. Thus, the general perception of fairness should be explored if, or when, the RobS is actually implemented as an assessment format to gain a more substantiated impression of PSETs' perception and compare it to other assessment formats.

Regarding RQ7, the potential implementation of RobS as an assessment format, the majority of the participants views this positively, for general, as well as EFL-specific, teacher education. Participants often emphasize the perceived lack of practice, i.e., action-orientation, in teacher education and see RobS as a potential tool to address this gap. However, the participants also point out that assessment requires adequate preparation in the respective courses, especially for such a new format. Thus, participants are aware that constructive alignment, as previously outlined, is essential and that assessment is not a means, in and of

itself. Skepticism towards the RobS relates, for instance, to the standardization of the situation. Possibly, the concept of the RobS could be made transparent to PSETs so that they are familiar with the extent of the standardization beforehand.

Some limitations should be addressed in the context of this study. First, there are some questions that have a substantial number of missing answers. This occurred primarily in more dynamic situations within the interviews, as other points that required clarification were discussed in more depth and the attempt to adapt to the flow of the interview led to missing these questions. Thus, the quality of the data is partially limited, especially for items with a large share of missing responses, like fairness. Second, the aspect of social desirability is also of concern. As the interviews are conducted in a social context, participants might be inclined to answer in a way that they consider favorable for the interviewer, who is also responsible for test development (Aguado, 2018, p. 19). Alternatively, these questions could have been addressed as part of the written survey, allowing for a greater extent of anonymity. However, the intention of the interviews is to understand the participants' perspective in more depth, which is why interviews are favored. Third, while the participants' perspective is extremely valuable, as they are the target group, the perspectives of other stakeholders are also relevant. For authenticity, for example, the perspective of actual learners and teachers might provide additional insights from those stakeholders that are portrayed in the RobS. For the use of the RobS, the perspective of university lecturers, who would likely be responsible for implementation, is essential as they also need to accept this new format of assessment. Lastly, the results of the qualitative content analysis are presented largely from a quantitative perspective. While this condensed approach is considered sufficient to discuss the research questions, this also reflects a loss information. Potentially, future research can address this gap and provide a more detailed picture of PSETs' perception of the RobS.

15.2.4 Summary

In this study, the perspective of the participants on authenticity, tested competences, fairness and use of the RobS as an assessment format is focused on. In guideline interviews, the participants are asked about their perception or opinion with regard to these aspects. Using the structuring approach to qualitative content analysis, as described by Kuckartz and Rädiker (2022, p. 129, 2023, p. 99), the interviews are analyzed. The results are presented from a primarily quantitative perspective with regard to the research questions. Concerning authenticity, a large majority of the participants perceive the actor and the situation to be rather authentic. Participants name several different competences they believe are being tested in the RobS, most of which are reflected in the framework to some extent. It seems, however, important to be transparent with regard to

the assessment criteria, as not to cause false expectations in the PSETs. A majority of participants also consider the RobS to be a fair potential form of assessment, considering the general challenges that any form of summative assessment has. When asked about a potential implementation, most of the participants have a favorable opinion, stating that the RobS addresses a gap that exists in current teacher education. Limitations relate to the method of interviews, e.g., social desirability, as well as the limited scope of perspectives, as only the perspective of PSETs is explored.

15.3 Study 2: Expert Interviews

This chapter focuses on the perspectives of experts with regard to relevant criteria and the operationalization of items in the scoring instrument. In validation studies, experts frequently serve as an external source that provides evaluative information on the current construct and operationalization of scoring (M. T. Kane, 2006, p. 28, 2013a, p. 449). The information that the experts provide can also be used to modify and improve the scoring instrument. Additionally, experts' assessment of the quality of a performance ideally correlates positively with the final test score generated by the scoring instrument, presupposing that experts are in agreement regarding the quality of the performances. The expert interviews are conducted as a two-tiered study. In the first part, experts are asked to assess selected performances on a global scale, so that their assessment can be used as a comparison for the test score that the (final) scoring instrument generates. Also, they are asked to provide evaluative comments to the reasoning of their assessment, allowing insights into the criteria the experts use. In the second part, experts are asked to evaluate the scoring instrument in qualitative interviews. The aim is to collect data to discuss the scoring instruments' quality and to use the evaluation to improve the scoring instrument by modifying item operationalizations, adding missing criteria, and excluding redundant or unnecessary criteria. The findings of the study can be used to discuss the validity of the following claims and research questions:

Claim 1: Model

The criteria that are used by the experts when evaluating the performances can be compared to the theoretically derived framework which is the foundation for the RobS construction. As the scoring instrument is also based on the framework, experts' evaluative comments on the instrument can be extended to the framework as well.

RQ8: To what extent do experts' assessment criteria align with the framework and the scoring instrument?

Claim 4: Item score

The appropriateness of the individual item operationalizations is usually investigated by expert judgements (M. T. Kane, 2006, p. 28, 2013a, p. 449). Based on their evaluation, items can be included, unchanged or modified, according to their suggestions. Also, redundant or unnecessary criteria can be excluded and missing aspects included.

RQ9: To what extent do experts consider the items of the scoring instrument to be appropriate for the purpose of assessing feedback on writing in the EFL context?

In this chapter, the full procedure of the expert interview study will be presented, but with a focus on the collection and analysis of the qualitative data for Claims 1 and 4, i.e., the qualitative evaluative comments on the performances in the first part, and the evaluation of the scoring instrument in the second part of the interview. This is done because in order to use the quantitative data, i.e., the rating of the performances, the final test score is required for statistical analyses. However, the results of the qualitative analysis of the expert interviews and a detailed discussion of the individual items' reliability (see Chapter 0) are preliminary steps that need to be taken to settle on the items for the final scoring instrument. Thus, the analysis and results for will be presented in Chapter 15.5.4.4, as part of the analyses for all the evidence for Claim 7, after the scoring instrument is finalized.

15.3.1 Method: Expert Interviews

The methodological approach of expert interviews will be discussed in this chapter. Criteria for choosing experts will be outlined and the rationale of the study procedure presented.

What makes an expert?

The question of expertise is of high importance, as this definition prescribes the criteria for selecting possible experts for this study. Generally, expertise can be conceptualized as *knowledge*, *ability* and *performance*¹²² (Krauss, 2020, p. 157; Misoch, 2019, p. 119). In a performance-oriented approach to define expertise, experts are often differentiated from novices, arguing that experts perform better in a certain field than novices; when referencing the expertise of teachers, this is also referred to as the "expertise paradigm" (Krauss, 2020, p. 157; translated by TJ) or "expert-novice-paradigm" (Hennes, 2018, p. 11). It is assumed that through their experience and performance in the field, as for example

¹²² Here, performance is used in the sense of achievement.

indicated by certificates, experts have implicit knowledge, which they can apply, for example, in evaluation processes (Hennes, 2018, p. 11). In a knowledge-oriented approach to define expertise, the specialized knowledge, which experts have acquired through experience in their field, is focused on (Misoch, 2019, p. 119). Situational demands can be used to reconstruct the specific knowledge that is required to solve a professional task, allowing for the determination of expertise more specifically (Krauss, 2020, p. 158). This also implies that expertise is socially constructed, as experts are assigned to a specific profession by their peers who are able to make that judgement (Misoch, 2019, p. 120).

While these conceptualizations provide guidance, it is ultimately “the researcher who according to his [sic!] research objective decides who she or he [sic!] wants to interview as an expert; [...] this is not an arbitrary choice but is related to the recognition of an expert as an expert within his own field of action” (Meuser & Nagel, 2009, p. 18). In order to identify experts of their field, usually several indicators of knowledge and performance are applied (Bromme, 2008, p. 161). Ideally, the expertise of the individuals for this context applies to both parts of the study. For the first part, experts should have experience in the field of teacher education and assessment, as the evaluation of feedback performances is in focus. For the second part, experts should have vast knowledge in at least one of the areas that are adjacent to the RobS, i.e., feedback, L2/FL writing, feedback on writing, feedback in the EFL classroom. Teachers, while they could undoubtedly provide valuable insights, are not considered to be experts for this particular study. This decision is primarily made with regard to the evaluation of the scoring instrument, as it cannot be assumed that teachers have the knowledge to appropriately evaluate the scoring instrument which is based on empirical findings in the field. Considering the review in Chapter 6, there is also a question whether teachers provide feedback to their learners with a consistently high quality. As a consequence, only one group has been identified that fulfills the outlined criteria: Researchers in the areas of feedback, L2/FL writing, feedback on writing or feedback in the EFL classroom. Through their research experience, it can be assumed that they have gained the required knowledge to adequately evaluate the content and structure of a scoring instrument, at least with regard to their specific discipline. Furthermore, researchers in this field also participate in teacher education and are therefore also experienced in the assessment of pre-service teachers. It can be assumed that their evaluation of the performances is more likely to be based on previous research. This assumption is corroborated in findings by Hennes et al. (2022). In a study comparing the weighting of criteria used to evaluate texts by different groups, they found the following:

Scientist [sic!] tended to rate criteria identified in theory/literature (e.g. Hennes et al., 2018) as critically relevant for text quality as more important than expected based on the other groups' ratings, i.e. they focus on the reader's perspective ("Using text structure to guide/support the reader", "Considering the reader's perspective") and text coherence ("making the text work as a whole"). (Hennes et al., 2022, p. 290)

Other studies focusing on the validation of test instruments for teacher education also rely on researchers, further supporting this approach (i.e., Gronostay et al., 2023, p. 173; Schumacher, 2017, p. 177).

Expert interviews

The mode of the expert interview is special with regard to its direction, as not the experts themselves are in the foreground, but their knowledge-based statements on a specific topic, product or performance (Misoch, 2019, p. 120). While experts are being defined and selected regarding specific criteria that arise from the research project, the specific research question also determines how the interview is designed. The overarching objective in this study is to reconstruct expert knowledge in the form of the criteria they use to assess performance, in the first part based on the performances, and in the second part based on a preliminary scoring instrument. Due to the "procedural nature and the non-explicitness" (Meuser & Nagel, 2009, p. 31) of expert knowledge, it is advised against "extracting knowledge from the interview by asking questions in the sense of a questionnaire" (Meuser & Nagel, 2009, p. 31). Meuser and Nagel therefore argue that an open format, in which "[experts] will report cases of decision-making and state principles they keep to" (2009, p. 31) is a preferable approach, allowing experts to verbalize their knowledge. Therefore, experts are not asked to state criteria they consider important, but rather encouraged to use their knowledge in evaluative scenarios.

15.3.2 Study Design and Data Processing

In the following, the design of the expert interviews is laid out. The data collection and preparation process are presented first. Then, the specific approaches to the qualitative content analysis as the method of evaluation are reasoned. Lastly, the preliminary scoring instrument is addressed briefly.

Data collection

For this study, ten experts who fit the criteria and have at least a PhD-equivalent certificate were initially contacted via e-mail. The experts that were contacted stem from several regions of the globe: Europe, North-America, Asia, Australia and Oceania. Five experts agreed to participate and thus constitute the sample as described in Table 29. The interviews were conducted and recorded from

October-November of 2023 via Zoom. Experts received information on data protection and the declaration of consent before the interview and stated their agreement to both at the beginning of the recording. Before the interview started, experts were greeted, introduced to the project and its aims, and issues of data protection and consent briefly presented. In the first part of the interview, experts are shown four selected performances which they have to rate on a global level. In the second part, the scoring instrument is presented and the experts provide an evaluation thereof. The interviews were planned for about 90 minutes and had an actual duration between 90 and 160 minutes. The first part was planned to last about 60 minutes, the second part about 30 minutes. The interviews were conducted in German with experts 1,2 and 3, and in English with experts 4 and 5, based on their Native language(s).

Table 29: *Overview of experts.*

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
Language of interview	German	German	German	English	English
General area of expertise	Foreign language education	Education and Multilingualism	Teaching English as a Foreign Language	Curriculum & Pedagogy	Education
Relevant expertise for this study	EFL teacher feedback competence	Feedback on L2/FL writing (German and English)	Feedback in the English language classroom	Writing & Feedback on writing	Feedback

In the first part of the interview, experts are shown four selected videos, asked to assess them using the German or US higher education grading system (see Table 69 in Chapter 15.5.3.3), and verbalize their evaluation, providing reasons. The videos have a duration of about eight – ten minutes each. Using this approach, experts are asked to verbalize the criteria they used for assessment in a specific case, providing access to their guiding principles (Meuser & Nagel, 2009, p. 31). After watching all the videos, the experts are able to settle on a final grade for each performance and are asked to determine a final ranking of the four performances. The videos are selected to represent a range of grades and thus performances (see Table 30).

Table 30: *Preliminary grades of the videos in the expert interview.*

Position in the rating procedure	Participant code	Grade (Germany)	Grade (US)
1	BNP09	3.0	C
2	MMM07	2.0	B
3	NDA16	5.0	F
4	CML16	4.0	D

The grades are based on a preliminary point-grade conversion and a score that is generated with the preliminary scoring instrument (see Appendix D.5). 50 % of the maximum points are considered the threshold for passing the test (i.e., 4.0), the rest of the points are distributed equally to represent the range of grades. As the videos are mostly in German with English being used only to a small extent, subtitled versions were created for the non-German speaking experts. The transcripts of the performances are transformed into a .vtt-file and used to amend the videos. In two pilot runs, $n = 2$ native speakers of English, one of whom does not speak German, watched all subtitled videos to ensure comprehensibility. Within the session, both participants articulated difficulties comprehending the content if the audio and subtitles are both used and reported comprehending more if only subtitles are shown, the audio is muted, and the videos are slowed down. They stated that hearing the occasional use of English language chunks disrupted their reading process and severely impeded their comprehension. Consequently, it was decided to only show the slowed down videos with subtitles without audio for the non-German speaking experts, while the German-speaking experts also received the audio information. In the procedure, the videos were shown in the same order so that the answers between experts are more comparable.

The second part of the expert interviews focuses on the evaluation of the preliminary scoring instrument. This particular order is supposed to counteract priming of the experts with regard to the evaluation of the videos. Also, this order allows experts to evaluate the appropriateness of the items better, as the specific context for which they are designed to be used is evident and items can be discussed against the backdrop of the performances. Before the scoring instrument is shown, experts are presented with the six guiding questions that are continuously referred back to in the interview (see Table 31). The questions cover the experts' general opinion of the scoring instrument (1) and any specific initial observations (2). The subsequent questions target potentially missing aspects (3), aspects that can be excluded (4, 5) and aspects they consider to be important and should consequently not be excluded (6) (see also Appendix D.5).

Table 31: Questions for the evaluation of the scoring instrument.

No.	Question
1	What is your overall impression of the scoring system?
2	Did you notice anything special in particular, which you would like to address?
3	In your opinion: Are any aspects missing?
4	In your opinion: Are there any redundant aspects?
5	In your opinion: Are there any dispensable aspects?
6	In your opinion: Which aspects would you consider to be of particular importance?

The experts are shown slides of a presentation which includes the preliminary version of the scoring instrument, which is based on the feedback framework, by screen sharing via Zoom. The relevant slides of that presentation are found in the guideline in Appendix D.5. Each criterion and the associated items are introduced separately and expert judgment is usually asked for after a set of items was shown. After all items are presented, experts are asked for a final evaluation. This two-tiered approach allows expert knowledge to be accessed implicitly through the evaluative comments, but also more explicitly, by asking experts to evaluate criteria and discuss them, ensuring an in-depth investigation of relevant criteria.

Data preparation

Before the analysis, the recordings of the interviews were prepared. The audio of the interviews is the basis for the transcript. The transcription guidelines are based on the content-semantic transcription as outlined by Dresing and Pehl (2018, pp. 21–22) and are included in Appendix E.2. The interview transcripts are pseudonymized.

Data evaluation – Part I

For the analysis of the evaluative comments of the performances, the approach of a summarizing qualitative content analysis according to Mayring (2022, p. 68) is primarily pursued. A qualitative content analysis allows for a systematic reconstruction of the criteria experts use in the assessment process and is specifically suggested for this purpose by Gläser and Laudel (2010, p. 47). As in the course of the interview reference to previous performances is possible, the first part of the interview that focuses on the performances is the evaluation unit and the interview itself can be considered to be the context unit. Statements by the experts are categorized for each participant, creating a matrix which includes the respective coding unit for all experts and performances. Then, the interview segments are paraphrased to condense meaning. The paraphrases are then the foundation for expert-specific summaries. These are created to evaluate the

specific focus that an expert has with regard to the performances. Paraphrases are also the foundation for the inductive codes. These are developed for a qualitative and quantitative representation of criteria that experts mention overall. Following Mayring's approach (2022, p. 85), codes are created in a first cycle, looking at the material in a chronological order. In a second coding cycle, the set of codes is applied to the whole material. It is not considered how often a criterion is mentioned by a single expert, as it could be assumed that experts might only mention a criterion once explicitly but use it for evaluation implicitly consistently. Rather, the number of experts that refer to specific criteria is of interest. Following the suggestion by Meuser and Nagel, expert knowledge is thus explicated in an open approach, in which interviewees verbalize their decision-making process, allowing for insights into the guiding principles and criteria that experts use to evaluate the performances (Meuser & Nagel, 2009, p. 31).

Data evaluation – Part II

For the analysis of the qualitative interviews in part two, the structuring qualitative content analysis according to Mayring (2022) is mainly followed. This specific approach is considered most useful in order to systematically summarize the different statements that the experts make about the scoring instrument, as it allows for the identification of similarities and contrasts between experts on the same criteria. Categories for the analysis are based on the specific criteria of the scoring instrument, with an additional category based on general comments regarding the RobS and the scoring instrument, subsuming guiding questions 2 - 6 as deductive subcategories. Inductive categories are also added for comments in the interviews that do not fit to these guiding questions. Hence, the strict structuring approach is slightly modified to account for unforeseen statements, an approach which Mayring endorses (2022, p. 107). More detailed categories that relate to the individual criterion, i.e., a positive or negative attitude, are avoided as a comprehensive summary that includes the essence and nuances of all expert judgements is more suitable for the small number of cases. Moreover, not all experts continuously refer to all criteria explicitly, let alone all guiding questions for each item or criterion.

In the coding process, each comment the experts make regarding a specific criterion or with reference to a guiding question is then coded as such. This may relate to full statements, but also mere fragments in other contexts. The coding unit is thus each fragment that specifically relates to a specific item or a guiding question. As these fragments may occur throughout the second part of interview, this is defined as the evaluation unit, while the interview as a whole is the context unit. In the first coding cycle, coding units are identified and assigned to the respective category. This is done to identify if further categories are

necessary for the coding process. In the second coding cycle, the inductively developed categories that refer to the scoring instrument in general, but do not fit to the guiding questions are applied to the whole material. Coding units are then consolidated by paraphrasing and a thematic matrix is created. Paraphrasing is used as a method to consolidate text fragments and make the expert judgements more comparable (Mayring, 2022, p. 70). Based on paraphrases, code summaries are created that condense the experts' judgements and are subsequently used to modify the scoring instrument.

Preliminary scoring instrument

The scoring instrument that is the focus in the expert interviews (see Appendix D.5) differs substantially from the final version presented in Chapter 8.3. In order to contextualize the expert judgements and the results, major differences will be summarized. Most importantly, the preliminary version differed with regard to the actual scoring. First, the maximum amount of points is three. This allows for more differentiation, especially, for the frequency-based items. The maximum of three points is usually scored if behavior is shown as *always*, e.g., "TP always involves the learner in the feedback conversation, e.g., through questions or prompts." (see Appendix D.5), providing an additional level of mastery. Also, criteria are weighted according to their contribution and the perceived difficulty to the feedback conversation. For example, the item *Self-evaluation* is scored with three points, *Positive feedback* with two points and *Self* with one point. Other deviations relate to preliminary item names and specific operationalizations. For consistency, related item titles of the final scoring instrument will be provided in parentheses in the results section.

15.3.3 Results of the Qualitative Expert Interviews

In the following, the results of the expert interviews will be presented. Firstly, the focus will be on the evaluative comments the experts made on the performances, i.e. the first part. Secondly, the evaluation of the scoring instrument will be focused on, i.e., the second part.

15.3.3.1 Experts' Evaluative Comments on the Performances

The results of the content analysis of experts' evaluative comments on performances are presented in this chapter. First, expert-specific summaries will be provided. Second, an overview of the inductively developed codes, which represent the criteria experts use, will be described. The coding manual for the interviews is found in Appendix F.3.

Expert 1

Expert 1 consistently mentions the use of the target language as a criterion, but is unsure how strong it should be weighted in the assessment:

E: Also eigentlich/ Ich war irritiert, dass es jetzt auf Deutsch stattgefunden hat und ich nehme an, die angehenden Lehrpersonen die müssen sollten das eigentlich in der Zielsprache Englisch durchführen. Also davon bin ich jetzt ausgegangen, dass sie das tun sollten und das hat sie nicht gemacht. (E1)

(E: So actually/ I was confused that it took place in German and I assume that the prospective teacher should do this in the target language English. So, that was my assumption that they have to do it and she did not do it.)

In the statement, the expert mentions being surprised that the majority of the conversation took place in German, as they assumed that it was supposed to be conducted in English. The expert also mentions dialogue orientation as an important criterion, which is rooted in recent research on feedback. Furthermore, the potential effect on the learners is highlighted in two regards. First, the recency effect is discussed, especially in the context of the writing process, and the final message that learners take in. Secondly, the expert discusses how learners can actually use the information that was presented to them in the conversation. Also, a deficit orientation is addressed negatively, whereas a focus on learner progression is preferred. Additionally, the expert refers to erroneous statements that the participants make in terms of language use (explicit/implicit) by the participants, but they do not seem to be weighted immensely. Also mentioned are text-specific aspects that are addressed in the conversation: Text surface-level elements do not seem to be important, whereas an overall focus on the genre and deep-level features seem to be more important and a contextualization of mistakes and the task also seem to be important. Lastly, activating prior knowledge and using learning objectives are addressed.

Expert 2

Expert 2 consistently mentions deep- and surface-level feedback of the text as two facets that should be addressed and sufficiently differentiated in the conversation:

E: [...] die Lernende wurde an verschiedenen Stellen mit einbezogen, um den Text zu verbessern, das ist also ein deutliches Plus, und zwar war das sowohl in Bezug auf deep level features als auch auf Ebene der surface level features. Ich fand auch die strukturierte Darstellung überzeugend, weil wir sind von den deep level features dann zu den surface level features gegangen. (E2)

(E: [...] the learner was included at different points to improve the text, so this is a clear plus and this was both in relation to deep-level features as well as surface-level features. I also think that that the structured

depiction was convincing because we moved from deep-level features to surface-level features.)

In the statement, the expert praises the fact that the respective participant addressed both levels and did so in a structured manner. The expert also refers to selective error correction as a criterion throughout the interview. Furthermore, the importance of a structure, like using the three aims of feedback as described in the model of Hattie and Timperley (2007), and focusing especially on the importance of the feed forward is mentioned. Also, the expert states that revision of the text is important and should be advocated for. It is also positively highlighted when the learner is integrated in the conversation (e.g., in a dialogic way) and when the participants are precise in their feedback.

Expert 3

Expert 3 puts a strong focus on the dialogic aspect of the feedback conversation, e.g., asking questions, engaging the learner and adapting to their needs:

E: [...] Also gerade in dem Gespräch wird ja auch der dialogische Charakter vom Feedback ziemlich deutlich, den man oft dann NICHT so hat vielleicht in anderen Situationen oder diejenigen, die noch einem alten Feedbackverständnis folgen. Und ich finde, sie hat das Format auch in der Hinsicht gut genutzt, also, dass sie Fragen gestellt hat an die Schülerin und sie auch aktiv eingebunden hat und, dass sie, ja, auf ihre Rückfragen dann auch eingegangen ist und/ Also das jetzt zu der Gesprächsführung. (E3)

E: [...] So especially in the conversation the dialogic character of feedback is evident, which you often DO NOT see, maybe in other situations or those that follow an old understanding of feedback. An I think, she has also used the format well, so, she asked the learner questions and actively included her and she addressed questions from the learner / so this for conversational skills.)

In the statement, the expert positively highlights the dialogue in the conversation, e.g., asking questions, adapting to learner prompts and engaging the learner in the conversation. Furthermore, the emotional-motivational level and student-teacher relationship is often mentioned as a factor. It is also positively evaluated when participants refer to the material, e.g., the model text or other examples and further resources like dictionaries, especially as a means to support learners. The expert also positively addresses the instances when participants refer to the overall learner progression, including the revision of the text, as well as refer to objectives. The information in the feedback conversation should focus on surface-level and deep-level aspects, but not be provided in a strictly chronological order. When it comes to the communicative aspect, the expert focuses on an overall structure or frame for the conversation and prefers a balanced approach between directly addressing errors and hedging of feedback. Also, participants' specificity and accuracy are mentioned.

Expert 4

The expert often focuses on deep-level features of the text that primarily refer to the genre (e.g., structure, register, tense, and audience orientation) when evaluating the feedback performances and considers these more important than feedback on language. Expert 4 also puts an emphasis on the language of the participant, as a lack of technical terms and use of meta language and specificity are reasons for failure as learners would not know what the mistakes were. Participants' feedback should thus be specific, precise and provided in a structured manner. Expert 4 positively evaluates the provision of positive feedback and the requirement for revision. The expert also comments negatively if self-evaluation is initiated:

E: [...] Now, and I did not like the notion, I mean, it is all about us and students and it was not a rhetorical question, like what is your view, are you happy with the text? What do you think about this? If they are happy with it, how are they going to be open for feedback? (E4)

Expert 5

Expert 5 focuses mostly on three aspects. First, the inclusion and balance of lower and higher-level information in the feedback conversation, comparable to surface- and deep-level features of a text, including positive feedback is mentioned. Second, providing specific feed forward, so that the learner can move forward in their learning progression, especially with regard to the learners' final impression of the conversation is noted. Third, creating a dialogue, establishing a relationship with the learner, and checking on the learners' understanding so that the participants know whether the feedback is understood is discussed:

E: She virtually now, with few exceptions, she never listened to the student at all. It was a straight monologue with very few exceptions and given my knowledge of feedback, it is about, not even, the quality of what is given but the nature of what is heard and understood and there was very little checking to see whether the student understood. (E5)

A sole focus on memorization and providing very detailed information on errors is criticized.

Overview of criteria

The experts mention a variety of different aspects when commenting on the performances. An overview is provided in Table 32. All experts refer to the necessity of addressing text deep-level features in the feedback conversation, four also mention surface-level features. Four experts also explicitly deem it to be positive when the teacher and learner engage in a dialogue, e.g., when the teacher asks questions and appropriately reacts to the learner's questions. The specificity of the information is mentioned by four experts as a criterion, too, as

is the necessity of prompting the revision of the text. Three experts address the provision of learning objectives as feed up, the overall structure of the feedback conversation and the mentioning of positive aspects.

Some criteria have been addressed by fewer than three of the experts. They focus more on the learner to a certain extent: the final message they receive, the general support of the learner, their ability to use the feedback and the teacher ensuring that the focus is on learner progression, and the emotional-motivational level. Other aspects that are addressed by two experts refer more to the teacher behavior: avoiding a deficit orientation, providing feed forward, providing accurate information, and using the material or referring to other external material like textbooks and dictionaries. Aspects mentioned once are the use of target language, contextualization, activation of prior knowledge, selecting errors, balancing hedged communication, using meta-linguistic language, and initiating self-evaluation.

Table 32: *Criteria for the assessment of feedback performance mentioned by the experts.*

Code	Code description	<i>n</i>
Text deep-level features	Expert addresses text deep-level features.	5
Dialogue	Expert addresses the engagement of the learner in a dialogue.	4
Text surface-level features	Expert addresses text surface-level features.	4
Revision	Expert addresses the advocacy of revision.	4
Specificity	Expert addresses specificity.	4
Learning objectives	Expert addresses learning objectives (Feed Up).	3
Structure of the feedback conversation	Expert addresses the structure of the feedback conversation.	3
Mentioning positive aspects (first)	Expert addresses mentioning positive aspects (first).	3
Final message (Recency)	Expert addresses the finale message that learners have of the feedback conversation.	2
Learner support	Expert addresses the support of the learner.	2
Learners ability to use the information / Checking understanding	Expert addresses the learner's ability to use the feedback and participants making sure of it.	2
Deficit orientation	Expert addresses deficit orientation, i.e., a detailed focus on mistakes and errors.	2

Code	Code description	<i>n</i>
Learner progression	Expert addresses the reference to learner progression, either to past or future.	2
Accuracy (Participant)	Expert addresses participants accuracy, either with regard to language or other professional aspects.	2
Emotional-motivational level	Expert addresses the emotional-motivational level and relationship.	2
Use of (external) material	Expert addresses the use of (external) material.	2
Feed forward	Expert addresses the provision of feed forward.	2
Use of target language	Expert addresses the use of target language.	1
Contextualization of errors and mistakes	Expert addresses the contextualization of errors and mistakes in the text context.	1
Activating prior knowledge	Expert addresses the activation of prior knowledge.	1
Selecting errors	Expert addresses the selection of errors.	1
Balancing hedged communication	Expert addresses balancing hedged communication with more direct feedback.	1
Self-evaluation	Expert addresses the initiation of self-evaluation.	1
Metalinguistic language	Expert addresses the use of metalinguistic language.	1

In conclusion, three main aspects are addressed by a majority of experts when confronted with the performances. The first aspect refers to the content of the feedback conversation, with a clear focus on providing both surface-level feedback and deep-level feedback. It is criticized when only the surface-level is addressed, though a focus on only deep-level feedback is not addressed by the experts. The fact that information should be provided with specificity, so that learners know where exactly the gap is, and what precisely they have to do in order to close that gap, and that they include statements with regard to what the learner has already achieved is also addressed by four experts. Also, most experts address the necessity of a revision of the text, using the feedback in that process to achieve the desired level. The second aspect is reflective of the general approach to feedback, for which experts clearly favor a dialogic approach. The learner should be included in the conversation, e.g., by asking questions and establishing a relationship between the teacher and learner that is conducive to learning is encouraged. The third aspect refers to structure, as three experts address the notion of an overt conversational structure, i.e., following the model by Hattie and Timperley (2007) and stating learning objectives.

15.3.3.2 Evaluation of the Scoring Instrument

In the following, the results of the structuring content analysis with regard to the evaluation of the scoring instrument are presented. Item-specific summaries are presented first, followed by the summaries of the remaining codes that refer to the guiding questions or the scoring instrument in general. The coding manual for the interviews is found in Appendix F.3.

Self-evaluation

This criterion is discussed by four out of five experts. Overall, those experts are in favor of including this notion in the evaluation sheet:

E: Ja, also ich finde ganz zentral auch auf jeden Fall diese den Einbezug von den Lernenden durch die Self-Evaluation, also Selbstevaluation gleich am Anfang. Also wirklich, dass es ein die Lernenden auch kognitiv auch aktiviert werden und ihnen nicht einfach nur Feedback (lachen) ja gegeben wird ohne, dass die das nur aufnehmen und nicht selber mitdenken. (E1)

(E: Yes, I think this is really key this inclusion of the learners by self-evaluation, so self-evaluation at the beginning. So that the learner is really cognitively activated and they do not (laughs) just receive feedback without taking it in and thinking along.)

Expert 1 emphasizes the relevance of self-evaluation occurring at the beginning as a means of cognitive activation. Expert 3 positively remarks on the higher weighting of this criterion with three points. On the contrary, expert 2 wonders whether this criterion should be scaled rather than operationalized as a dichotomous item so that the variance with which the criterion can be fulfilled is depicted in a more nuanced way. Expert 4 also comments that the exact formulation of the self-evaluation prompt could be considered and argues against the high weighting of this item.

Positive feedback

This criterion is addressed by three of five experts and is generally not argued against. Expert 2 questions whether this notion should receive less points than self-evaluation, as it is scored with two points. Expert 1 and 4 are rather in favor of including this aspect, but add some caveats. For expert 1, this aspect has to be considered within the whole context and structure, discussing potential negative effects from the feedback sandwich and recency effect. Expert 4 emphasizes the importance of specifically mentioning what has been done well and not providing generic praise:

E: [...] I do not care about the positive, I think (inaudible) for always trying to / people that always tell kids they are doing well, but mentioning specifically what you have done well is good. (E4)

Learning objectives (Feed Up)

Two out of the five experts address this criterion. Expert 1 generally regards this criterion positively. Expert 4 questions the different point distribution between feed back and feed up, as the former is the key component of the scenario. The expert further argues that it is more important that feedback is given against the objectives than to simply state them:

E: I would have given, for the state the learning objectives one, i.e., knowing what they are, because they should have known them already and then they give feedback against them as being the BATTLE thing you know rather than / you know, that just seems very odd. [...]

E: If you are going to give them different points, I think from my perspective these are not weighted they I would weight them. I would not put so much emphasis on just stating what the learning objectives are, maybe revisiting briefly but the IMPORTANT thing is, the learner receives feedback about performance AGAINST the learning objectives, because you are stating the learning objective and then you only give them hint for learning objective. There are several objectives, we need to know how we are going against them. For me, the important thing, it is often the feedback is not related to what the aims were, what the learning objectives were, but you got to give three for feed forward. (...) Anyway. (E4)

Feed Back

The comments for this criterion resemble that of the previously described *Learning objectives (Feed Up)*. Two out of the five experts address this criterion and do so generally positively. Expert 1 addresses this criterion positively as a facet of relevant theory:

E: [...] [es ist] euch gelungen, aus meiner Sicht, sehr viele Aspekte abzubilden, die relevant sind in diesem Kontext. Auch mit Bezug auf die relevanten Theorien. Ich denke jetzt gerade an zum Beispiel an Feed Up, Feed Back, Feed Forward. (E1)

(E: [...] you really managed, in my point of view, to depict many aspects that are relevant in the context. Also the reference to the relevant theories. I am thinking about feed up, feed back, feed forward.)

Expert 4 questions the different point distribution between feed back and self-evaluation:

E: I am also looking at the left-hand column, here you have the points allocated and those vary from one to three. Three for self-evaluation and only one for feedback. (laughs) That seems very odd to me. Is that based on what is on the next slide or /. (E4)

As stated above, expert 4 also argues that feedback is supposed to be given against specific learning objectives.

Feed Forward

Four out of five experts address this criterion. Three experts regard this aspect generally positively. Expert 4 also emphasizes the appropriateness of the high weighting of three points. Expert 2 is generally in favor of including feed forward, but proposes that points should be awarded for each instance in which a 'good' feed forward is provided, stopping at three feed forward instances to ensure focus:

E: Okay. Was für mich nicht ganz klar ist, ist warum manche Sachen mehr Punkte bekommen als andere. Also wäre es nicht eigentlich so, dass jedes Mal, wenn ein gutes feed forward gegeben würde, einen Punkt bekommen muss. Also da ist ja so zusagen so eine Schwerpunktsetzung in den Punkten und da gibt es vielleicht einen Grund für, aber das fällt mir jetzt erstmal auf. Also da würde ich mich fragen, aha wieso bekommen wir da, also es kann ja sein, dass da jemand nur ein feed forward gibt, ist vielleicht nützlich, aber der andere hat vielleicht zwei gute feed forwards und ab einer bestimmten Zahl geht es ja aber auch wieder runter, weil wenn wir zu viel feed forward geben, dann gibt es einfach keinen Fokus mehr, keinen Feedbackfokus und dann ist der Lernende überfordert. Da müsste man sich vielleicht ein bisschen was einfallen lassen, also was ist hie sozusagen das optimale Level, ich finde immer so drei Punkte ist sicherlich was, was man ganz gut sagen kann, ich habe jetzt four aspects oder three aspects, two in a deep level und vielleicht two in a surface level oder so, dass man das irgendwie deutlich macht und für sich selber müsste man dann auch irgendwie klar kriegen, was wäre eigentlich zu viel für so ein Feedback. (E2)

(E: Okay. What is not quite clear to me is why some things get more points than others. So, would it not actually be the case that every time a good feed forward was given, it has to get a point. So, this is kind of an emphasis in the scoring and maybe there is a reason for that, but this is what I am noticing. So, I would ask myself, aha, why do we get, well, it could be that someone only gives one feed forward, which may be useful, but the other person may have two good feed forwards and from a certain number it goes down again, because if we give too much feed forward, then there is just no more focus, no feedback focus and then the learner is overwhelmed. One might have to think of something, so what is the most optimal level here, so to speak, I always think three aspects is certainly something that you can say quite well, I now have four aspects or three aspects, two on a deep-level and maybe two on a surface-level or something, so that it is clear and then one would also have to somehow make clear what would actually be too much for such feedback.)

Task & Process

None of the experts addressed these criteria explicitly.

Self-Regulation

Two out of five experts address this criterion. Expert 2 criticizes that the quality of self-regulative feedback is not addressed in this criterion, using examples from the previously shown performances:

E: [...] Auch bei der Selbstregulation, wie hilfreich ist es jetzt im Oxford dictionary nachzugucken, wenn wir im Zeitalter von Digitalisierung sind, ja machen wir einen tick Strategie, aber ist diese wirklich noch so der Brenner. Oder der Hinweis, guck mal wir haben hier am Ende des Buches nochmal die Liste der irregular verbs, das finde ich ist eigentlich eine ganz gute Sache, geh die nochmal durch und vielleicht gibt es die dann auch auf irgendeiner App was auch immer, wo man die dann nochmal lernen könnte. (E2)

(E: [...] Also with the self-regulation, how useful is it to look something up in the Oxford dictionary, when we are in the age of digitalization, yes we tick strategy, but is this really working. Or the hint, look here, at the end of the book is the list of irregular verbs again, I actually think this is a good thing, look at them one more time maybe there is an app or whatever, in which one could learn them again.)

Expert 4 also views this critically, as this aspect is rarely seen in the seen sample videos and the quality of information on self-regulation can differ.

Self

One expert addresses this notion. Expert 4 discusses the challenges when it comes to differentiating personal praise from positive feedback on the text:

E: Oh, this is something negative. You get a point if you do not say it (laughs). Okay, sorry, I misinterpreted that entirely. So it does not address the learner's / there is no generic feedback about, on a /

I: Yes, no generic praise.

E: Although, some say, in the videos that it was a good text.

I: Yeah, this is a grey area right. It is somehow related to the task because the task was the text but it is still generic, so.

E: It has to not related to the text. A good text is alright but all good effort which someone / effort like in video four, he made a good effort, that is a personal thing is it not?

I: That is true. (E4)

Furthermore, the expert states that a learner's affective dimension is not included in the set of criteria which *Self* could address:

E: [...] So, what is missing is, yes maybe, something that has to do with the effect of aspects of both self-evaluating, of getting feedback, of the self of the writer or even, I mean, some students may find this really engaging and others sort of was not, I do not know. (E4)

General comments on the items for General aspects of feedback I (Feedback functions)

Expert 5 is generally positive towards all these previously outlined items without providing specific comments:

E: No, I think that is pretty much going on all different levels. [...]

E: No, I am sorry. It is good. Obviously, I like the criterion. (E5)

Dialogue orientation

All experts address this notion and do so generally positively. Expert 1 emphasizes the importance of this criterion in light of recent research on feedback. Expert 5 adds that the dialogue should also include checking on the learner's understanding:

E: [...] I like the dialogue orientation because it should be/ it is not just a dialogue it is does both parties hear and understand the other party. And so, after the (inaudible) conversation where you reinforce you know, no that is not what I meant, this is what I meant, et cetera. So, I can see how that comes out. So, I (inaudible) with the dialogue as much as involving, that checking the learners understanding. (E5)

Expert 4 criticizes that this aspect does not consider the quality of the questions or prompts, arguing that cognitively engaging questions should be asked and frequency should not be the focus of the scaling but rather quality.

Quality of Feed Back

This criterion is addressed by four of the five experts. Expert 4, however, only questions whether the criterion is highly inferential¹²³. Expert 3 and 2 are generally positive towards the criterion, though the latter proposes checking if this scale is applicable to the scenarios:

E: Quality of feedback, accurate, sometimes accurate, often accurate. Ja das finde ich ist eigentlich ganz gut, hier hätten wir dann auch diese Skalierung, wo man sagen könnte, so in den Bereichen müsste man nochmal gucken, wie gut das eigentlich passiert. (E2)

(E: Quality of feedback, accurate, sometimes accurate, often accurate. Yes, I think this is actually quite good, we have the scaling in which one could say, so you in these areas one would have to see, how good does this actually happen.)

Expert 5 argues that being descriptive alone is not sufficient for quality feedback, and that it should be balanced with regard to self-regulative aspects, which are not necessarily descriptive.

¹²³ In previous iterations of the scoring instrument, a preliminary distinction between highly and lowly inferential criteria was made and included in the slides.

Quality of Feed Forward

This criterion is explicitly addressed by one expert. Expert 5 positively evaluates this criterion, though adding that it should include checking if learners understood what was being said:

E: [...] Now quality of feed-forward, now I like that, it is the one that matters the most to me and it is also coming back to the dialogue. Do the teacher check that the student understood what the call of action was or just go back and do it again. Well, this is what I want you to find out, I would like a little bit more of that in that criteria. (E5)

Securing understanding / content adaptation (Checking understanding)

Three of the five experts address this criterion. Expert 5 and 1 are favorable towards the criterion. Expert 1 adds that the use of external material, like the texts, paper, pens or gestures, could be included to ensure understanding, although stating that this is not central:

E: [...] Und das ist ein schöner Ansatz, das mal ein bisschen zu explizieren, was genau könnte das alles beinhalten und deckt eigentlich Dinge ab, die aus meiner Sicht, zentral sind. Was vielleicht nicht drin ist, oder wenn ich oder vielleicht habe ich es übersehen, sind so Strategien, die der Adaption dienen können oder bei der vielleicht Verständnissicherung wie Hilfsmittel, Text, Papier, Stifte, Gesten irgendwie so. Aber das ist vielleicht auch nicht zentral. Ja. (E1)

(E: [...] And this is a nice approach to explicate, what could all this include and does it cover aspects that are key in my view. Maybe what is not included or maybe I did not catch it, are strategies that could serve adaptation or maybe with securing understanding like scaffolds, text, paper, pens, gestures or so. But maybe this is not so important. Yes.)

Expert 4 also suggests the use of the terms 'explanations' or 'rephrasing' instead of 'answers,' as the latter would imply a focus on corrective feedback.

Linguistic adaptivity (Adaptation of language)

All experts address this criterion but do so to a varying degree. Expert 1 in this context emphasizes positively that the appropriate use of both languages is focused on here. Expert 3 discusses the amount of points within the larger context of the assessment. Experts 2 and 3 consider the scaling of the criterion as generally logical. Expert 2 is critical of the general idea of functional monolingualism in the EFL classroom, citing a lack of evidence, questioning the effectiveness of speaking English for the feedback conversation, and framing it more as a normative objective:

E: Ja, finde ich auch interessant. Ich meine hier kann man ein bisschen, hier kommt man in so einen Graubereich, der ja nicht gut erforscht ist. Das ist sozusagen ein didaktisches Desiderat möglichst viel auf Englisch zu machen, ob das jetzt sozusagen im Feedback den Lernenden hilft, ist

ja so ein bisschen unklar und wenn wir jetzt von code switching ausgehen würden, dann würde das wahrscheinlich hilfreich sein, wenn man switched, aber so richtig hilfreich war das noch nie, deswegen finde ich, also das ist sozusagen auf einer anderen Ebene, das ist eher auf einer didaktisch normativen Ebene, als einer empirisch gesicherten Ebene. Ich glaube das ist sozusagen ein Bereich, den man für sich selber so einmal klarkriegen muss, denn hier geht man auf eine andere Ebene, das heißt nicht unbedingt, dass also vielleicht wären die Lernenden auch fähig das auf Englisch zu verstehen, also das weiß man halt einfach nicht so. Deswegen habe ich jetzt hier keine direkte Anregung, ich finde es nachvollziehbar, gerade aus fachdidaktischer Richtung, würde jetzt aber keine Aussage darüber machen, was wirklich so wichtig ist jetzt für den Lernenden, außer das Zeitargument ist immer eins, also je mehr Kontakt die mit dem Englischen haben ist natürlich desto besser. (E2)

(E: Yes, I also think that is interesting. I think you can get into a bit of a gray area here that is not well researched. It is, so to speak, a didactic desideratum to do as much as possible in English, whether that helps the feedback for the learners, so to speak, is a bit unclear and if we were to consider code switching, then that would probably be helpful if you switched, but it has never been really helpful, so I think this is on a different level, so to speak, it is more on a didactically normative level than an empirically proven level. I think that's is an area that one has to make clear, so to speak, because this is going on a different level here, which does not necessarily mean that, maybe the learners would also be able to understand it in English, so you just do not really know that. Therefore, I do not have a direct suggestion here, I think it is understandable, especially from a didactic point of view, but I would not make a statement about what is really so important for the learner now, except the frequency argument is always one, so the more contact they have with English is of course the better.)

Like expert 2, expert 5 also comments that it is important that the learner understands what is being said. Expert 4 assumes that this criterion could be rated reliably, though questions whether it could be constructed in a less complex manner.

Adaptive error correction

Four experts address this aspect. Expert 1 is generally positive towards the differentiation of the adaptivity construct and expert 5 only asks a comprehension question. Expert 3 sees a connection between this aspect and that of dialogue orientation. Expert 4 views the complexity of this criterion critically, especially concerning the pragmatics of the rating process, due to the difficult differentiation between implicit/indirect feedback and unspecific feedback. They argue that, potentially, the provision of opportunities for self-correction could be the important aspect here:

E: I would not like to be doing this on the fly as it were, in other words in real-time. I would not like to be (in account?) to make all of these decisions at the time I am viewing the video. [...]

E: The second category there is / gives them an opportunity, and when they do not get it right, I just go straight to the direct but this one you are saying the opportunity is THROUGH the implicit and indirect and not just by invitation. So, the second one could be, you know, you see any problem with this particular reader's perspective, with this particular construction? And then I say NO, you see this (laughs)? On the first one I have got to do it much more indirectly and I am not sure what that would look like and if I see it. So, maybe those, the important aspect is, is there an opportunity for self-correction? (E4)

Prioritization of content/structure (Prioritization of deep-level features)

Two experts address this criterion. Expert 4 only peripherally addresses this criterion, but does so positively. Expert 3 argues that participants should be able to fulfill this aspect as the material directs them towards this notion:

E: [...] Ja genau und mit diesem Fokus auf den Inhalt und, also Struktur, das war ja auch das, was ich eingangs fragte, wie stark soll das denn jetzt gewertet werden? Weil das ja auch in der Aufgabenstellung auch so rüberkam, also hätte es eigentlich den Kandidatinnen da auch klar sein sollen. Ich glaube aufgrund dieser Menge an Fehlern die da waren, haben sie sich oft daran aufgehalten an den sprachlichen Sachen. (E3)

(E: [...] Yes exactly and with this focus on the content and, so, structure, this is what I asked at the beginning, how relevant is this for the evaluation? Because looking at the task it seems like the participants should have known this. I believe based on the number of mistakes they often stuck to the linguistic aspects.)

Appropriate amount of information (Amount and balance of information)

Four experts address this aspect, generally more critically. Expert 2 questions the potential inter-rater reliability, as an 'appropriate' amount is not defined more closely. At a different point in the interview, expert 2 argues that three to five aspects are an appropriate number:

E: [...] da würde man glaub ich eine super inter-rater reliability hinbekommen bei allen Punkten bis auf appropriate amount (unv.), vielleicht könnte man da noch eine (unv.) irgendwie geben, also wie viel soll es denn sein in einem zehn minütigen Gespräch müssten zumindest irgendwie drei Punkte genannt werden, aber es dürfen auch nicht zwanzig genannt werden, also das müsste man irgendwie doch, weil ich könnte mir sonst vorstellen, dass man nicht mit diesem Bogen auf eine gute inter-rater reliability kommt. (E2)

(E: [...] there, I think, one would get a great inter-rater reliability with all points except amount (inaudible) maybe there could be a (inaudible) somehow, so how much does it have to be in a ten-minute conversation at least three points must be mentioned but it should not be twenty, so that is something that because I imagine that with this instrument one does not achieve good inter-rater reliability.)

Similarly, expert 4 addresses the challenge of adapting to a learner who the participants know little about, also proposing to focus on three different aspects. Expert 3 is generally positive towards the criterion and mentions specific aspects that should be addressed with regard to the text, e.g., structure, style, error patterns, false friends or the German quote, while other aspects could only be mentioned but not explicitly focused on. Expert 5 argues that not the amount, but rather the balance between lower-level information and higher-level information is to be considered, as both aspects should be discussed in a balanced matter:

E: I would say the TP provides an appropriate amount and BALANCE of information. And whatever words you used to service at depth to improve the text quality. Because you said comparing on the length the conversation was is a separate way. It is that balance that is missing, otherwise person three would come out very high and person three was overly emphasised the amount of information. (E5)

Language correctness (Surface: Language correctness) & Language adequacy (Surface: Language appropriateness)

Only expert 4 addresses the criteria. The expert argues that both criteria could be combined and the genre-aspect could be emphasized more:

E: The one and two seem to be almost something, you could subsume them, language accuracy and language appropriacy, inappropriate use of tense, forming of the simple past, well / at some point or some sort of grammatical, anyway. [...]

E: I have not heard of genre conventions. I would put we¹²⁴ in the genre conventions. (E4)

Content (Deep: Content)

Only expert 2 addresses this notion and does so generally positively when addressing the diagnostic dimension:

E: Ne, also ich meine der Punkt mit den Punkten da haben wir ja schon drüber gesprochen, dass das schwierig ist und ansonsten die Idee zum content, structure und zum Prozess was zu sagen, finde ich auch eine gute Idee auf jeden Fall. (E2)

(E: No, so I think the issue with the points we already have addressed, that this is difficult and otherwise the idea of content, structure, and saying something about process, I think it is a good idea.)

¹²⁴ The expert refers to the use of the pronoun *we* in the news report as an indication for a less neutral style of writing.

Structure (Deep: Structure)

Two experts address this notion. Expert 2 generally positively, as indicated above. Expert 4 highlights the importance of structure when it comes to text production:

E: [...] I might think that the notion of / the features of text normally associate with a specific purpose or certain genre. This could be perhaps THE most important. So those will be things like the headline and the structure and the useful things like WE are appropriate which some people might put into / oh, I do not know I would put it there. I do not know, I just think you are giving a lot of emphasis to process when in these videos I do not think I saw any discussion of the process. (E4)

The expert argues that this aspect should be extended with regard to addressing the audience properly, suggesting a restructuring of this aspect as well as *language adequacy* to emphasize the relevance of the audience:

E: [...] There is no category that deals with the reader and the rhetorical purpose of the text which is one that we / If you are thinking about hierarchy, we start with the rhetorical purpose and then we might move to the content and structure, and then we might move to language resource use and then we might move to the mechanics of punctuation, spelling, grammar. [...]

E: I think I would address the audience for this and then extend to if all of this, is that actually appropriate for the audience, not just accurate and appropriate use of tense but what it / the way it is structured, so the headline / those who talked, I mean our teachers and classes will immediately talk about hooking the reader into your piece and you would do that with a headline. And some of them talk about something that was interesting and grabbing attention and so on. (E4)

Process (Deep: Process)

Two experts address this notion. Expert 2 generally positively as shown previously. Expert 4 questions the need for this aspect as it is rarely seen in the performances and also wonders about its high weighting:

E: Is there any evidence of that one, when you are giving that three points? I might think that the notion of / the features of text normally associate with a specific purpose or certain genre. This could be perhaps THE most important. So those will be things like the headline and the structure and the useful things like WE are appropriate which some people might put into / oh, I do not know I would put it there. I do not know, I just think you are giving a lot of emphasis to process when in these videos I do not think I saw any discussion of the process. (E4)

Process orientation (Revision)

Three experts address this notion. Expert 4 only states that they imagine that this aspect is easy to rate. Experts 2 and 3 evaluate this aspect generally positively. Expert 2 agrees that the focus on revision is sufficient in this context for

the consideration of the writing process, as the planning aspects of writing do not seem to be a problem in the presented scenario:

E: [...] und bei der process orientation das finde ich gut, also das klar wird, dass man es auf jeden Fall überarbeiten muss. Jetzt müsste man sich natürlich fragen, ob man auch fragen soll, wie die Lernende den Text geplant hat, also wenn man jetzt prozessorientiert vorgehen würde, könnte man auch nochmal sagen in diesen Bereich Planung zu gehen, aber ich finde es auch legitim, sich nur auf die revision zu stürzen, weil wir ja sehen, dass eigentlich das Planen nicht das Problem ist, das machen ja auch Lehrkräfte häufig im Unterricht, aber sie fordern zu wenig Überarbeitung ein. (E2)

(E: [...] and with process orientation that I think is good, so that it is clear that you have to work on in in any case. No one would have to wonder whether one should ask, how the learner has planned the text, so when you take a process-oriented approach, you could include the area of planning, but I think it is legitimate to only rely on revision because we see that the planning is not really the problem, teachers do that frequently in classes, but they demand to few revisions.)

They argue that, based on empirical findings, the revision aspect is an issue in the classroom while planning does occur, legitimizing the focus on revision here. Expert 3 positively mentions this aspect when asked which aspects stood out to them, adding that this should be considered against the backdrop of the series of lessons.

Error pattern

Four experts address this aspect. Expert 1 and 3 are generally positive towards this notion:

E: Ja auf jeden Fall mit der Identifizierung von den Mustern, das finde ich wichtig, weil das hilft ja am meisten und das andere sind eher error, mistake, nur diese Unterscheidung, dass das hier auf jeden Fall auch Stellenwert hat. (E3)

(E: Yes, absolutely with the identification of patterns, I think this is important because this helps the most and the rest are rather errors, mistakes, just this distinction, that this also has value.)

Expert 3 states that the identification of patterns is important as it is most helpful and adds that it must be ensured that the distinction between errors and mistakes is represented. Experts 2 and 4 are not generally against the idea of using or focusing on error patterns. However, they strongly criticize the frequency-based scaling and the operationalization of exclusively focusing on patterns, as this would exclude other potentially more relevant aspects of the text that cannot be grouped into certain patterns:

E: Why would they get three points if they do that exclusively? Which is really poor-quality feedback if you do that exclusively (laughs).

E: No, no. You could say PREDOMINANTLY (...) addresses error patterns, which I do not think is good. I mean, I would not give them three points for that (laughs). This is the problem with your frequency scale. You are implying that it is good to do this a lot. (E4)

Expert 4 adds that the way these patterns are communicated should also be addressed. Both experts 2 and 4 are open to the idea of merging this aspect with *Target-oriented error selection* and expert 4 proposes to have that as a dichotomous category:

E: Anyway, that is just my sense, I am not sure what, going through this, what you could DUMP. Maybe simplify some of the ones like the language in which the interaction is conducted. Maybe those categories could be simplified, I do not know, and the one that we talked about that could be a dichotomy rather than, the error pattern. [...]

E: The target and the pattern, the target-oriented (inaudible) that goes into one. (E4)

Target-oriented error selection

Four experts addressed this aspect. Expert 1 is generally positive towards this criterion:

E: [...] Was ich besonders toll finde hier ist das target-oriented error selection, also eigentlich die Kombination zusammen mit Error pattern. Das finde ich sehr gut, weil es sehr oft halt in diesen ja Feedbackgesprächen, wenn es um Texte geht. Ja, ich meine ein Beispiel von diesem vom Video, das durchgefallen ist, so dieses so ein klassisches Beispiel von einfach mal alles aufzählen, was auffällt und halt eben alles gar nicht mit im Blick hat. (E1)

(E: [...] What I find really great here is the target-oriented error selection, so the combination with error pattern. I think that is very good because these feedback conversations are often about texts. Yeah, I think one example in the video, that failed, so this classic example of just recounting everything that stands out and not have everything in mind.)

The expert points out that in this facet, in combination with *Error pattern*, it is important to take into account that not all text features are addressed. Expert 2 rates this aspect as more important than *Error patterns*. Expert 4 proposes a reformulation, so that the relation to the learning objective is mandatory for the frequency. Both Experts 2 and 4 are open to the idea of merging this category with error patterns, as discussed above. Expert 5 adds that the selection of errors should also be based on contrasting errors with instances in which the error has been avoided.

Detailedness

Two experts address this aspect. Expert 5 is in favor of this category and proposes to include the aspect of balancing information here. Expert 2 comments

on the term 'elaborated feedback' as it is not clearly defined in research. They argue that the possible types of elaboration should be made explicit for the rating process:

E: [...] hier also das ist ja auch so ein schwieriges Feld mit dem elaborated feedback, hier würde ich ein bisschen genauer sein und sagen, also was Sie unter elaborated verstehen. Also geht es darum, dass sie z.B. eine Regel sagen, weil manche Sachen haben ja eine Regel tatsächlich, also wenn es jetzt die regelmäßige past tense gibt, dann gibt es die -eds aber leider gibt es auch die unregelmäßigen Verben, wo man sie einfach auswendig lernen muss und so weiter, also, weil elaborated wird ja von manchen Forschenden wie Ferris die denken dann eher an metakognitive und so, aber andere denken eher an besonders ausführlich und vielleicht könnten Sie sich da festlegen, was Sie als elaboriert hier finden, also soll sozusagen eine Regeln erklärt werden oder worum geht es genau bei dem elaborierten, weil das wird ja auch in der Literatur tatsächlich ein bisschen, [...] also auch sehr unterschiedlich manchmal genutzt dieser Begriff, was ja auch verwirrend ist. (E2)

(E: [...] here so this is also a difficult area with elaborated feedback, here I would be a bit more precise and say what you mean by elaborated. So is it about saying, for example, a rule, because some things actually have a rule, so if there is the regular past tense, then there are the -eds, but unfortunately there are also the irregular verbs, where you just have to learn them by heart and so on, so, because elaborated is used by some researchers like Ferris who think more in terms of metacognitive and so on, but others tend to think of really detailed and perhaps you could define what you consider elaborated here, i.e., should a rule be explained, so to speak, or what exactly is elaborated on, because this is actually in the literature, [...] so this term is sometimes used very differently, which is also confusing.)

Linguistic accuracy (Language accuracy)

All experts address this criterion. Experts 5 and 1 are generally positive towards it. Expert 1 wonders about the applicability in terms of the amount of English that is spoken in the dialogue and proposes a merger with pedagogical accuracy. Expert 2 views it more critically, is doubtful concerning to the amount of English spoken, and also proposes a merger with pedagogical accuracy. Expert 2 further explains that a subsequent differentiation regarding the correctness of the feedback in general, possibly with regard to surface- and deep-level could be an option. Also, expert 2 states that this aspect is more appropriate for evaluating classroom instruction than a feedback scenario:

E: Ja, also ich glaube das ist eine super Kategorie, wenn Sie jetzt Unterricht bewerten würden, aber ausgerechnet in der Feedbacksituation ist es ja sogar ganz gut, wenn die wahrscheinlich auf Deutsch zurückgreifen, weil also was jetzt schon scheinbar schwierig ist für die Lernenden und dann sie wahrscheinlich gerade deswegen auch auf Deutsch zurückrufen, weil die Schüler es nicht können und dann können Sie es halt gar nicht bewerten. Ich meine da müsste man halt überlegen inwieweit

Sie diese Kategorie der linguistic (unv.) überhaupt bewerten können in Feedbackgesprächen, weil es vielleicht häufig nicht gegeben ist, also da müsste man halt gucken. (E2)

(E: Yes, I think this is a great category if you were to evaluate instruction but especially in feedback situations it may be quite good when they use German because what is apparently quite difficult for the learners and when they then use German because the learners are not able to do it and then one cannot evaluate it. I think one has to think about the extent to which this category of linguistic (inaudible) is assessable in feedback conversations because often it is just not a given, so there one would have to take a look at.)

Expert 3 favors the inclusion of this aspect as the teacher is a language model, but deems the inclusion of technical terms to be lacking. Expert 4 also criticizes that the use of meta-language or technical terms is missing here. Furthermore, expert 4 strongly criticizes the combination of a pass/fail and frequency scale.

Pedagogical accuracy (Professional accuracy)

All experts address this aspect and are generally positive about it. Expert 3 and 5 are generally positive towards this aspect. Expert 4 wonders about the focus of learning strategies in the examples, as they are rarely evident in the performances they saw. Experts 1 and 2 propose a merger with the previous category. Expert 1, for example, proposes putting that criterion on a continuum with pedagogical accuracy as the two criteria are not necessarily clearly distinguishable:

E: Fachlichkeit genau. Ja. (...) Gut und das finde ich auch das finde ich toll auch hier diese Ausdifferenzierung zwischen diesen beiden Kategorien, was auf jeden Fall Sinn macht und/ Das finde ich auch spannend, weil ich habe manchmal so das Gefühl, es ist fast wie ein Spektrum oder ein Kontinuum von Linguistic accuracy und pedagogical accuracy. Wo bewegt man sich dazwischen? Schwierigkeit ist wirklich das so klar auseinander zu halten, weil man das ja irgendwie/ Man muss ja irgendwie. (E1)

(E: Professionalism. Yes. (...) Good, and that I also think is great, this differentiation between these two categories, which makes a lot of sense and/ I think this it is interesting because I sometimes have the feeling it is like a spectrum or a continuum of linguistic accuracy and pedagogical accuracy. Where are you in between? The difficulty is really to distinguish this clearly, because some / yeah you have to.)

Overall evaluation

Three experts comment on the scoring instrument in an overall evaluative manner. Expert 3 describes the instrument as logical and novel. Expert 4 also considers this approach to be novel, but is skeptical overall with regard to pragmatics. Expert 5 is generally positive towards the instrument:

E: No, I like where you are going. I like what you are doing. (E5)

Comprehensiveness

Two experts address the comprehensiveness of the scoring instrument explicitly and point to possible challenges for scoring as a result. Expert 1 addresses the comprehensiveness tendentially positively, but points out that raters might encounter difficulties when rating performances using this instrument due to cognitive load. Expert 4 is also skeptical as the instrument may not be pragmatic for use:

E: I think you are being quite exhaustive in terms of coverage. My concern will be pragmatics. [...]

E: [...] I think if you train expert raters to utilize this it is possible, I think this is going to be quite complex (...) and I do not know that it would then translate into something rather professional even an experienced professional could really use. (E4)

Scoring

Three experts also address the general aspect of scoring validity and reliability. Expert 2 generally questions the different weightings of criteria as there is not necessarily evidence that would support this and proposes dichotomous items that are weighted equally. In some instances, the expert also argues for providing points for instances in which a successful feedback event occurred instead of considering the relative frequency (see the section on *feed* forward in this chapter for a direct quote of E2). Expert 4 proposes not having all raters evaluate all facets, but rather have them focus on one in order to achieve higher degrees of inter-rater reliability. Expert 5 says that the evaluation sheet offers the potential for evaluating feedback criteria that are evident in the sample performances. With regard to a pass/fail scheme, expert 5 proposes an alternative approach, determining a pass level based on the items that are relevant for passing and accumulating these points, rather than using a simple 50 % level of overall points.

Based on previous research

Expert 1 and 2 positively address that the criteria reflect current research in the field:

E: [...] Aber ich glaube nicht, dass sich etwas doppelt, also mein Gesamteindruck ist gut, die Kategorien ergeben sich ja aus der Literatur, finde ich sind schön herausgearbeitet. (E2)

(E: [...] But do not believe that something is a duplicate, so my overall impression is good, the categories are derived from literature, I think they are well worked out.)

Positively emphasized

Three experts emphasize different items in the scoring instrument. Expert 1 singles out the dialogue orientation as it reflects their understanding of feedback as literacy. Expert 3 singles out self-evaluation as important. Expert 5 emphasizes the aspect of checking understanding as it points to making sure that learners can use the feedback:

E: Overall impression, [...] I like/ We did not see it very often at all, but I like the notion of trying to understand how the student was receiving and understanding the feedback. Because as I do more and more work in this area, it becomes more and more critical. We can get teachers, and you know this, who can give brilliant feedback that is not understood. And my mantra is, was not heard or understood (unv.). When I look at one of your slides, it comes across that you at least (unv.) about was it heard, none of that is really checked whether the student understood it. And the only example of action, it was: do it again. But in the scoring, there are options for that. (E5)

Missing aspects

Three experts address aspects that they would add. Experts 1 and 3 propose that the use of paralinguistic means like the material could be further included in the evaluation sheet:

E: [...] In den Kriterienraster weiß ich jetzt nicht, ob es schon irgendwo untergebracht war, dass man diese 3D Perspektive hat, so zu sagen. Also das man die Materialien, die Arbeitsblätter einbindet, dass man auf konkrete Textstellen verweist, weil das findet ja immer in einem bestimmten Kontext statt, in dem Materialien sind, sei es jetzt hier in dem Vorortkontext oder online, und diese Materialien, die haben ja auch eine Rolle in dem Feedbackprozess. Ist das irgendwo drin? Und zum anderen halt dieses, was man halt sehen konnte, diese Beziehungsebene, sehen und hören konnte, wie jetzt auf die Person eingegangen wird und das sah man teilweise dann, wie sie sich da hinwenden mit dem Arbeitsblatt zu den Personen zum Beispiel auch oder natürlich, was schon berücksichtigt wurde, diese Beziehungsebene in dem sprachlichen. Aber auch dieses Materielle/ [...] (E3)

(E: [...] in the scoring instrument I do not know whether this is included that one has this 3D perspective, so to say. So that one includes the material, the task sheets, that one refers to specific aspects in the text, because this always occurs in a specific context in which there are materials, be it in the in-person context or online, und these materials, they have a role in the feedback process. Is this included somewhere? And what is more, this, what was evident, this relationship level, what one could see and hear, how one person is addressed and partially one saw that they turn with the task sheet towards the persons for example or, of course, what is considered, the relationship in the linguistic. But also this material/ [...])

Expert 3 states that the use and inclusion of the material in the conversation should be considered, as well as the positioning of the person via body language, being turned towards the learner. Also, expert 3 states that the relationship between the teacher and learner was a focus of their evaluation and this aspect is not included in the scoring instrument. Expert 5 proposes a stronger focus on elaboration of feedback, implemented throughout the scoring instrument.

Redundant aspects

Expert 1 states that while aspects are not redundant, some consider the same objective but approach it from different perspectives. The expert states, however, that none of the mentioned aspects should be neglected:

E: Gut, das sind ja teilweise vielleicht schon, aber es sind dann Aspekte, die aus unterschiedlichen Perspektiven angegangen werden. Von dem her wahrscheinlich, ah eher nicht. Also nicht jetzt, dass mir jetzt wirklich aufgefallen ist, dass was doppelt vorkommt. (E1)

(E: Good, yes, maybe be partially but then these are aspects that are regarded from different perspectives. Thus, probably, not really. So, I did not really notice something as a duplicate.)

Context of the RobS

Only expert 3 addresses the boundaries of the assessment format when discussing the scoring instrument. The expert discusses the notion that the scoring instrument is restricted to the simulation context, as it does not include the possibility to refer back to previous situations:

E: [...] Also ich finde die Skala die ist jetzt auch schon so darauf zugeschnitten, was für eine Art Kontext das jetzt ist: so eine einmalige Simulation ohne viel Rückbezugsmöglichkeiten, weil sonst wäre vielleicht dieser Rückbezugsmöglichkeiten da nochmal bisschen prominenter, ne, wenn man jetzt da hat, die ist im Klassenkontext, die Schülerin, hätte man halt sehr gut und da kann man natürlich viel stärker die Entwicklung aufzeigen. Aber jetzt mit dem Hintergrund, dass es jetzt sowas einmaliges ist, da kann ich das dann schon nachvollziehen, dass da jetzt weniger Punkte vergeben worden sind. (E3)

(E: [...] So I think the instrument is tailored to the type of context: a singular simulation without many possibilities for references to the past because otherwise, these references would be more prominent, when you now have the class context, the learner, one could much stronger indicate the development. But against the backdrop that this is singular, I can understand that there are less points provided for this.)

Expert 3 also expresses the wish that participants get the chance to reflect and argue their decisions as an additional part of the assessment, especially concerning the aspects of writing and diagnosis, a follow-up conversation could allow participants to explain their action.

In conclusion, experts have a generally positive attitude towards the scoring instrument and most items, but offer some options for modifications. Some experts emphasize specific criteria and items, like *Dialogue orientation* and *Securing understanding / content adaptation (Checking understanding)*. On an item-specific level, items *Error pattern* and *Amount of information (Balance and Amount of information)* are discussed most critically, while for most other items, rather minor specifications are addressed. On a general level, the comprehensiveness, and associated with this the issues of reliability and pragmatics, is criticized most. Moreover, the differentiated weighting is also regarded with skepticism. However, experts do not explicitly propose the exclusion of items for reasons of content. The results and specific modifications that result from this study will be discussed in the next section.

15.3.4 Discussion of the Results and Modifications to the Scoring Instrument

The results will be discussed regarding the research questions and the modifications for the scoring instrument. A discussion that relates to the test validity will be provided in the overall validity argument in Chapter 17.

First, results will be discussed regarding RQ8. The experts mentioned 24 different criteria, with only one criterion mentioned by all five experts and only eight by a relative majority of at least three experts. Thus, 18 criteria are mentioned by less than three experts. While the experts (mostly) agree on issues like text deep- and surface-level features, the necessity of specificity, a requirement of revision and a dialogic approach to feedback, several criteria are mentioned only by individual experts. For instance, expert 1 focuses on the use of the target language in the conversations as a main criterion, while expert 3 singularly focuses on the issue of hedging. Expert 4 puts a lot of emphasis on specificity, also with regard to meta-linguistic language use, and regards the initiation of self-evaluation negatively, while the results of part two indicate that other experts consider this rather positively. The qualitative comments thus indicate that there is a consensus on about 30 % of the criteria, while 70 % of the criteria are used by a minority of the experts.

When comparing the criteria that the experts mention regarding the framework and the scoring instrument, several similarities can be identified. Of the eight criteria mentioned by a majority of experts, seven are also included in the framework: text deep- and surface-level features, dialogue, revision, specificity, learning objectives and mentioning positive aspects (first) are explicitly addressed. Only the structure of feedback is not represented in the framework or scoring instrument. However, the reason to not include this aspect is discussed against the specific context of the RobS in detail in Chapter 7.4. Six other aspects mentioned by a minority of experts are also included in the framework: The

general notion of supporting the learner, checking learner's understanding, the participants' accuracy, use of material, feed forward and selecting errors. *Deficit orientation*, which the experts view negatively, is implicitly found in the framework within the notion of *Focused error selection* and *Target-orientation*, which requires participants to make choices concerning the extent to which errors are addressed. Moreover, *Positive feedback*, also takes into account whether the achieved goals are addressed. It is this criterion in which the emotional-motivational level is also most clearly included, although it has been consciously excluded from the framework, as discussed in Chapter 7.4.

Some aspects are rather indirectly addressed in the framework or scoring instrument or not included. Activating prior knowledge is only included in the scoring instrument to the extent that this might relate to cognitively demanding questions. It is not further considered in the framework or scoring instrument, as the constraints of the RobS generally hinder the addressing of learner history. Likewise, the criterion of learner progression, while inarguably relevant for any real-life feedback scenario, is excluded from the framework and scoring instrument. The notion of the target language is indirectly addressed in *Adaptation of language*, in which an initial use of the English language is addressed favorably. While one expert addresses self-evaluation, they do so negatively. Thus, it cannot be stated that there is agreement with the framework, as it explicitly takes the opposite view. The use of metalinguistic language is also not explicitly addressed, but it might be considered as part of *Detailedness*, which might require these technical terms for explanations, and the accuracy aspects, as they consider the correctness of these statements. Similarly, the contextualization of errors is also subsumed in *Detailedness* as it considers the extent to which information is provided, including that of a contextualizing nature. Not included are the balancing of hedged communication and the relevance of the final message to the learner. Concerning the latter, the structure of the conversation has generally been avoided in the framework due to the dynamic nature of the RobS with its time limitations (see Chapter 7.4).

Concerning RQ9, overall, most experts have a positive attitude to the approach taken in the scoring instrument. Experts 1 and 2 also explicitly acknowledge the theoretical foundation as positive. None of the experts explicitly expresses that specific criteria should be excluded and specific items are positively highlighted by some experts. Thus, the criteria that are used to assess performance in the scoring instrument seem to be appropriate. Expert 3 also discusses the scoring instrument in the light of the RobS, adding that it suits the context, but mentions that aspects outside the context are not included, like reflection of participants as to their reasons behind performance. Nonetheless, the general scope of the instrument is regarded as quite comprehensive and

possibly too complex for a reliable scoring. Similarly, the differentiated weighting of the items is discussed critically. Also, the experts provide more differentiated comments with regard to the operationalization of discrete items. Especially the operationalization of *Error pattern* and *Amount of information* is criticized. Additionally, experts propose that paralinguistic criteria could be added to the scoring instrument, providing a more holistic evaluation of the performance. The modifications that are subsequently made to the scoring instrument will be presented in the following.

Modifications

After the expert interviews, several modifications were made to the scoring instrument. Most significantly, the complexity is reduced with regard to the range and weighting of points, as numerous experts, especially expert 2 and 4, argue that the comprehensiveness of the scoring instrument might interfere with reliability and pragmatics. By reducing the maximum points from three to two, the scoring options are reduced for all tiered criteria. The decision was made to eliminate the top tier, which was designed as an excellence level and used qualifiers such as 'always'. This step is considered appropriate as the RobS is designed for PSTs who are still in the progress of acquiring the necessary competence. The new maximum level of two points thus uses qualifiers such as 'often', allowing for some room for improvement on the participants' side. With regard to weighting, the decision was made not to emphasize specific criteria and to weigh all criteria equally. As expert 2 stated, there might be little empirical evidence to argue these weightings in terms of learner achievement. Also, this contributes to an overall reduction of complexity.

Several items were also modified based on the results. First, those related to feedback aims are discussed. *Positive feedback* is made more specific so that it refers to the task and does not deflect from the focus of improving performance. *Learning objectives* are included as is, but the condition of providing feedback against learning objectives at least once is added for *Quality of Feed back*. *Feed Back* and *Feed Forward* are excluded as they did not account for the number of instances in which qualitative feed back or feed forward occurred, as suggested by expert 2 and are, in addition, also operationalized in the quality-oriented criteria. Similarly, *Task* is excluded, as experts did not address it and it is represented in other items as discussed in Chapter 7.3. *Self-regulation* is modified to more closely define the quality of the self-regulative feedback.

Several substantial changes were made for criteria within the dimension of feedback strategies. While *Dialogue orientation* is positively evaluated by four experts, and also used as a criterion in the assessment of the performances, expert 4 is skeptical as it does not specify what the dialogue is about. Thus, a

condition is added to ensure that at least once, the dialogue is cognitively demanding for the learner. For *Securing understanding / content adaptation (Checking understanding)*, the term 'answers' is exchanged with 'alternative explanations' as suggested by expert 4. For *Linguistic adaptivity (Adaptation of language)*, the complexity of the different tiers is reduced to provide more general and accessible descriptions. Similarly, *Adaptive error correction* is simplified and the quality of the intermediary step (i.e., implicit/explicit and indirect/direct) removed so that only the final step of explicit feedback remains. The item *Appropriate amount of information (Amount and balance of information)* is modified more substantially. First, based on the criticism by expert 2, a discrete range of three to five aspects is added as an orientation for an appropriate amount of information. Moreover, it is added that in addition to the amount, the information has to be balanced between surface- and deep-level features. This modification is made as expert 3 addresses the fact that the kind of information should also be considered and expert 5 explicitly states that the balance of information is important. The item is now operationalized as tiered, with zero points scored if the amount is inappropriate, one point being scored if the amount is appropriate but the balance is off, and two points being scored if the kind of information is also balanced. A specification is added to *Process orientation (Revision)* that addresses that the revision can also occur in the subsequent lessons, which is indicated in the material. The operationalization for *Error pattern* is modified substantially. Taking into account the criticism levelled by expert 2 and 4, rewarding an exclusive focus on error patterns with the most points is avoided. In order to measure the ability to identify and provide feedback on error patterns separately, it is reconceptualized as a dichotomous item with two points being scored if error patterns are focused on at least once in the conversation, as suggested by expert 4. Also, the notions of error patterns are more explicitly included in the item description of *Target-oriented error selection* as proposed by expert 2 and 4.

Some changes were also made to the criteria in the dimension of feedback topic. Both items *Language adequacy (Surface: Language appropriateness)* and *Structure (Deep: Structure)* are described in more detail so that the notion of genre is included more clearly, for the former on a micro-, and for the latter on a macro-level. *Detailedness* is also amended by specifications as to what kind of elaboration is considered in this context, e.g., explanations. The criteria *Linguistic accuracy (Language accuracy)* and *Pedagogical accuracy (Professional accuracy)* are modified similarly. The item descriptions are adapted to a consistent frequency scale as proposed by expert 4. While expert 1 and 2 also consider a combination of both criteria, it was decided to keep them separate, but to more clearly define the range of features that fall within the category.

No criteria were excluded, as no expert strongly recommended that a criterion is not relevant to cover the competence holistically. Even with the more strongly criticized criteria of *Error pattern* and *Appropriate amount of information* (*Amount and balance of information*), the issue is rather with the concrete operationalization than with the criterion itself. Two items are added based on the expert judgements. The criterion *Use of material* is added to account for proposals made by expert 1 and 3. Both experts advocated for the inclusion of paralinguistic aspects and how the material is used in the RobS. Also, the notion of *Body language* is added as another facet of paralinguistic phenomena and dialogue orientation. It is also added to the framework after the expert interviews. It is thus conceptualized to describe the degree to which the participant's posture, facial expressions and eye contact show attentiveness that contributes to that dialogue.

15.3.5 Limitations

Limitations of the study and consequently the results need to be addressed. First and foremost, this study only considers the perspective of researchers in the field as experts. This increases the likelihood that the evaluation of the performances and the scoring instrument is based on scientific expertise, generating the required evidence to discuss the respective claims. However, this study does not include the perspective of pure practitioners, e.g., teachers and teacher educators from the second phase. Their expertise is different in nature, generating evidence that can be used to discuss other claims that might also contribute to validity, but not less valuable. Similarly, the perspective of learners is not included here, although there is an argument that learners might be well suited to argue whether the feedback that is provided is actually understood. Potentially, future research might focus on their evaluative perspective of the performances and the scoring instrument.

Another limitation is represented by the sample, as, arguably, only a relatively small number of experts participated. This might decrease the generalizability of the findings. However, the experts that participated were asked based on their experience and qualification in the field. Thus, the quality of their statements is likely high, compensating for a potential lack in quantity. Although the overall number of potential experts is limited due to the narrowly defined field, the sample can be considered sufficient for the purpose. Still, the modifications that were made might be specific to the evaluation of the specific sample.

Furthermore, technical issues need to be addressed. Experts 2 and 3 stated that the audio quality was not optimal, resulting in difficulties with the comprehension of the participants' statements. Similarly, although the pilot with native speakers indicated sufficient comprehensibility, experts 4 and 5 stated that the

subtitles were not easily accessible. Especially expert 4 stated that the translation was, at times, not understandable. Additionally, experts 4 and 5 did not receive an audio transmission. Their assessment is thus not fully comparable to that of experts 1, 2 and 3. While the criteria used to assess these performances and the evaluation of the scoring instrument is not strongly affected by these circumstances, the actual assessment of the videos, which is discussed in Chapter 15.5.4.4, has to be reflected against these challenges.

15.3.6 Summary

The aim of the study presented in this chapter is to collect data from experts in order to first, evaluate the current state of the scoring instrument and second, modify the scoring instrument accordingly. For that purpose, expert interviews are conducted. In this study, experts are defined as researchers that have experience in the adjacent fields of the RobS, i.e., feedback, L2/FL writing, feedback on writing or feedback in the EFL classroom. Five experts agreed to participate and interviews were conducted in a two-tiered approach. In the first part, experts are asked to watch four performances and provide both a global, quantitative rating as well as to reason their assessment in an evaluative comment. In the second part, experts are asked to evaluate the preliminary scoring instrument. The focus of the analyses is on the qualitative data, as the quantitative data will be further focused on in Chapter 15.5.4.4. Qualitative data is analyzed using approaches of qualitative content analysis according to Mayring (2022).

The data of the first part is analyzed with regard to the RQ8, aiming to identify criteria that experts use for assessment and compare them to the framework and scoring instrument. Findings indicate that the experts use several different criteria to assess the performances. Experts mainly address three aspects: the content of the feedback conversation, the general approach participants take (i.e., monologic vs dialogic) and the structure. Interestingly, of the 24 criteria mentioned by the experts overall, only eight are mentioned by a majority of them. This indicates that the experts only partially apply the same criteria in their assessment of the RobS. Comparing the results to the framework and the scoring instruments, it can generally be stated that many of the experts' criteria can be found, to a certain extent, in them. While some criteria are explicitly mentioned by both the experts and the framework, others are arguably implicitly or indirectly addressed. Only a few criteria, mostly those that address the structure, are not included, which is a conscious decision, as outlined in Chapter 7.4. The second part of the study aims at structuring experts' evaluations of the preliminary scoring instrument in order to collect information with regard to its appropriateness and options for modifications. Generally, while there is a positive reception of the scoring instrument as a whole, the comprehensiveness is seen as a

challenge for scoring. However, none of the experts advocates for the exclusion of items. Contrarily, some experts propose the inclusion of paralinguistic criteria. Experts are most critical of the items *Error pattern* and *Amount of information (Balance and Amount of information)* as their operationalization is not considered to be appropriate.

Based on these findings, several modifications were made. The complexity of the instrument is reduced not by excluding criteria, but by decreasing the maximum score for each item to two points and weighing all criteria equally. The critically evaluated items are modified according to the experts' suggestions and items are included for the paralinguistic dimension of the feedback conversation. Limitations of the study relate to the singular perspective of experts in this study, as well as technical and methodological constraints of this study with regard to the quality and inclusion of the audio information.

15.4 Study 3: Scoring Reliability

In order to investigate the reliability of the scoring instrument, a two-tiered approach is taken. First, the individual items are focused on in order to create a set of items which are reliable in terms of the internal consistency as well as inter-rater reliability. Therefore, an exploratory approach was taken that first focused on the preliminary set of 27 items. Based on the results, a set of items is finalized. With the final set of items, revised reliability coefficients are presented. In the second part, investigations into the reliability of the test score are made. This separation is necessary as the test score is generated by an aggregation of item scores. Thus, the final set of items has to be determined before a further investigation of test scores can be made. Using the final test score, the additional test-related factor of the actors' influence can be discussed. The findings of this study can be used to discuss the following claims and research questions:

Claim 4: Item score

In order to assess the performance with finalized set of items, they need to be applicable for the test format and be generated reliably. In order to generate evidence to discuss this claim the frequency of item scores as well as the consistency with which items are applied to performances by different raters need to be considered.

RQ10: To what extent can the scoring instrument generate reliable item scores?

Claim 5: Test score

The test score represents the participants' levels of performance if only relevant items are used to aggregate the test score and the scale is internally consistent. The generated test scores should depict the universe of potential scores and

allow for the differentiation of participants. Moreover, the test score should be generated reliably by different raters.

RQ11: To what extent can the scoring instrument generate reliable test scores?

Claim 3: Performance

Within the course of the study, three different actors portrayed the SL. As the actors are a part of the RobS, they should ideally not cause any variance in the generated test scores and allow all participants to perform equally.

RQ12: To what extent do the actors influence the generated test scores?

15.4.1 Scoring Reliability on the Item Level

The purpose of this study is twofold. First, the number of items of the preliminary scale is potentially too high for practical use. Thus, items that do not have a sufficient ability to differentiate between performances or display insufficient reliability coefficients can be excluded. Also, the rating manual can and should be improved based on results of the inter-rater reliability investigation. Potential rating challenges can thus be mitigated by modifying instructions or clarifying item descriptions. Second, this study does provide information about the extent to which the final scale is reliable – both in terms of the internal consistency and the application by different raters.

15.4.1.1 Rating of the Full Sample on the Item Level

In order to identify items that are not sufficiently discriminating between performance levels, the rating results of the full sample ($N = 51$) are focused on. The methodological procedures, including the rating process and statistical guidelines, will be touched upon first before descriptive results of item frequencies and co-occurrences as well as internal consistencies are presented.

15.4.1.1.1 Methods

Rating process

First, the video-recorded RobS were rated based on the scoring instrument. As described in the instructional manual for scoring, each video was watched at least twice with an optional third time. In the first viewing, the focus was put on the event-based and dichotomous criteria. In the second viewing, the focus was put on the frequency-based measures. The third viewing was used to review uncertain decisions. The results of the rating are documented twofold. The numeric results were entered into an Excel spreadsheet, which includes checkboxes for each item. Each performance's rating was documented in an individual tab and the results of each item, as well as the sum, transferred into an overview table. This table was transferred to the statistics program SPSS (Version 29.0.0.0)

for further calculations. Additionally, the data is also coded qualitatively. This step was taken in order to improve subsequent reconstruction of the rating. With regard to the non-event-based measures that require an estimation of quality of frequency, brief comments were made in the same excel file if the maximum number of points was not achieved. Event-based codes, but also specific events, which are the foundation for the evaluation of non-event-based measures, are also documented in MAXQDA (Standard 2020). The use of MAXQDA allows for the coding, comparison and analysis of the data in both the video and transcription format as not all video-documented performances are transcribed. Segments were coded according to the descriptions in the rating manual. The focus was on documenting at least one clear event so that scoring could be traced back. Consequently, while potentially multiple events could be coded for a single item in one performance, it is usually only coded once. This was done because the event only needed to occur once for the purpose of rating. As the aim was not to provide an extensive overview of occurrences in the videos, it was refrained from coding the performances in such detail. The rating and coding process consisted of two steps. First, rater 1 rated and coded all $n = 51$ videos chronologically. No more than four videos were rated and coded in a single day and breaks of at least 15 minutes were taken between videos to limit biases based on previous performances. After all videos were evaluated, each item and its corresponding coded segments were revisited and evaluated for all potential 51 cases to ensure that the individual items were rated consistently. If necessary, the rating and coding was adapted. The results of rater 1 are the basis for the descriptive statistics of the whole sample and used for further analyses, including Chapter 15.5.

Descriptive statistics

First, descriptive analyses are conducted on the item level with the aim to reduce the overall number of items. Specific items that occur relatively rarely or frequently in the sample are possibly not differentiating enough for the test score and bear potential for exclusion. Items which are scored with two points extremely frequently (and consequently rarely scored with zero points) indicate that the behavior described by the item is a given for the participants and is not necessarily suitable for discriminating between performances. Conversely, the inclusion of items that are concerned with rarely occurring events also needs to be discussed. Such items are potentially not relevant for the sampled performances and also do not necessarily contribute to the discriminatory requirements. Exclusion of items with (in-)frequently occurring scores could potentially facilitate the rating process considering the overall number of items. However,

exclusion should be argued against the relevance of the item for the construct as not to impede the construct validity (Döring, 2023, p. 473).

In addition, it is explored to what extent items co-occur. Items that frequently occur together in a segment of a performance potentially measure the same facet of the construct. The focus of the exploration is on items that are similar in their conceptualization. A comprehensive approach in which co-occurrences of all items are considered is not expedient, as it potentially creates results that are coincidental and not logically justifiable. Items that are considered are

- 1) "Process" and "Deep: Process", as both consider the creation of texts;
- 2) "Process" and "Self-Regulation" as both focus on feedback levels that extend on the task at hand;
- 3) "Surface: Language correctness" and "Surface: Language appropriateness", as both focus on surface-level features;
- 4) "Deep: Structure", "Deep: Content" and "Deep: Process" as those focus on deep-level features;
- 5) "Checking understanding", "Dialogue orientation" and "Adaptive error correction" as those have a communicative focus.

Cross tabulations of these items indicate simultaneous occurrence of these items in the sample and allow for an examination of co-occurrence within a single performance. Additionally, the qualitative analysis of the coded segments can be drawn upon to indicate if there is actual overlap. As not all instances of an event are however documented in MAXQDA, these results only allow for a tentative indication.

Internal consistency

For measurement of internal consistency, Cronbach's α is calculated based on the rating of rater 1. This coefficient considers the average inter-item correlations, assuming all items aim at measuring one construct (Döring, 2023, p. 274). A value of $\alpha > .80$ is considered acceptable (Döring, 2023, p. 274; Streiner, 2003, p. 103). α -values $> .90$ are increasingly questioned, as these could be indicators for redundant items in a scale (Streiner, 2003, p. 103). In order to potentially improve the internal consistency, item-total statistics are considered in order to deliberate the exclusion of items. Negatively correlating items are potentially harmful for the overall reliability of the scale and should be excluded. A threshold of .30 is indicated as desirable for all items. However, items should not be eliminated systematically based on that score, rather their exclusion should be reasoned from a content-perspective (Döring, 2023, p. 473). Especially items in tests that seek to measure a formative construct, such as domain knowledge "do not have to be highly correlated" (M. Stadler et al., 2021, p. 2). Such constructs are often constituted, i.e., formed, of several distinct disciplines that do not have

to show a large extent of correlation (M. Stadler et al., 2021, p. 2). Arguably, the construct measured in the RobS qualifies as such because it considers behavior holistically and includes multiple perspectives. Consequently, lower levels of correlation of individual items are to be expected (Döring, 2023, p. 473).

15.4.1.1.2 Results

In the following, the item-specific reliability result will be presented, based on the rating of rater 1. First, the item frequencies are focused on. Second, specific instances of co-occurrences are addressed. Lastly, the internal consistency and the corrected-item correlations of the items are described.

Item frequencies

Table 33 shows the item-specific means, standard deviations and the distribution of cases for each item¹²⁵. Looking at all the items, each possible item score has been used at least once in the sample. For the dichotomous items, the least frequently used item scores (< 10 %) are zero points for *Use of material* and *Surface: Language accuracy* ($n = 2$), *Surface: Language appropriateness* ($n = 3$), *Self* ($n = 4$), and two points for *Deep: Process* ($n = 5$). The most frequently used item scores (> 90 %) are the respective opposites. For the three-tiered items, the least frequently used item scores (< 10 %) are zero points for *Detailedness* ($n = 1$) and *Quality of Feed Forward* ($n = 5$), and one point for *Adaptation of language* ($n = 4$). The most frequently used item score for the three-tiered scores are two points for *Detailedness* ($n = 42$) and no other item score was used more than 80 %. This trend is naturally also reflected in the means for the items. For the dichotomous items, the lowest means (< .20) is displayed by the item *Deep: Process* ($M = 0.20$). The highest means (> 1.80) are evident for the items *Use of material* ($M = 1.92$), *Surface: Language accuracy* ($M = 1.92$), *Surface: Language appropriateness* ($M = 1.88$), and *Self* ($M = 1.84$). For the three-tiered items, the lowest means (< 1.00) are *Checking understanding* ($M = 0.69$), *Adaptive error correction* ($M = 0.71$) and *Amount and balance of information* ($M = 0.98$). The item with the highest mean (> 1.80) is *Detailedness* ($M = 1.80$).

¹²⁵ Items with an asterisk (*) are not included in the final scale.

Table 33: Item-specific means, standard deviations and distribution of points for each item.

Item	<i>M</i>	<i>SD</i>	0		1		2	
			<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Self-evaluation	0.39	0.80	41	80.4			10	19.6
Positive feedback	1.61	0.80	10	19.6			41	80.4
Learning objectives (Feed Up)	1.61	0.80	10	19.6			41	80.4
Process	0.75	0.98	32	62.7			19	37.3
Self-regulation	0.78	0.99	31	60.8			20	39.2
Use of material*	1.92	0.39	2	3.9			49	96.1
Prioritization of deep-level features	0.94	1.01	27	52.9			24	47.1
Error patterns	1.41	0.92	15	29.4			36	70.6
Surface: Language correctness	1.92	0.39	2	3.9			49	96.1
Surface: Language appropriateness	1.88	0.48	3	5.9			48	94.1
Deep: Structure	1.49	0.88	13	25.5			38	74.5
Deep: Content	1.69	0.74	8	15.7			43	84.3
Deep: Process*	0.20	0.60	46	90.2			5	9.8
Self	1.84	0.54	4	7.8			47	92.2
Body language	1.76	0.65	6	11.8			45	88.2
Adaptation of language*	1.25	0.94	17	33.3	4	7.8	30	58.8
Revision	1.31	0.76	9	17.6	17	33.3	25	49.0
Adaptive error correction	0.71	0.83	27	52.9	12	23.5	12	23.5
Dialogue orientation	1.20	0.90	16	31.4	9	17.6	26	51.0
Quality of Feed Back	1.22	0.76	10	19.6	20	39.2	21	41.2
Quality of Feed Forward	1.37	0.66	5	9.8	22	43.1	24	47.1
Checking understanding	0.69	0.68	22	43.1	23	45.1	6	11.8
Amount and balance of information	0.98	0.93	22	43.1	8	15.7	21	41.2
Target-oriented error selection	1.27	0.78	10	19.6	17	33.3	24	47.1
Detailedness	1.80	0.45	1	2.0	8	15.7	42	82.4
Language accuracy	1.29	0.70	7	13.7	22	43.1	22	43.1
Professional accuracy	1.29	0.73	8	15.7	20	39.2	23	45.1

Item Co-occurrences

First, co-occurrences in the rating of the event-based items are presented. Cross tabulations indicate the occurrence of item scores within the same performance. The actual overlap of the coded segment can potentially be distinctly located in the performance through the code-relations-map in MAXQDA (see Figure 24).

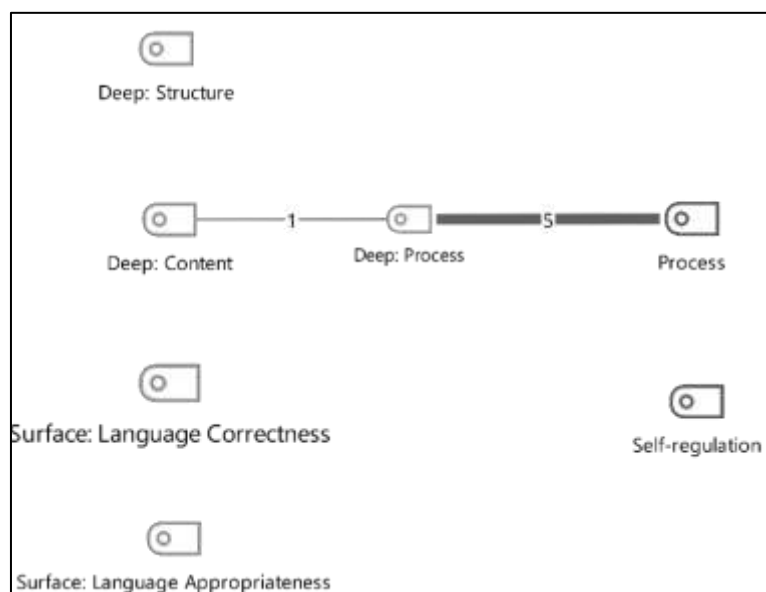


Figure 24: Code overlap of relevant event-based items.

As Figure 24 shows, only the item *Deep: Process* shows overlap with their coded segments, in one case with *Deep: Content* and in another with *Process*.

The item *Deep: Process* is scored with two points in $n = 5$ performances, all of which occur in performances in which the item *Process* is also scored with two points (see Table 34). Results in Figure 24 indicate that all five segments that are coded as two points for *Deep: Process* are also coded as two points for *Process*. One example is as follows:

[0:05:24] TP: und das fällt etwas schwierig weil das sind auch sachen die man nicht vielleicht immer unbedingt in den texten verwendet aber du warst ja so mutig und hast es da wirklich getan und probiert damit zu arbeiten also da auf jeden fall gucken. (GBP25)¹²⁶

The participant provides information on the process of writing in the form of feed back, e.g., how the learner is going, as is required for *Deep: Process*. This extends to the coding of *Process*, which encompasses, feed up, feed forwards and feed back on the process, highlighting the fact that the codes overlap.

Table 34: Cross tabulation *Deep: Process* - *Process*.

	Process			
		0	2	Total
Deep: Process	0	32	14	46
	2	0	5	5
	Total	32	19	51

¹²⁶ Quotes from the simulation are not translated but paraphrased.

In $n = 11$ performances, two points are scored for both *Process* and *Self-regulation* (see Table 35). Two points for either of the two items are scored in $n = 8$ performances for *Process* and in $n = 9$ performances for *Self-regulation*. As is indicated in Figure 24, these codes do not seem to overlap.

Table 35: Cross tabulation *Self-regulation* - *Process*.

	Process			
Self-regulation		0	2	Total
	0	23	8	31
	2	9	11	20
	Total	32	19	51

Two points are scored for both *Surface: Language appropriateness* and *Surface: Language correctness* in $n = 47$ performances, while in one performance only *Surface: Language appropriateness*, and in two performances only *Surface: Language correctness* is scored (see Table 36). As is indicated in Figure 24, the coded segments do not seem to overlap in the performances.

Table 36: Cross tabulation *Surface: Language correctness* - *Surface: Language appropriateness*.

	Surface: Language appropriateness			
Surface: Language correctness		0	2	Total
	0	1	1	2
	2	2	47	49
	Total	3	48	51

In $n = 32$ performances, two points are coded for both *Deep: Content* and *Deep: Structure*, while in $n = 11$ performances only *Deep: Content*, and in $n = 6$ performances only *Deep: Structure* receive two points (see Table 37). As is indicated in Figure 24, these codes do not seem to overlap.

Table 37: Cross tabulation *Deep: Structure* - *Deep: Content*.

	Deep: Content			
Deep: Structure		0	2	Total
	0	2	11	13
	2	6	32	38
	Total	8	43	51

The item *Deep: Process* is scored with two points in $n = 5$ performances, all of which co-occurred in performances in which two points for the item *Deep: Content* were also scored (see Table 38). Figure 24 shows that in at least $n = 1$ instance, the codes overlap. This instance is found in AEM23, in which the participant provides feed back on the inclusion of facts from the task sheet and how

the learner should paraphrase the statements for their own news report in the future [0:05:37 – 0:06:06].

Table 38: Cross tabulation *Deep: Process* - *Deep: Content*.

Deep: Process	Deep: Content			
		0	2	Total
	0	8	38	46
	2	0	5	5
	Total	8	43	51

In $n = 4$ performances, two points are scored for both *Deep: Process* and *Deep: Structure*, while in $n = 1$ performances only *Deep: Process*, and in $n = 34$ performances only *Deep: Structure* was coded (see Table 39). As is indicated in Figure 24, these codes do not seem to overlap.

Table 39: Cross tabulation *Deep: Process* - *Deep: Structure*.

Deep: Process	Deep: Structure			
		0	2	Total
	0	12	34	46
	2	1	4	5
	Total	13	38	51

For investigations into co-occurrences that include frequency-based items, only cross tabulations are considered as the code refers to the whole performance and localizations are not possible. The scores for *Checking understanding* and *Dialogue orientation* are identical in $n = 10$ performances for zero points; in $n = 2$ performances for one point and in $n = 5$ performances for two points, which leads to a total overlap of scores in $n = 17$ performances (see Table 40). Generally speaking, the item *Dialogue orientation* is scored more often with two points ($n = 26$) than *Checking understanding* ($n = 6$), while it is the reverse for the scores of zero points or one point. In $n = 6$ instances, *Dialogue orientation* was scored with zero points while *Checking understanding* was scored with one point or two points.

Table 40: Cross tabulation *Dialogue orientation* - *Checking understanding*.

Dialogue orientation	Checking understanding				
		0	1	2	Total
	0	10	5	1	16
	1	7	2	0	9
	2	5	16	5	26
	Total	22	23	6	51

The scores for the items *Adaptive error correction* and *Dialogue orientation* are identical in $n = 16$ performances for zero points which encompasses all instances in which the code was given for the latter item (see Table 41). In $n = 2$ performances there are identical scores for one point, and in $n = 12$ performances the scores are identical for two points, which includes all instances in which the score is given for *Adaptive error correction*. There is a total of identical scores in $n = 30$ performances. Generally, there is no instance in which *Adaptive error correction* is scored with one or two points and a score worth less is given for *Dialogue orientation*.

Table 41: Cross tabulation *Dialogue orientation* - *Adaptive error correction*.

Dialogue orientation	Adaptive error correction				
		0	1	2	Total
	0	16	0	0	16
	1	7	2	0	9
	2	4	10	12	26
	Total	27	12	12	51

The items *Adaptive error correction* and *Checking understanding* have identical scores in $n = 16$ performances for zero points; in $n = 7$ performances for one point and in $n = 3$ performances two points (see Table 42). There is a total of identical scores in $n = 26$ performances. Generally, there are code overlaps for each possible combination.

Table 42: Cross tabulation *Checking understanding* - *Adaptive error correction*.

Checking understanding	Adaptive error correction				
		0	1	2	Total
	0	16	4	2	22
	1	9	7	7	23
	2	2	1	3	6
	Total	27	12	12	51

Internal consistency

In order to evaluate the internal consistency of the scoring instrument, Cronbach's alpha is calculated. For the 27 items in the instrument, α is = .79, which is slightly below the acceptable value of .80 (Döring, 2023, p. 274; Streiner, 2003, p. 103). The corrected item-total correlation is depicted in Table 43, the complete inter-item correlation matrix is found in Appendix H.3.

Table 43: Statistics for internal consistency of the 27-item scale.

Item	Corrected item-total correlation	Cronbach's α if the item is deleted
Self-evaluation	.31	.78
Positive feedback	.35	.77
Learning objectives (Feed Up)	.55	.77
Process	.43	.77
Self-regulation	.35	.78
Use of material*	-.08	.79
Prioritization of deep-level features	.28	.78
Error patterns	.31	.78
Surface: Language correctness	.20	.78
Surface: Language appropriateness	.17	.78
Deep: Structure	.25	.78
Deep: Content	.30	.78
Deep: Process*	.20	.78
Self	.15	.79
Body language	.35	.78
Adaptation of language*	-.17	.81
Revision	.15	.79
Adaptive error correction	.45	.77
Dialogue orientation	.32	.78
Quality of Feed Back	.48	.77
Quality of Feed Forward	.56	.77
Checking understanding	.38	.78
Amount and balance of information	.43	.77
Target-oriented error selection	.44	.77
Detailedness	.40	.78
Language accuracy	.20	.78
Professional accuracy	.45	.77

The two items that correlate negatively are *Use of material* ($\rho = -.08$) and *Adaptation of language* ($\rho = -.17$). There are nine items that correlate below a desired threshold of $\rho = .30$: *Prioritization of deep-level features*, *Surface: Language correctness*, *Surface: Language appropriateness*, *Deep: Structure*, *Deep: Process*, *Self*, *Revision* and *Language accuracy*. This includes all five criteria that focus on the elements that could be diagnosed in the text (Becker-Mrotzek, 2014). The

deletion of *Adaptation of language* alone would result in an increased α -value of .81, indicating a sufficient internal consistency.

15.4.1.2 Inter-Rater Reliability on the Item Level

In this chapter, the results of the inter-rater reliability investigations are presented on the item level. Firstly, the rater training, the rated subsample and the rating procedure is described. Secondly, results of rating I are described and the modifications to the rating manual presented. Lastly, results of the subsequent rating II are described.

15.4.1.2.1 Methods

The objective of the inter-rating investigations is twofold. First, the degree to which the items are applied consistently by different raters is examined. Second, based on these findings, the instructions for the rating process should be modified to increase reliability. In order to investigate inter-rater reliability, two further independent raters were tasked with rating a portion of the data. These two raters are student assistants who, at the time, were students of English. Rater 2 was enrolled in the Bachelor of Education program and rater 3 in the Master of Education program. Rater 3 was involved in the project in multiple roles as she portrayed the SL in several RobS and also transcribed multiple RobS, thus having had more experience within the project overall.

Rater training

Both raters received an extensive introduction and training with regard to both rating and coding (see Figure 25 for an overview)¹²⁷.

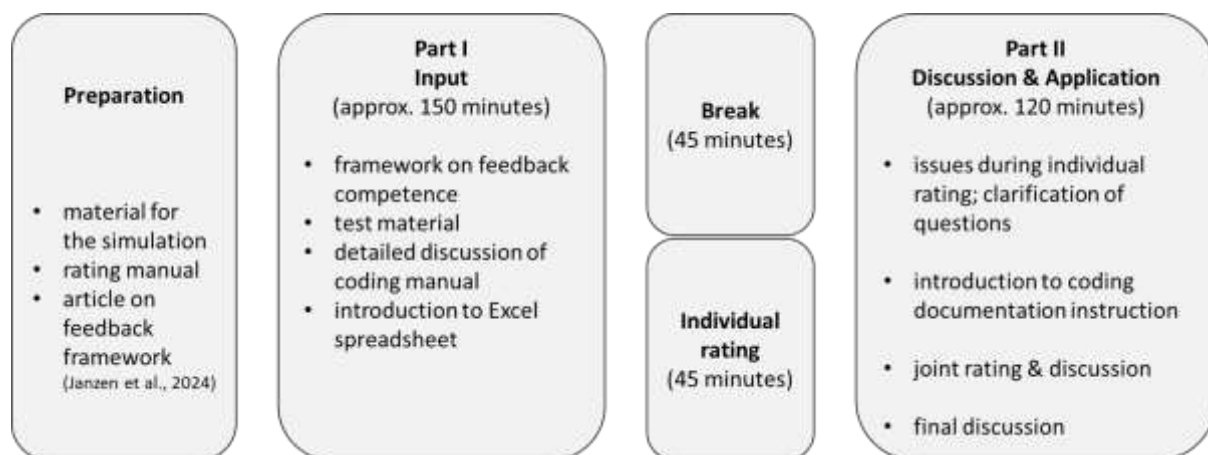


Figure 25: Overview of the rater training.

¹²⁷ Each rater received an individual, but identical, training. Training of rater 2 was done in-person; training of rater 3 digitally, via Zoom.

In order to prepare for the training, each rater received the material for the test, a research paper that introduces the framework for feedback (Janzen et al., 2024), and a preliminary version of the rating manual (see Appendix B for the final version)¹²⁸. With regard to the test material, the raters were asked to identify potential sources for a feedback conversation in the learner text M1 so that they are able to anticipate participants' comments in the performances. The training itself was done in one day and took about six hours in total. The first part took approximately two and a half hours and was input-focused. First, the framework of feedback was briefly reviewed (see Chapter 7). Second, the learner text and the material were discussed and results of the intercoding process of the text presented and introduced as a source for coding and rating (see Chapter 14.1.1). Thirdly, the rating manual was introduced - each item was presented and explained, and questions discussed. Lastly, raters were introduced to the Excel-spreadsheet used to rate the videos. Following the input, raters were given a 45 minutes break and another 45 minutes to individually rate a first video in order to become familiar with both the rating manual and the design of the rating spreadsheet. The video was taken from the sample of the pilot survey that had the same learner text. The second part was focused on discussion and application. Firstly, potential issues with the Excel-spreadsheet and the rating manual were discussed with regard to that first video. Secondly, raters were introduced to the coding file for qualitative documentation and the corresponding instructions. Similar to the rating process described above, the raters were asked to document the event-based codes and provide brief comments for other items so that a certain degree of transparency and traceability was ensured. Other than rater 1, raters 2 and 3 were asked to enter either the coded segments or time stamps, depending on whether a transcript was available for the performance, and comments for the individual item in a separate Excel file. This was done, as it allows for an easier comparison for the consultation sessions after the rating. Thirdly, another video from the pilot survey was rated and discussed jointly. The raters were asked to stop the video when they thought to have identified an event from the rating manual, enter the rating in the spreadsheet and document the code, as instructed. Throughout, questions were discussed regarding the rating manual.

Rating sample

Following the training, a set of identical $n = 10$ videos of the main survey sample is rated by each rater. A portion of about 20 %, i.e., ten videos, and not the full

¹²⁸ Participants' codes were anonymized in raters' rating manual as to not influence their rating based on the examples.

sample, is rated, as is conventionally accepted (Döring, 2023, p. 557). Similar studies focusing on the rating of feedback situations had a similar amount or share of inter-rated material (Holstein et al., 2022, p. 5; Kürzinger & Pohlmann-Rother, 2019, p. 390). For the purpose of investigating the reliability of the items and potentially modify the rating instructions, this sample size is thus considered sufficient. The videos for the inter-rating process are selected based on several aspects so that the subsample represents the whole sample in certain aspects. Firstly, the videos needed to depict a variety of different performance levels, thus the test scores needed to show the range of the sample. Secondly, the entire variety of possible ratings for an item should be included. This includes the rating of both 0 and 2 points for the dichotomous items, as well as the rating of 0, 1 and 2 points for the three-tiered items. Therefore, the selected videos needed to cover all potential ratings, using rater 1 as a benchmark. Thirdly, the sample should include a relatively equal share of participants with regard to their gender as well as phase of their study. Fourthly, the videos should include a sample of all three actors, and as many of the used videos as possible should have a transcript so that documentation and reporting is facilitated. The rating sample consisted of the following ten videos (see Table 44).

Table 44: Description of the inter-rating subsample.

TP	Test score ¹²⁹	Actor	Phase of study	Gender	Transcript
AEM23	30	3	Bachelor	Female	No
CCF27	17	1	Bachelor	Female	Yes
KMP05	26	3	Bachelor	Male	Yes
MMM07	42	2	Master	Male	Yes
MMK21	27	2	Bachelor	Male	Yes
CHS10	21	1	Bachelor	Male	Yes
GJS22	34	2	Bachelor	Female	No
EME03	47	3	Master	Female	No ¹³⁰
EEH01	39	3	Master	Female	No
MJO28	32	2	Master	Male	Yes

Inter-rating procedure

The inter-rating procedure consisted of several steps (see Figure 26). Raters were first asked to rate four videos that had a transcript. While the focus of the rating

¹²⁹ The selection was made based on the results of the test score that included 27 items, the test scores presented here are those based on the final set of 24 items.

¹³⁰ The performance was transcribed after the inter-rating procedure was completed.

was the video, the transcripts could be used for the rating, if necessary. After four videos were rated, a first individual consultation took place (Consultation I). Raters were able to ask questions on the rating manual, and specific codes that indicated that further clarification was necessary were also discussed. Raters then reevaluated those four videos and rated the other videos as well for the first full rating (Rating I).



Figure 26: Overview of the inter-rating procedure.

In order to investigate inconsistencies between raters 1 and 2 as well as raters 1 and 3, another individual consultation session took place (Consultation II). In these consultations, each individual discrepant rating was discussed by comparing the ratings to the descriptions in the rating manual and notes taken to make adjustments to the coding manual. The focus was mostly on discrepancies that included the rating of zero points, as this indicates that the described event did not occur while discrepancies between one point and two points are assumed to be of rather inferential nature. Based on these findings, the rating manual was modified in order to improve clarity. Subsequently, raters re-evaluated their rating based on these discussions for a final rating (Rating II). In a final step, a trilateral consultation was scheduled. In it, final modifications of the rating manual were presented, discussed and specific formulations agreed upon.

Methodologically, this iterative approach is also reminiscent of design-based research (DBR). In DBR approaches such as Educational Design Research (EDR), several cycles of “investigation, development, testing, and refinement” (McKenney & Reeves, 2019, p. 15) are used in the research process. Similarly, the inter-rating process here draws upon this idea and seeks to refine the rating manual based on the rating results. Consequently, results of reliability are expected to change between the first full rating and final rating, as the results of the first rating are used to improve the rating manual and the reliability scores. This iterative cycle is also described by Döring, specifically for test or scoring instruments (2023, p. 558). She argues that discrepancies in categories with insufficient reliability coefficients must be discussed with the raters, the instrument be modified

accordingly and the changes documented. Following, the revised instrument should be tested again and, potentially, modified again, if reliability is still lacking. However, it should be cautioned not to create an instrument that fits only to a certain sample.

To discuss the inter-rater reliability of specific items, the data is primarily viewed as nominal, which requires the use of the agreement coefficient Kappa (κ). Conversely to the simple calculation of agreement in percent, Kappa coefficients are corrected to account for random agreement (Döring, 2023, p. 559). For the agreement between two raters, Cohen's κ is used and for more than three raters, the similar Fleiss' κ is calculated. For the frequency-based items, the data can also be analyzed as ordinal, as those items specify the extent to which a behavior is evident. This perspective requires the calculation of Spearman's Rho (ρ), which is a rank correlation coefficient. A threshold value of .60 is proposed for Kappa or Spearman's Rho coefficients for a reliability that is "moderate" (Döring, 2023, p. 559; translated by TJ) or "good" (Döring, 2023, p. 562; translated by TJ), while reliability coefficients below .40 are low. Nonetheless, the decision to exclude items should not only be made based on the reliability coefficients alone because they need to be discussed in the context of content and construct. Especially in complex situations, as RobS, and with complex category systems, as found in the rating manual, lower reliability coefficients might be acceptable (Döring, 2023, p. 562).

15.4.1.2.2 Results of Rating I on the Item Level

The results from the inter-rater reliability investigations will be presented according to the rating iteration in which they were generated. For each iteration, the quantitative reliability coefficients will be depicted first. After the results of the first full rating are presented, the rating discrepancies will be described using selected examples from the rated performances. This is done to provide context for the modifications that are made to the rating manual as documented thereafter. Subsequently, results of rating II are shown.

The results are based on the rating of $n = 10$ performances by rater 1 and the rating I of raters 2 and 3. In order to estimate an inter-rater reliability coefficient which subsumes all 27 items, the overall value for Fleiss' κ was calculated. For that, each raters' item-related codes were put into a list and the value generated (see Table 45). For rating I, the overall agreement is $\kappa = .62$ ($p < .001$; 95 % CI [.57, .67])¹³¹.

¹³¹ If not indicated otherwise, p -values always refer to two-tailed tests of significance.

Table 45: Calculation of overall Fleiss' Kappa.

Rater 1	Rater 2	Rater 3
Item 1 – Participant 1	Item 1 – Participant 1	Item 1 – Participant 1
Item 1 – Participant 2	Item 1 – Participant 2	Item 1 – Participant 2
...	...	...
Item 27 – Participant 10	Item 27 – Participant 10	Item 27 – Participant 10

Table 46 presents the Fleiss' κ values for the individual items and the three raters. The individual items range from $\kappa = -.03$ to $\kappa = 1.00$, indicating a spectrum from "poor" to "almost perfect", using the terminology of Landis & Koch (1977). *Use of material* has the lowest agreement with $\kappa = -.03$, followed by *Process* ($\kappa = .20$), *Revision* ($\kappa = .29$) and *Self-regulation* ($\kappa = .31$), though none of these unsatisfying values reaches the level of significance. *Dialogue orientation* ($\kappa = .39$), *Quality of Feed Back* ($\kappa = .35$), *Quality of Feed Forward* ($\kappa = .33$) and *Language accuracy* ($\kappa = .30$) all show significant, though only "fair" agreement (Landis & Koch, 1977, p. 165). Items *Error patterns* ($\kappa = .44$), *Amount and balance of information* ($\kappa = .44$), *Target-oriented error selection* ($\kappa = .42$), *Detailedness* ($\kappa = .45$), *Surface: Language appropriateness* ($\kappa = .52$), *Deep: Process* ($\kappa = .58$), *Body language* ($\kappa = .58$) and *Professional accuracy* ($\kappa = .52$) all show significant moderate agreement. The other eleven items exceed the threshold of $\kappa = .60$ and show at least "substantial" agreement (Landis & Koch, 1977, p. 165).

Table 46: Fleiss' κ of rating I.

Item	Fleiss' κ
Self-evaluation	.83 ($p < .001$)
Positive feedback	.71 ($p < .001$)
Learning objectives (Feed Up)	.68 ($p < .001$)
Process	.20 ($p = .282$)
Self-regulation	.31 ($p = .094$)
Use of material*	-.03 ($p = .850$)
Prioritization of deep-level features	.86 ($p < .001$)
Error patterns	.44 ($p = .016$)
Surface: Language correctness	.71 ($p < .001$)
Surface: Language appropriateness	.52 ($p = .004$)
Deep: Structure	.83 ($p < .001$)
Deep: Content	.68 ($p < .001$)
Deep: Process*	.58 ($p = .001$)
Self	.83 ($p < .001$)
Body language	.58 ($p = .001$)

Item	Fleiss' κ
Adaptation of language*	1.00 ($p < .001$)
Revision	.29 ($p = .079$)
Adaptive error correction	1.00 ($p < .001$)
Dialogue orientation	.39 ($p = .004$)
Quality of Feed Back	.35 ($p = .007$)
Quality of Feed Forward	.33 ($p = .018$)
Checking understanding	.64 ($p < .001$)
Amount and balance of information	.44 ($p < .001$)
Target-oriented error selection	.42 ($p = .002$)
Detailedness	.45 ($p = .001$)
Language accuracy	.30 ($p = .042$)
Professional accuracy	.52 ($p < .001$)

Table 47 shows the Spearman's ρ and Cohen's κ values comparing raters pairwise. Spearman's ρ is included to allow for a rank-oriented discussion of the categories that can be interpreted from an ordinal perspective.

Table 47: Pairwise comparison of inter-rater reliability scores of rating I.

Item	Rater 1 - Rater 2		Rater 1 - Rater 3		Rater 2 - Rater 3	
	Spearman's ρ	Cohen's κ	Spearman's ρ	Cohen's κ	Spearman's ρ	Cohen's κ
Self-evaluation	1.00 ($p = \text{n. a.}^{132}$)	1.00 ($p = .002$)	.76 ($p = .010$)	.74 ($p = .016$)	.76 ($p = .010$)	.74 ($p = .016$)
Positive feedback	.67 ($p = .035$)	.62 ($p = .035$)	.67 ($p = .035$)	.62 ($p = .035$)	1.00 ($p = \text{n. a.}$)	1.00 ($p = .002$)
Learning objectives (Feed Up)	.76 ($p = .010$)	.74 ($p = .016$)	.61 ($p = .060$)	.55 ($p = 0.53$)	.80 ($p = .005$)	.78 ($p = .011$)
Process	.25 ($p = .486$)	.23 ($p = .429$)	.58 ($p = .077$)	.58 ($p = .065$)	-.17 ($p = .645$)	-.15 ($p = .598$)
Self-regulation	.20 ($p = .580$)	.20 ($p = .527$)	.50 ($p = .141$)	.40 ($p = .114$)	.50 ($p = .141$)	.40 ($p = .114$)
Use of material*	n. a. ¹³³	n. a.	n. a.	n. a.	n. a.	n. a.
Prioritization of deep-level features	.82 ($p = .004$)	.80 ($p = .010$)	.82 ($p = .004$)	.80 ($p = .010$)	1.00 ($p = \text{n. a.}$)	1.00 ($p = .002$)
Error patterns	.22 ($p = .545$)	.21 ($p = .490$)	1.00 ($p = \text{n. a.}$)	1.00 ($p = .002$)	.22 ($p = .545$)	.21 ($p = .490$)

¹³² n.a. = not available; i.e., no level of significance is reported. This occurs, for instance, if to both raters rate the item as a constant and, thus, no variance occurs in the data.

¹³³ No coefficients are generated due to the "kappa paradox" (Feinstein & Cicchetti, 1990), as explained in Chapter 15.4.1.3.1.

Item	Rater 1 - Rater 2		Rater 1 – Rater 3		Rater 2 – Rater 3	
	Spearman's ρ	Cohen's κ	Spearman's ρ	Cohen's κ	Spearman's ρ	Cohen's κ
Surface: Language correctness	.67 ($p = .035$)	.62 ($p = .035$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	.67 ($p = .035$)	.62 ($p = .035$)
Surface: Language appropriateness	.51 ($p = .133$)	.41 ($p = .107$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	.51 ($p = .133$)	.41 ($p = .107$)
Deep: Structure	.76 ($p = .010$)	.74 ($p = .016$)	.76 ($p = .010$)	.74 ($p = .016$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Deep: Content	.61 ($p = .060$)	.55 ($p = .053$)	.76 ($p = .010$)	.74 ($p = .016$)	.80 ($p = .005$)	.78 ($p = .011$)
Deep: Process*	.51 ($p = .133$)	.41 ($p = .107$)	.67 ($p = .035$)	.62 ($p = .035$)	.76 ($p = .010$)	.74 ($p = .016$)
Self	.76 ($p = .010$)	.74 ($p = .016$)	.76 ($p = .010$)	.74 ($p = .016$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Body language	.38 ($p = .286$)	.38 ($p = .236$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	.38 ($p = .286$)	.38 ($p = .236$)
Adaptation of language*	1.00 ($p = n.a.$)	1.00 ($p < .001$)	1.00 ($p = n.a.$)	1.00 ($p < .001$)	1.00 ($p = n.a.$)	1.00 ($p < .001$)
Revision	-.05 ($p = .895$)	.04 ($p = .867$)	.79 ($p = .007$)	.64 ($p = .014$)	.41 ($p = .242$)	.29 ($p = .197$)
Adaptive error correction	1.00 ($p = n.a.$)	1.00 ($p < .001$)	1.00 ($p = n.a.$)	1.00 ($p < .001$)	1.00 ($p = n.a.$)	1.00 ($p < .001$)
Dialogue orientation	.72 ($p = .020$)	.27 ($p = .145$)	.88 ($p < .001$)	.40 ($p = .045$)	.90 ($p < .001$)	.61 ($p = .012$)
Quality of Feed Back	.35 ($p = .324$)	.27 ($p = .204$)	.14 ($p = .703$)	.18 ($p = .302$)	.88 ($p < .001$)	.69 ($p = .001$)
Quality of Feed Forward	.80 ($p = .006$)	.53 ($p = .019$)	.47 ($p = .168$)	.24 ($p = .147$)	.56 ($p = .093$)	.33 ($p = .077$)
Checking understanding	.76 ($p = .011$)	.49 ($p = .049$)	.81 ($p = .005$)	.64 ($p = .019$)	.96 ($p < .001$)	.82 ($p = .003$)
Amount and balance of information	.71 ($p = .022$)	.25 ($p = .235$)	.72 ($p = .019$)	.70 ($p = .001$)	.72 ($p = .018$)	.41 ($p = .051$)
Target-oriented error selection	.64 ($p = .048$)	.23 ($p = .285$)	.70 ($p = .024$)	.51 ($p = .029$)	.52 ($p = .121$)	.54 ($p = .015$)
Detailedness	.73 ($p = .016$)	.39 ($p = .066$)	.68 ($p = .031$)	.52 ($p = .014$)	.65 ($p = .041$)	.49 ($p = .034$)
Language accuracy	.70 ($p = .025$)	.35 ($p = .160$)	.60 ($p = .068$)	.33 ($p = .018$)	.67 ($p = .035$)	.29 ($p = .035$)
Professional accuracy	.65 ($p = .042$)	.48 ($p = .045$)	.55 ($p = .098$)	.47 ($p = .056$)	.79 ($p = .007$)	.64 ($p = .014$)

In terms of pairwise inter-rater agreement, no values could be calculated for *Use of material* as raters 2 and 3 both rated all ten cases identically, which prohibits the calculation of the reliability coefficients, due to the "kappa paradox" (Feinstein & Cicchetti, 1990). For *Error patterns*, *Surface: Language appropriateness*

and *Body language* there is significant absolute agreement between rater 1 and 3. Reliability values for these items that include rater 2 are, however, below the value of $\kappa = .60$. Similarly, *Revision* and *Amount and balance of information* show substantial agreement between raters 1 and 3, but not if rater 2 is included. *Dialogue orientation*, *Quality of Feed Back* and *Professional accuracy* show substantial agreement between raters 2 and 3, while the inclusion of rater 1 leads to an insufficient reliability measure for these items. The agreement for *Quality of Feed Forward*, *Target-oriented error selection*, *Detailedness* and *Language accuracy* is also not above $\kappa = .60$ for any of the pairwise examinations.

As the frequency-based items can also be considered ordinal, Spearman's ρ provides another perspective on the consistency between raters. Interestingly, items *Dialogue orientation*, *Amount and balance of information*, *Detailedness* and *Language accuracy* have consistency coefficient values above the threshold of $\rho = .60$ (Döring, 2023, p. 559), although the κ -value is insufficient. For both *Quality of Feed Back* and *Quality of Feed Forward*, consistency values are not sufficient in two pairwise comparisons, while for *Target-oriented error selection* and *Professional accuracy*, it is insufficient in one pairwise comparison, respectively.

15.4.1.2.3 Modifications

The focus of consultation II was to discuss discrepancies, identify those which can be resolved and modify the rating manual accordingly¹³⁴. A detailed overview of each individual discrepancy in rating I is found in Appendix H.2. Modifications are only of specifying nature and do not inherently change the items, their scoring or their operationalization, but amend them by reformulating explanations or adding examples and details. With regard to the event-based items and the scoring of zero points or one point/two points, the majority of discrepancies were due to raters not identifying relevant events during the rating process. A reviewing of the instances in the video and rating manual lead to an agreement in most cases and raters were free to modify their judgements for rating II. In the following, the modifications to the rating manual are presented as the result of the consultations. This detailed description is warranted, as it also relates to the exact item descriptions in the final scoring instrument presented in Chapter 8.3 and the subsequent use of the Robs score in Chapter 15.5. The final version of the rating manual is found in Appendix B.

For some items, modifications were made that were similar. For the items *Error patterns*, *Surface: Language correctness*, *Surface: Language appropriateness*, *Deep: Structure*, *Deep: Content*, *Deep: Process*, it is added that the explicit nature

¹³⁴ Consultations were not recorded. Results are thus based on notes taken during the consultations and the documented codes and comments by the raters.

is also evident through the use of material, e.g., the physical pointing at aspects in the text. For the event-based items *Positive feedback*, *Process*, *Self-regulation*, *Surface: Language correctness*, *Surface: Language appropriateness*, *Deep: Structure*, *Deep: Content*, *Deep: Process*, the amendment of 'at least once' is added to specify more clearly that the occurrence of one instance of the event is sufficient for the scoring of two points. No modifications are made to *Self*, *Body language*, *Prioritization of deep-level features*, *Adaptation of language*, *Quality of Feed Back*, *Quality of Feed Forward*, *Language accuracy* and *Professional accuracy*.

First, individual modifications for event-based items are presented. The item *Self-evaluation* is amended with the specification that the self-evaluation must occur before *specific* feedback is given and that "[opening] remarks that do not contain specific information are acceptable" (Appendix B) as raters stated that there were uncertainties regarding these events.

For *Learning objectives (Feed Up)*, the term learning objectives is specified in the explanation as 'context-related'. This is done to emphasize and reiterate this notion as it is already described in the description of the item. The modification is made based on a discrepancy in the rating of GJS22. The participant stated that the target audience for the news report is English language speaking, which requires the appropriate mediation or translation of the German quotes [05:39 – 06:14]. This event was not coded by rater 2 and 3, as rater 3 argued that this is not directly stated as a learning objective in the material. The event was, however, coded by rater 1, as the description of the category included learning objectives that are relevant in the context.

For *Process*, the specifications are made more explicit by adding that the focus of the process can either be on 'writing', in a more general sense, or on 'writing the news report', as long as the comments refer to the task at hand. This modification is made because of a discrepancy between rater 1 and 2, in which the same event was coded once for *Self-regulation* (rater 1) and once for *Process* (rater 2). In the event, the participant GJS22 refers to the use of dictionaries when writing texts, however it is not clear if they refer to the task specifically or texts in general [0:05:08 -0:05:27]. The referral to the scaffolding in the sample task is included as an unsuccessful attempt in the manual, as it is not an adequate piece of feedback in this context because the scaffolding in question is neither particularly useful for the task at hand, nor does the participant address it specifically for the task, but rather as a general strategy and a form of self-regulation. This addition is made because rater 3 coded the following event, in which the participant falsely relates to the scaffolding in M4 as a source for future writing, with two points:

[0:03:01] TP: [wenn also also hier hier] habe ich jetzt zum beispiel so hilfstellung gegeben

[0:03:07] S: ah okay [ja]

[0:03:08] TP: [und] genau (.) so hilfestellung auf die kannst du vielleicht auch mehr achten

[0:03:13] S: mhmh. (MMK21)

The key phrase "You tried out new linguistic formulations" (Appendix B) is added as another example, which is adapted to also include feed back on the process level as does *Deep: Process*.

For *Self-regulation*, it is added that the strategy should be "specific". Specificity is then laid out in more detail as referring "to the fact that the strategy should directly address the question or topic" (Appendix B). Additionally, the provision of general advice on language learning is added as an unsuccessful attempt as it is not specific enough. Thus, the already indicated specificity is expanded upon and it is stated more clearly that vague comments are not coded here. This modification is made based on the following event, in which raters 1 and 2 initially disagreed on the specificity of the provided strategies:

[0:03:36] S: [okay also einf]ach auswendig lernEN

[0:03:38] TP: genau genau und äh schreibe sie dir immer wieder auf immer wieder und also (.) was was noch anderen schülerinnen und schülern (.) ähm geholfen hat oft ist halt immer wieder dieses (.) wenn man was sagt vielleicht mal kurz nur in sich gehen und sagen wie hätte ich das jetzt auf englisch geSAGT

[0:03:58] S: also wenn ich generell was auf deutsch sage oDER

[0:04:00] TP: genau genau also wenn du jetzt so einfache sätze oder zum beispiel auch ähm (.) wenn du musik hörst

[0:04:07] S: mhmh

[0:04:08] TP: dann du hörst ja bestimmt auch ganz viel (.) verschiedene

[0:04:11] S: ja

[0:04:11] TP: englischsprachige musik [oDER]

[0:04:12] S: [mhmh]

[0:04:13] TP: das du dann halt anfängst so immer wieder die musik zu hören und dich zu fragen was sagt der da eigentlich geraDE

[0:04:18] S: ah okay

[0:04:19] TP: das ist dann halt die nutzen ja auch alle möglichen (.) zeitenformen und so

[0:04:23] S: mhmh

[0:04:24] TP: und sowas hilft eigentlich immer den meisten schülerinnen und schülern weiter

[0:04:30] S: ja okay

[0:04:31] TP: und was halt auch vielleicht cool wäre ist die meisten gucken ja immer netflix serien

[0:04:35] S: [mhmh]

[0:04:35] TP: [oder] sowas

[0:04:36] S: ja

[0:04:36] TP: das du da einfach mal vielleicht anfängst den ähm englischen untertitel anzumachen du musst die ja nicht komplett auf englisch gucken oder [sowas]. (CCF27)

In this situation, the participant presents two possible strategies after the learner asked how to practice irregular verbs. One is to reflect on statements the learner makes in German and translate them into English, the other is to consume more media (songs, TV series) in the English language with subtitles, but neither directly addresses the learner's concrete issue.

For the item *Use of material*, the notion of accessibility is added to the explanation and used to replace the previous description "directed at the learner" (Appendix B) in the specifications. The aspect of accessibility emphasizes that the material should be appropriately arranged spatially, so that a potential learner could actually trace what the participants are showing. It is then explicated that having the text upside down from the learner's perspective, despite pointing at the text, is an unsuccessful attempt as the text is not accessible for the learner. This modification is based on a discrepancy in the rating of GJS22, as rater 1 rated the performance with zero points while raters 2 and 3 rated the performance with two points. In the performance, the material is always positioned so that only the participant is reasonably able to read the text correctly, while it is upside-down for the learner (see Figure 27). Raters 2 and 3, however, focused on the actual pointing at the material as an appropriate use and considered this sufficient.



Figure 27: Screenshot of the performance by GJS22 (on the left) [0:05:25].

Regarding *Error patterns*, 'describe' is replaced with 'address' after rating I, prior to the consultations. The previously used verb has a more illustrative nature and implies that the pattern has to be descriptively presented to the learner, which is not required. Rather, the participants need to focus on these patterns and provide feedback on them. In this regard, it is also added that the kind of feedback (KR, KCR or elaborated feedback) is irrelevant here, as long as information on the pattern is provided. Lastly, addressing repeated errors (e.g., multiple instances of the exact same error in the text) or error types (e.g., spelling mistakes) is excluded, as these are not necessarily patterns, as envisioned. The last modification is made based on a discrepancy in which rater 2 rated the following event as appropriately addressing an error pattern, while raters 1 and 3 did not:

[00:04:05] TP: das ist mir noch aufgefallen (.) zwei mal become anstatt got (.) become ist etwas werden (.) nicht etwas erhalten

[00:04:14] S: achso also wäre das jeTzT

[00:04:16] TP: genau das wäre im englisch we got thousand euros prize money nicht become das hieße wir werden oder wir wurden (.) das ist also nicht eins zu eins übertragbar auf englisch. (CHS10)

The participant here addresses the fact that the learner used the verb *become* incorrectly in two instances in the text. While rater 2 considered this a pattern, rater 1 and 3 viewed this as addressing a repeated error and not a pattern.

The item *Deep: Content* is amended in the description to include 'content-related aspects'. Thus, aspects that are logical within the given context, and not only content aspects that are evident in the task sheet, are coded here. This modification is based on a discrepancy in which participant CHS10 argues that the text does not include specific information on the date and location, which is atypical for a news report and should be included by the learner [0:00:45 - 0:01:50]. Rater 1 considered this notion as an aspect of content, while raters 2 and 3 did not, as the participant did not refer to aspects specifically from the task sheet.

The explanation and specification for *Adaptive error correction* are modified so that the initial question is no longer specified as being 'open'. The exact degree of when a question is open was discussed in the initial consultations and discarded subsequently for rating I.

The item *Revision* is first of all nominally changed, as the previous term 'Process orientation' was considered too close to the other items *Process* and *Deep: Process*. The specifications with regard to contradictory messages for the scoring of two points is also added explicitly to the score of zero points, to make this specification more prominent in the rating process. This modification is made

based on the performance of KMP05, in which all three raters identified the same event but coded it differently:

[0:08:30] S: mhmh okay (.) muss ich den text denn jetzt nochmal neu schreibEN

[0:08:34] TP: also komplett neu schreiben musst du den nicht also ich habe wie gesagt noch eine überschrift die fehlt

[0:08:40] S: mhmh

[0:08:41] TP: dann halt so ein bisschen einen absatz machen nach dieser (.) introduction (.) und das größte ist dann halt noch dieses zitat von dem ähm (.) das war von dem miniSTER ja von dem minister das du das irgendwie wenn du das nicht übersetzen willst das du es dann halt umschreibst weil zu deutsch da drin ist halt ein bisschen schlecht (.) aber ansonsten (.) so von der grundform ähm ja (.) also. (KMP05)

The participant answers the VT “Do I have to completely rewrite the text?” by first stating that no complete rewriting of the text was necessary. They then go on to point out that a headline was missing, a paragraph could be made after the introduction and that, most importantly, the German quote had to be translated or paraphrased, while indicating that, overall, the text was sufficient. Rater 1 coded this event with zero points, as the message of whether the learner has to rewrite the text is vague or even contradictory, considering that the participant first indicates that no complete rewriting has to occur, but then addresses aspects on a feed back level that need improvement. Rater 3 considered this instance to be of advisory nature, thus coding one point, while rater 2 regards this instance as a clear instruction for revision and coded two points.

Some of the frequency-based categories are also modified. For *Dialogue orientation*, amendments are made to the condition, as most discrepancies resulted from a lack of identifying the corresponding event. Specifically, a definition of ‘cognitively demanding’ is added that specifies that “[cognitively] demanding questions or prompts engage the learner more deeply with the context-relevant substance” (Appendix B). This modification is based on rater 1 identifying the following events from different participants as cognitively demanding for the learner, while raters 2 and 3 did not:

[0:01:04] TP: äh in den newspaper article (.) ähm (..) ja (.) ansonsten wie ist dir denn die aufgabe gefalLEN also fandest es war LEICHT oder nicht so LEICHT (CCF27)

[0:00:27] TP: genau ähm vielleicht zu erst einmal was glaubst du denn wo deine stärken und schwächen bei dem text laGEN (MMK21)

[00:05:58] TP: selling art and picking up garbage that the bank bought das wusste ich nicht ganz genau was du meiNST. (CHS10)

In their statements, CCF27 and MMK21 require the learner to evaluate difficulties in the process, or reflect on weaknesses in the text, while CHS10 prompts the learner to explain a statement in her text. Thus, clearer guidelines for when the condition is fulfilled are included.

For *Target-oriented error selection*, a clarification of chronology is added in the specification. It states that “[starting] with aspects that occur at the beginning of the text does not necessarily indicate that the errors and/or error patterns are addressed in a chronological manner, as a subsequent conscious selection is still possible” (Appendix B). Such a clarification is deemed necessary, as, for example, in the performance of AEM23, the chronological nature was evaluated based on the first few initial feedback statements by raters 2 and 3 [0:00:39 – 0:00:58], although a focus on relevant errors is evident later in the feedback conversation [00:59 – 01:41].

The description for *Detailedness* is slightly modified. It is added that the extent of elaboration is not a relevant factor here, as the explanation indicates that only the kind of feedback is relevant. This modification is made as rater 1 often rated the performances higher than raters 2 and 3, who stated that they also considered the extent to which elaborated feedback had occurred as evident, e.g., in the performance of KMP05.

The description for *Amount and balance of information* is amended by a clarification of what constitutes an ‘aspect’. The specifications already state that an aspect is context dependent and can include several pieces of information, depending on how it is communicated. An example which serves to illuminate the issue is found in the performance of AEM23, which rater 1 rated with one point, while raters 2 and 3 rated it with zero points. Rater 3, for example, understood eight different aspects to be addressed by the participant in the performance: Addressing the preposition before ‘Tuesday’, tense, the incorrect use of the word ‘action’, spelling mistakes, the forming of irregular verbs, content, copying phrases from the task sheet and the erroneous item ‘speeck’¹³⁵. Rater 1, however, considered some of the aspects that were distinguished by rater 3 to be a single aspect which then includes several pieces of information. Thus, this amendment is made to further emphasize that point.

For *Checking understanding*, the qualifying frequency expressions are replaced. The expression ‘often’ is replaced by ‘several times’ and ‘sometimes’ is replaced by ‘a few times, but at least once’. This change is necessary, as the relational nature of ‘often’ and ‘sometimes’ led to uncertainties during the rating process. Thus, the item is changed to focus on the absolute instances. In line

¹³⁵ The aspects are not necessarily listed in a chronological order.

with the other codes, the inferential nature is still retained to account for the different lengths of the conversations. Additionally, the key phrase “‘Do you know what that/... means?’ In the sense of ‘Do you understand?’” is added.

15.4.1.2.4 Results of Rating II on the Item Level

After the described second consultation, raters were free to modify their rating. The overall agreement is again calculated for three raters, as described for rating I. For rating II, the overall agreement is $\kappa = .83$ ($p < .001$; 95 % CI [.78, .88]). Results of the item-specific agreement are depicted in Table 48. In the final rating, the spectrum of agreement between three raters ranges from $\kappa = .54$ to $\kappa = 1.00$, indicating an improved range from “moderate” to “almost perfect” (Landis & Koch, 1977, p. 165). With the exception of *Learning objectives* ($\kappa = .66$), all items improved in their agreement. There is an absolute agreement (between all three raters and thus also in pairwise comparison) for eleven items. With regard to Fleiss’ κ , *Deep: Content* ($\kappa = .81$), *Revision* ($\kappa = .89$) and *Language accuracy* ($\kappa = .86$) have an “almost perfect” agreement (Landis & Koch, 1977, p. 165). Items *Deep: Process* ($\kappa = .58$) and *Quality of Feed Forward* ($\kappa = .59$) narrowly miss the $\kappa = .60$ threshold (Döring, 2023, p. 559), while *Amount and balance of information* ($\kappa = .54$) fails to reach the threshold more clearly. These latter three items do show a level of reliability that can still be considered “sufficient” (Döring, 2023, p. 562), all the other items have at least “substantial” agreement (Landis & Koch, 1977, p. 165).

Table 48: Fleiss’ κ of rating II.

Item	Fleiss’ κ
Self-evaluation	1.00 ($p < .001$)
Positive feedback	.76 ($p < .001$)
Learning objectives (Feed Up)	.66 ($p < .001$)
Process	.60 ($p = .001$)
Self-regulation	.73 ($p < .001$)
Use of material*	1.00 ($p < .001$)
Prioritization of deep-level features	1.00 ($p < .001$)
Error patterns	1.00 ($p < .001$)
Surface: Language correctness	1.00 ($p < .001$)
Surface: Language appropriateness	1.00 ($p < .001$)
Deep: Structure	1.00 ($p < .001$)
Deep: Content	.81 ($p < .001$)
Deep: Process*	.58 ($p = .001$)
Self	1.00 ($p < .001$)

Item	Fleiss' κ
Body language	1.00 ($p < .001$)
Adaptation of language*	1.00 ($p < .001$)
Revision	.89 ($p < .001$)
Adaptive error correction	1.00 ($p < .001$)
Dialogue orientation	.70 ($p < .001$)
Quality of Feed Back	.60 ($p < .001$)
Quality of Feed Forward	.59 ($p < .001$)
Checking understanding	.76 ($p < .001$)
Amount and balance of information	.54 ($p < .001$)
Target-oriented error selection	.77 ($p < .001$)
Detailedness	.66 ($p < .001$)
Language accuracy	.86 ($p < .001$)
Professional accuracy	.66 ($p < .001$)

Table 49 contains the pairwise agreement and consistency coefficients. With regard to the pairwise comparison, sufficient agreement ($\kappa = .60$) is not achieved for all of the three comparisons for *Learning objectives* ($\kappa = .52$ for rater 2 – rater 3), *Process* ($\kappa = .40$ for rater 2 – rater 3), *Dialogue orientation* ($\kappa = .56$ for rater 1 – rater 2), *Quality of Feed Back* ($\kappa = .40$ for rater 1 – rater 2; $\kappa = .57$ for rater 1 – rater 3), *Quality of Feed Forward* ($\kappa = .56$ for rater 1 – rater 2 and for rater 1 – rater 3) and *Detailedness* ($\kappa = .52$ for rater 1 – rater 2 and for rater 2 – rater 3). However, for the frequency-based measures (i.e., the latter four items), consistency in terms of Spearman's ρ is sufficient.

Table 49: Pairwise comparison of inter-rater reliability scores of rating II.

Item	Rater 1-Rater 2		Rater 1 – Rater 3		Rater 2 – Rater 3	
	Spearman's ρ	Cohen's κ	Spearman's ρ	Cohen's κ	Spearman's ρ	Cohen's κ
Self-evaluation	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Positive feedback	1.00 ($p = n. a.$)	1.00 ($p = .002$)	.67 ($p = .035$)	.62 ($p = .035$)	.67 ($p = .035$)	.62 ($p = .035$)
Learning objectives (Feed Up)	.76 ($p = .010$)	.74 ($p = .016$)	.76 ($p = .010$)	.74 ($p = .016$)	.52 ($p = .120$)	.52 ($p = .098$)
Process	.82 ($p = .004$)	.80 ($p = .010$)	.67 ($p = .035$)	.62 ($p = .035$)	.41 ($p = .242$)	.40 ($p = .197$)
Self-regulation	.82 ($p = .004$)	.80 ($p = .010$)	.82 ($p = .004$)	.80 ($p = .010$)	.67 ($p = .035$)	.62 ($p = .035$)
Use of material*	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)

Item	Rater 1-Rater 2		Rater 1 – Rater 3		Rater 2 – Rater 3	
	Spearman's ρ	Cohen's κ	Spearman's ρ	Cohen's κ	Spearman's ρ	Cohen's κ
Prioritization of deep-level features	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Error patterns	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Surface: Language correctness	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Surface: Language appropriateness	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Deep: Structure	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Deep: Content	.76 ($p = .010$)	.74 ($p = .016$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	.76 ($p = .010$)	.74 ($p = .016$)
Deep: Process*	.51 ($p = .133$)	.41 ($p = .107$)	.67 ($p = .035$)	.62 ($p = .035$)	.76 ($p = .010$)	.74 ($p = .016$)
Self	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Body language	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)	1.00 ($p = n. a.$)	1.00 ($p = .002$)
Adaptation of language*	1.00 ($p = n. a.$)	1.00 ($p < .001$)	1.00 ($p = n. a.$)	1.00 ($p < .001$)	1.00 ($p = n. a.$)	1.00 ($p < .001$)
Revision	.87 ($p = .001$)	.83 ($p < .001$)	1.00 ($p = n. a.$)	1.00 ($p < .001$)	.87 ($p = .001$)	.83 ($p < .001$)
Adaptive error correction	1.00 ($p = n. a.$)	1.00 ($p < .001$)	1.00 ($p = n. a.$)	1.00 ($p < .001$)	1.00 ($p = n. a.$)	1.00 ($p < .001$)
Dialogue orientation	.85 ($p = .002$)	.56 ($p = .008$)	.89 ($p < .001$)	.70 ($p = .001$)	.95 ($p < .001$)	.85 ($p < .001$)
Quality of Feed Back	.65 ($p = .043$)	.40 ($p = .059$)	.78 ($p = .007$)	.57 ($p = .004$)	.93 ($p < .001$)	.85 ($p < .001$)
Quality of Feed Forward	.89 ($p < .001$)	.69 ($p = .002$)	.78 ($p = .008$)	.56 ($p = .005$)	.77 ($p = .008$)	.56 ($p = .004$)
Checking understanding	.87 ($p = .001$)	.66 ($p = .010$)	.94 ($p < .001$)	.82 ($p = .003$)	.94 ($p < .001$)	.82 ($p = .003$)
Amount and balance of information	.70 ($p = .023$)	.39 ($p = .066$)	.72 ($p = .019$)	.70 ($p = .001$)	.80 ($p = .005$)	.57 ($p = .004$)
Target-oriented error selection	.70 ($p = .024$)	.66 ($p = .010$)	1.00 ($p = n. a.$)	1.00 ($p < .001$)	.70 ($p = .024$)	.66 ($p = .010$)
Detailedness	.68 ($p = .031$)	.52 ($p = .014$)	1.00 ($p = n. a.$)	1.00 ($p < .001$)	.68 ($p = .031$)	.52 ($p = .014$)
Language accuracy	.85 ($p = .002$)	.80 ($p < .001$)	.85 ($p = .002$)	.80 ($p < .001$)	1.00 ($p = n. a.$)	1.00 ($p < .001$)
Professional accuracy	.76 ($p = .011$)	.67 ($p = .004$)	.55 ($p = .098$)	.47 ($p = .056$)	.87 ($p = .001$)	.83 ($p < .001$)

The remaining discrepancies will be presented for categories that are below the required threshold value of $\kappa = .60$ (Döring, 2023, p. 559) (see Table 50). These are the items *Quality of Feed Forward* ($\kappa = .59$), *Deep: Process* ($\kappa = .58$) and *Amount and balance of information* ($\kappa = .54$). As discrepancies for *Deep: Process* remained identical to rating I, which is outlined in the previous chapter, the category will not be discussed in detail again.

Table 50: Selected discrepancies between rater 1 and rater 2 / 3 in rating II.

TP	Rater 1 - 2		Rater 1 - 3	
	Quality of Feed Forward	Amount and balance of information	Quality of Feed Forward	Amount and balance of information
AEM23		R1=1; R2=0		R1=1; R3=0
CCF27				
KMP05		R1=0; R2=1		
MMM07				
MMK21		R1=2; R2=1		
CHS10	R1=1; R2=0			
GJS22				
EME03	R1=2; R2=1		R1=2; R3 =1	
EEH01			R1=2; R3 =1	
MJO28		R1=2; R2=1	R1=2; R3 =1	R1=2; R3=0

The reliability coefficient for *Quality of Feed Forward* has improved compared to rating I. Rater 2 and 3 no longer disagree on the scoring of zero points for MMK21, while rater 3 also agrees with rater 1 and 2 on the scoring of zero points for CCF27. Generally speaking, remaining discrepancies in the scoring of the category are due to rater 1 coding the performances with one more point than raters 2 and 3. Looking at the pairwise comparison, agreement is insufficient in those that include rater 3. Both rater 2 and 3 score the performance of EME03 with one point, while rater 1 scores it with two points. There is only one remaining discrepancy which includes the scoring of zero points: Rater 2 scores CHS10 with zero points while rater 1 scores the performance with one point. Rater 3 and rater 1 also diverge for EEH01 and MJO28.

The category *Amount and balance of information* has only slightly improved reliability values in rating II. Rater 3 did not modify their rating and, thus, the same discrepancies remain. Rater 2 changed the rating of CCF27, but four discrepancies remain. For instance, both rater 2 and 3 continued to score the performance of AEM23 with zero points, while rater 1 scores it with one point.

15.4.1.3 Discussion of Items for the Final Scale

Based on these results, the exclusion or inclusion of items with insufficient reliability values is discussed. Items that are not discussed here are satisfyingly reliable with regard to frequency, internal consistency and inter-rater agreement.

15.4.1.3.1 Exclusion of Items

In the following, the reasoning for the exclusion of the items *Deep: Process*, *Adaptation of language* and *Use of material* is given.

Exclusion of Deep: Process

The item *Deep: Process* is the item that is least frequently scored with two points ($n = 5$). Furthermore, it co-occurs with the scores of two points for the items of *Process* and *Deep: Content* in $n = 5$ performances, and with *Deep: Structure* in $n = 4$ performances. Additionally, it always overlaps with the coded segments for two-points for *Process* and once with *Deep: Content*. These results are somewhat expected, as the operationalizations of *Deep: Process*, as a focus on diagnosis and feed back of process-related aspects in the learner product, is a subset of *Process*, which includes feed back on the process, but also includes feed up and feed forward. As it was not foreseeable to what extent each item would generally occur in the sample, both items were included in the rating process. Furthermore, the corrected item-total correlation for *Deep: Process* is .20 and thus below the targeted value of .30, indicating that it is also not substantially relevant for the total score. Regarding to inter-rater reliability, the item is fairly reliably with an initial κ of .58 ($p = .001$), although it also does not improve in rating II. Still, it does not reach the threshold level of $\kappa = .60$. As a consequence of these results, *Deep: Process* is excluded from the final set of items as it does not add any substantially new facet to the overall scoring instrument and merely complicates the scoring process as an additional item.

Exclusion of Adaptation of language

All three codes of the item *Adaptation of language* are used, although it is scored with one point in only $n = 4$ performances. As this score does not occur that frequently, the operationalization of the item could potentially be modified so that behavior that falls into the one-point category is included in one of the other two categories. The item has consistently perfect agreement between the three raters ($\kappa = 1.00$, $p < .001$). With regard to internal consistency, however, the item has the worst value with $\rho = -.17$ and is the sole item responsible for the insufficient internal consistency (i.e., $\alpha < .80$). Looking at the inter-item correlation matrix (see Appendix H.3), it correlates negatively with 16 of 26 items. Most substantially, it seems to negatively correlate with *Language accuracy*

($\rho = -.36$), *Positive feedback* ($\rho = -.29$), *Dialogue orientation* ($\rho = -.28$) and *Quality of Feed Back* ($\rho = -.22$). Potentially, those participants that achieve two points in *Adaptation of language* and at least speak English initially are more inclined to make mistakes that could lead to the scoring of less than two points for *Language accuracy*. Furthermore, due to their focus on speaking English, they might be less inclined to seek a proper dialogue with the learner and are less likely to qualitatively inform the learner about their performance in general, and with a focus on positive aspects. As this item was derived from the normative setting of functional monolingualism in the German EFL classroom (see Chapter 7.3.2.1), this facet might not be transferable to this feedback situation, which is adjacent to the EFL classroom but not a distinct part of it. As a conclusion, the item is excluded from the final set of items.

Exclusion of Use of material

The item *Use of material* is only scored in $n = 2$ performances with zero points, indicating that the vast majority of participants included the material in the conversation properly. The inter-rater reliability is estimated as $\kappa = -.03$ for all three raters. Pairwise agreement cannot be calculated as raters 2 and 3 scored all performances with two points in rating I and only rater 1 scored a single performance of the subsample with zero points. The negative or non-existent agreement is the result of a phenomenon described as *kappa paradox*, which occurs if there is a high agreement between raters, but there is an imbalance in rating, e.g., one scoring prevails, as is the case here (Feinstein & Cicchetti, 1990). As the rating II indicates, reliability has improved after consultations about the performance in which zero points are scored. Internal consistency of this item is estimated as $\alpha = -.08$, indicating a small, but still negative correlation with the other items in the scale. The item correlates negatively with 19 of 26 items. Possibly, as the item is coded so frequently, negative correlations are generated with items that occur less frequently and thus have a more differentiating contribution within the scale. Consequently, the item is excluded as it presents an additional item that requires scoring but does not contribute substantially to the differentiation of the performances.

15.4.1.3.2 Inclusion of Items Despite Scoring Frequency

Some item scores are used relatively frequently or infrequently within the scored performances but are still included in the final scale. The text surface-related items *Language correctness* and *Language appropriateness* are scored with two points in $n = 49$ and $n = 48$ times, respectively. Both items are also identically coded as zero or two points in $n = 48$ performances. The almost consistent scoring of two points for surface-level features is expected, as these features are

more easily identifiable in the text. The frequent co-occurrence of surface-level features is also not surprising considering that a main error pattern in the text is the use and forming of past simple. While the use relates to the appropriateness, the form relates to correctness and naturally, both aspects are often addressed in a performance, presumably in close proximity¹³⁶. These two items are the most frequent ones that are scored with two points, but remain in the set of items as they are relevant to assess participants' ability to provide feedback on a range of issues in the text. Also, they are easily identifiable, as indicated by the inter-rater reliability coefficients in rating II, so their exclusion would not necessarily facilitate rating. Exclusion is thus not warranted, considering that both items represent relevant facets of the text that require feedback.

Zero points are only scored in $n = 4$ performances for the item of *Self*. Nonetheless, the item is included in the final set of items. Generally, the lack of feedback on the self-level is to be seen as positive due to its potential harmful impact on the learner (Hattie & Timperley, 2007, p. 97). Score frequency is thus not an appropriate indicator for a potential exclusion, considering the importance of this facet for the overall quality of the feedback conversation.

The item *Quality of Feed Forward* is scored with zero points in $n = 5$ performances, indicating that participants rarely gave feed forward that did not fulfill the criteria of the item, while one point and two points are scored in a relatively equal manner. Possibly, the intentional use of the VT "How can I practice that?" provides most participants with the opportunity to provide at least one instance of acceptable feed forward, which is then scored with at least one point. Considering this, the item should not be excluded from the set of items as it does not indicate that the behavior is not seen in the sample, rather, that even when given the opportunity, some participants are unable to provide adequate feed forward.

Similar in distributive pattern but opposite in the prevalence, *Checking understanding* is scored with two points in $n = 6$ performances, while scoring for one point or zero points is distributed relatively equally. Conversely to *Quality of Feed Forward*, this behavior cannot be directly triggered, rather, the participants must actively enquire if the learner has questions. The distributive pattern indicates that the majority of participants do check for understanding at least once, though not 'several times'. Potentially, one could modify the item to be a dichotomous item that only takes into account whether there was at least one instance of checking understanding, or not. However, this would reduce the sensitivity of the item: As different facets of the text are discussed in the ten-minute

¹³⁶ An investigation into that relation was not done as it is not directly relevant for the validity argument.

feedback dialogue, each piece of information presents a potential new challenge of comprehension for the learner for which understanding could be checked on. While learners might have understood feedback on the content, i.e., add more facts, they might not have understood the reasons for avoiding personal opinion in a news report. As the item is still rather sensitive towards the frequency of checking understanding, and the frequency of those events is considered important for the context, the item is kept in the set of items.

The item *Detailedness* is scored with two points in $n = 42$ performances, which is the highest occurrence for any of the three-tiered items. This frequency suggests that the dialogic one-on-one situation naturally encourages participants to elaborate on specific aspects of feedback rather than merely providing feedback at the level of *Knowledge of Results* (KR) or *Knowledge of Correct Response* (KCR). While it is positive that participants predominantly provide elaborated feedback, the item lacks the necessary variance to discriminate effectively between participants, likely due to the inherent nature of the RobS. Nevertheless, the item addresses a relevant facet of feedback by evaluating whether participants consistently offer some form of elaboration. Therefore, it remains included in the final set of items. For future adaptations, the operationalization of the item could be revised, and the levels of elaboration could be modified. For instance, zero points could be assigned when elaboration occurs rarely (not necessarily never), one point if it occurs sometimes, and two points if it occurs often.

15.4.1.3.3 Inclusion of Items Despite Insufficient Corrected Item-Total Correlation

Several items lack have an insufficient corrected item-total correlation value of $\rho < .30$. With the exception of *Deep: Content* ($\rho = .30$), and leaving the already discussed and excluded *Deep: Process* aside, none of the diagnosis-related items *Surface: Language correctness* ($\rho = .20$), *Surface: Language appropriateness* ($\rho = .17$) and *Deep: Structure* ($\rho = .25$) meet the desired threshold, neither do the items *Prioritization of deep-level features* ($\rho = .28$), *Self* ($\rho = .15$), *Revision* ($\rho = .15$), and *Language accuracy* ($\rho = .20$).

Although the items do not sufficiently correlate with the corrected item total, they are not excluded. Low item discrimination values are not appropriate exclusion criteria if the construct that they are used to represent is modelled broadly (Döring, 2023, p. 473; M. Stadler et al., 2021). This is the case for the RobS, as the framework is specifically designed to holistically include a wider range of aspects that are relevant for evaluating the feedback situation. With regard to the diagnosis-related items, each focuses on a different aspect of the text, and is therefore necessary to adequately depict participants' diagnostic capabilities. The *Prioritization of deep-level features* also indicates the relevance

that participants attribute to errors and aspects diagnosed in the feedback conversation. It possibly does not show sufficient correlation because it relates strongly to the deep-level-related items, which just manage to meet the threshold or are slightly below it. The notion of *Self* covers a distinct level in the feedback model proposed by Hattie and Timperley (2007) and should remain in the set of items, also, because of the possible negative impact such feedback has on the learning process. *Revision* focuses on participants' understanding of the writing process, which requires the use of feedback to improve the text (Hyland, 2019; Lee, 2017, p. 78). It is a quintessential item that explicates the underlying assumptions that the participants have of feedback and should be included in the set of items. Lastly, *Language accuracy* considers the participants' appropriate use of the English language. Possibly, language capability does not directly relate to the other items in the set of items, as it is not directly relevant for the feedback conversation, similar to the excluded *Adaptation of language*. As future EFL teachers, assessment should however take into account their mastering of the language, not least because the provision of inaccurate language-related information is potentially harmful for a learner's learning process. Thus, all discussed items are included in the final set of items.

15.4.1.3.4 Inclusion of Items Despite Potentially Insufficient Inter-Rater Reliability

The results of rating I show that the agreement between the three raters is insufficient (i.e., $\kappa < .60$) for 14 of the included items, while agreement in rating II is only insufficient for two items. The initial low agreement potentially indicates that either the rating manual was not specific enough or that the events are hard to identify in the videos. In the following, potential reasons for lack of reliability in rating I are discussed and modifications reflected upon with regard to the results of rating II and the inclusion of these items.

The results of rating I for the items *Process* ($\kappa = .20$, $p = .282$) and *Self-regulation* ($\kappa = .31$, $p = .094$) are to be considered as "fair" but below the threshold of $\kappa = .60$ (Döring, 2023, p. 559; Landis & Koch, 1977, p. 165). They are also below the required level of significance, impeding the interpretation of these results. The modifications that are made for both items of a more specifying nature, in that, for *Process*, it is more clearly defined as to what the process might refer to and for *Self-regulation*, it is emphasized that the comments had to provide specific and topic-related information. After clarification, adaptation of modified coding rules and a discussion of discrepant events, both items achieve sufficient agreement: *Process* ($\kappa = .60$, $p = .001$) and *Self-regulation* ($\kappa = .73$, $p < .001$). For *Process*, the agreement between raters 2 and 3 remains insufficient ($\kappa = .40$, $p = .197$), while it is also lower for *Self-regulation* ($\kappa = .62$, $p = .035$) between the two raters. Both items refer to distinct levels of feedback in the model

of Hattie and Timperley (2007) that expand upon the feedback that is directed at the text at hand. Potentially, it is more difficult for raters that are still in earlier stages of their professionalization to actually identify instances in which these levels are addressed in the videos. Other studies that also rate videos of feedback performances using similar operationalizations but employ more experienced coders, report higher agreement values of $\kappa \geq .90$ (Holstein et al., 2022; Lipowsky et al., 2019). However, these studies have a narrower focus in their coding process in general, concentrating distinctly on three of the four levels of feedback (Holstein et al., 2022) or on fewer categories (Lipowsky et al., 2019, p. 391) in the coding process. The lower values for the reliability coefficients are thus possibly also to be reasoned with the overall higher number of categories in the rating manual, as well as the experience of the raters. As the items are quite relevant from a content-perspective the items are included despite presenting a possible challenge to the reliability.

Other items that are within the "fair" range are more inferential: *Revision*, *Dialogue orientation* and *Quality of Feed Back* (Landis & Koch, 1977, p. 165). The interpretation of the results for *Revision* is complicated, as the results are not significant for rating I ($p = .079$). A clarification of the item's rule that contradictory statements with regard to text revision are to be scored with zero points as one cause of discrepancy in rating I is resolved in the consultation. The remaining disagreement in rating II ($\kappa = .89$) can be considered acceptable, as the discrimination between coding of one point and two points is based on the perceived rigor with which a revision is communicated. For *Dialogue orientation*, the condition is specified with a definition of what 'cognitively demanding' constitutes. This is necessary because the discrepancies in coding the condition are the main reason for the low agreement values. This is supported by the analysis of the data from a rank-oriented perspective. As the scoring refers to the degree of dialogue orientation, ranking two points, one point and zero points based on the frequency, the item can also be understood as ordinal. The rank-correlation coefficients indicate that the overall tendency of dialogue orientation is rated more consistently when all three raters are compared pairwise. As this holds true for both ratings, the item itself can be considered sufficiently reliable. With the specified amendments regarding the condition, the item is included in the final set. With regard to *Quality of feedback*, the consultations focused again on the condition as the identification of relevant events seemed to be the main reason there was a discrepancy between the scoring of zero points vs. one point or two points. Raters, however, did not find the rating manual to be lacking, rather, it is challenging to identify the situation in the videos. Thus, the manual was not modified and the item included unchanged.

The categories for *Error patterns*, *Surface: Language appropriateness*, *Body language*, *Target-oriented error selection*, *Detailedness*, *Professional accuracy* and *Amount and balance of information* achieved “moderate” agreement in the first rating (Landis & Koch, 1977, p. 165). With the exception of *Amount and balance of information*, all reached sufficient values in the second rating. For *Error patterns*, disagreement is the result of deviations in the rating of rater 2, as rater 1 and 3 agreed perfectly. A potential issue for disagreement in the item of *Error patterns* is the distinction between repeated errors and error patterns, which was resolved and also implemented in the rating manual. After a review of the identified events in the consultation, relevant events are identified. Similarly, the disagreements in rating I for *Surface: Language appropriateness* and *Body language* are also only evident in comparisons that include rater 2, as rater 1 and 3 display perfect agreement. After discussion and reconsideration in the consultation, the items show perfect agreement. Accordingly, these three event-based dichotomous items are included in the final set of items.

The category *Target-oriented error selection* shows consistent insufficient agreement values in rating I. A main reason that is identified for this finding is the notion of ‘chronological’ feedback. Especially performances that initially seem like a chronological discussion of the text, but include a conscious selection of errors later in the conversation, are not identically identified by the raters. A subsequent modification to the rating manual and consultation were necessary to increase reliability values to a satisfying degree. The remaining discrepancies are acceptable, considering the item allows for a certain degree of inference as the item is operationalized as a relative-frequency item and it is consequently included. The category for *Detailedness* is specified, in that the extent of the feedback elaboration is specifically excluded in the rating process, as the focus is only on the type and frequency of the information provided. After discussion of that notion, the item has sufficient reliability values. The rank-oriented perspective of the pairwise comparisons shows that the consistency of coding was above $p = .60$ in rating I and II, indicating the raters scored performances sufficiently consistently in that regard. As total agreement is reached with rating II and consistency values are consistently sufficient the item is included in the final scale.

Neither of the teacher accuracy-focused categories are modified after the first scoring, although both *Language accuracy* and *Professional accuracy* do not indicate sufficient reliability coefficients in rating I. However, the rank-related coefficients for most of the pairwise comparisons of both categories indicate that the performances are rated consistently similarly. Potentially, the raters, being in earlier stages of their professionalization, had initial difficulties to appropriately assess these other PSETs on knowledge-based facets as they are in their

peer group. This is supported by the fact that agreement was highest among rater 2 and 3 for *Professional accuracy*, as they seemed to agree on a different level than rater 1. Both items show, however, that the consultations, in part substantially, improved the agreement to a value above $\kappa = .60$ and, thus, the items are included in the final scale.

The agreement for the item *Quality of Feed Forward*, while improving after the consultations, only reaches a value of $\kappa = .59$, narrowly missing the threshold. The raters do not indicate any issues with the description of the category in the rating manual, thus no changes are made. The slightly insufficient reliability value is potentially a result of the more complex nature of this item as it requires the rating of feed forward instances using three different criteria – positivity, unambiguity and specificity. The item is kept in the set of items nonetheless. It has the highest correlation value in the item-total correlations, indicating that it is relevant for the overall test score. This does make sense considering that, according to Hattie, the feed forward stage is the most relevant one for the whole feedback conversation as it gives the learner the appropriate information to further improve (Zierer, 2019). The item, thus, also presents a quintessential facet for the construct of feedback – from an empirical, as well as a theoretical perspective. Considering that the value for Fleiss κ is also just slightly below the required value, it is justifiable to include the item.

The category *Amount and balance of information* is the second category not to achieve the threshold level of .60 (Döring, 2023, p. 559). After the first rating, the manual is adapted specifically to account for the discrepancies that arose with regard to the definition of an ‘aspect’. While the values improved in rating II, only one discrepancy is resolved. Looking at the pairwise comparisons, those that include rater 2 seem to have less agreement. The consistency values, while tending to overestimate reliability (Döring, 2023, p. 559), show that the raters have a sufficient degree of consistency. It appears that the clarifications regarding what constitutes an aspect did not substantially change the raters’ understanding of the appropriateness of the amount of information, as still four of the six discrepant ratings include the rating of zero points, which indicates that the amount of information is not appropriate. Arguably, the instructions could be further and more substantially modified in two ways to potentially improve reliability. First, the concept of aspect could be exchanged with the smaller unit of ‘pieces of information’. This might increase reliability, as these events are easier to distinguish and potentially more clearly countable. Naturally, this would require a reconsideration of the amount, as three to five pieces of information are not sufficient for a feedback conversation of ten minutes. Fixating on single pieces of information might, however, put participants at a disadvantage who focus on error patterns and elaboration of these by providing specific examples

from the text in which the patterns is evident, which requires a lot of single pieces of information. These participants might, then, exceed a specified amount of information, while participants that simply list an appropriate amount errors from the text are put at an advantage. Secondly, a more strictly defined number of aspects could be implemented. However, distinct absolute numbers could complicate the rating process, as the events need to be reliably identified and counted. Also, this would not necessarily allow for the consideration of the duration of the dialogue, as the same absolute amount of information might be appropriate for a shorter conversation, but not enough for a longer conversation. The actual balancing of information is apparently not a substantial issue, as only the comparison of rater 1 and 2 shows two instances of discrepancies which is justifiable considering the inferential nature of this category. While the implemented coding instructions do not lead to a sufficiently reliable agreement between the three raters, the discussed options for modifications have their own challenges, are quite substantial and potential benefits need to be explored separately. Due to the fact that the success of the modification is not guaranteed and would not only require another iteration of coding by the three raters but also a complete reevaluation of the whole sample due to its substantial nature, this was not done. Exclusion of the item, as another option, would however drastically degrade construct validity. Providing an amount of feedback which is not impeding learner's cognitive load is essential. As Lee states, often "[less] is more" (Lee, 2017, pp. 75–76) when it comes to teacher feedback. However, too little feedback is also not justifiable, considering the effort the learner has put into the text and into seeking feedback. As a consequence, the item is included in its current state because of its relevance for the construct despite insufficient reliability coefficients.

15.4.1.4 Revised Reliability Values on the Item Level

Based on the findings, the items *Deep: Process, Adaptation of language* and *Use of material* are excluded from the final set of categories for the generation of test scores. In the following, the revised reliability coefficients for internal consistency, as well as the overall agreement of the three raters with regard to rating I and II will be presented. Item frequencies will not be presented as they are unchanged and have already been described.

The revised value for Cronbach's alpha is $\alpha = .81$, which is an acceptable value (Döring, 2023, p. 274; Streiner, 2003, p. 103). The corrected item-total correlation matrix indicates that no item correlates negatively with the final scale (see Table 43). Cronbach's α is not improved by the deletion of further items. Seven items remain below the level of $p = .30$: *Prioritization of deep-level features* ($p = .27$), *Surface: Language correctness* ($p = .19$), *Surface: Language*

appropriateness ($\rho = .17$), *Deep: Structure* ($\rho = .24$), *Self* ($\rho = .14$), and *Revision* ($\rho = .17$) but are kept due to the previously discussed reasons.

In order to estimate an inter-rater reliability coefficient which subsumes all 24 items, the overall value for Fleiss' κ is calculated again. For rating I, the agreement is $\kappa = .60$ ($p < .001$; 95 % CI [.55, .66]). For rating II, the agreement is $\kappa = .82$ ($p < .001$; 95 % CI [.77, .88]). While the first value indicates "substantial" agreement in the first rating process, the agreement in rating II is "almost perfect" after the consultation and thus satisfactory (Landis & Koch, 1977, p. 165). The inter-rater reliability of the items can thus be considered sufficient, especially considering the complexity of the situations that are evaluated and the number of items that the raters have to consider during the rating process. While some individual items are below the desired threshold, their inclusion is discussed in detail. The reliability of the item-rating potentially benefits from multiple raters of one performance. As many of the discrepancies are the result of a lack of identification of an event, the general use of two raters might reduce the probability that events are not identified.

15.4.2 Scoring Reliability on the Test Score Level

This chapter focuses on the overall reliability of the test scores. First, the data presented descriptively. The results are discussed in terms of the scoring instrument's ability to discriminate between performances. Secondly, the inter-reliability of the generation of the test score is investigated. The test score is of special importance as it is the foundation for a subsequent grade and further investigations.

15.4.2.1 Descriptive Statistics and Tests on Normal Distribution for the Full Sample

In the following, the results of the final test scores are described. The results are the foundation for the subsequent discussion with regard to the scoring instrument's potential to discriminate between performances. An overview of the descriptive statistics is given in the boxplot in Figure 28 and Table 51. The mean of the test score in the full sample ($N = 51$) is $M = 31.22$ ($SE = 1.12$; $SD = 8.00$) and the median is identified at 33 points. The minimum number of scored points is 10, the maximum is 47, with the achievable maximum amount of points being 48. The listed percentiles in Table 52 provide a first impression of the distribution of test scores in the sample.

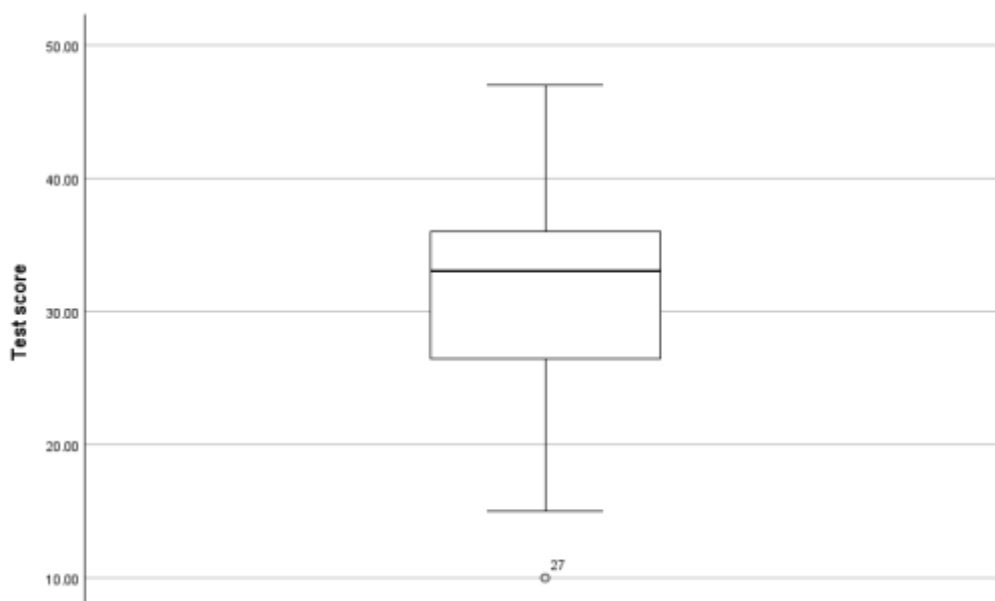


Figure 28: Boxplot of the RobS test scores.

Table 51: Descriptive statistics of RobS test scores.

	<i>n</i>	<i>M</i>	<i>SE</i>	<i>SD</i>	95 % CI		Median	<i>Min</i>	<i>Max</i>
					lower	upper			
Test score (points)	51	31.22	1.12	8.00	28.96	33.47	33	10	47
Test score (in %)	51	65.04	0.02	16.67	60.33	69.73	68.75	20.83	97.92

Table 52: Percentiles of the RobS test scores.

	<i>n</i>	Percentile						
		12.5 %	25 %	37.5 %	50 %	62.5 %	75 %	90 % ¹³⁷
Test score (points)	51	20	26	31	33	35	36	40
Test score (in %)	51	41.67	54.17	64.58	68.75	72.92	75.00	83.33

Table 53 shows the overall distribution of the test scores, grouped in five point-steps and Table 54 shows the number of participants that reached specific percentage-range of test scores. Of all participants, $n = 9$ or 17.64 % of participants do not achieve 50 % or more, with one participant achieving exactly 50 % (24 points). 14 participants achieve points in the range of 30-34 and 35-39 points, which are the largest single groups in the five-point distribution table. Figure 29 presents a graphic overview of all test scores¹³⁸.

¹³⁷ The 90 % percentile was chosen to have the test score as an even number.

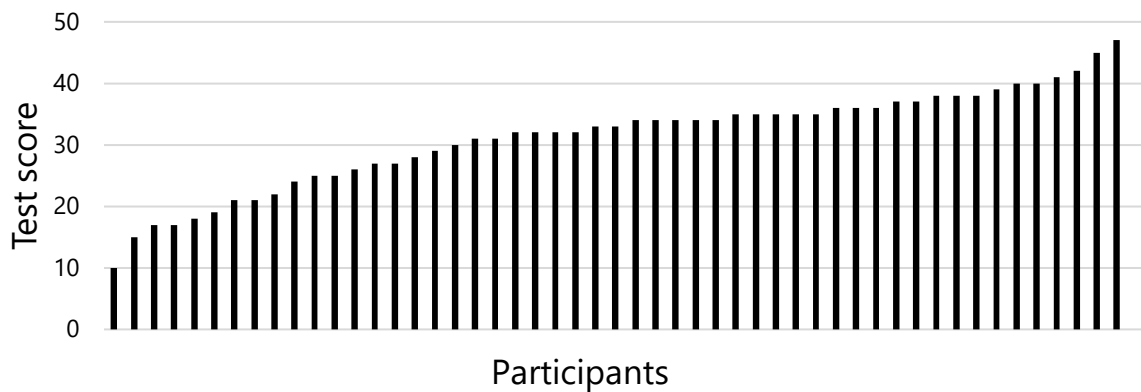
¹³⁸ A tabular overview is depicted in Table 85 in Appendix H.4.

Table 53: Descriptive statistics of participants in five-point intervals of the RobS test score.

Points	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-48
<i>n</i>	1	5	4	7	14	14	4	2
%	1.96	9.80	7.84	13.73	27.45	27.45	7.84	3.92

Table 54: Achieved RobS test score in %.

	0 - 19 %	20 - 29 %	30 - 39 %	40 - 49 %	50 - 59 %	60 - 69 %	70 - 79 %	80 - 89 %	90 - 100 %
<i>n</i>	0	1	5	3	7	10	18	5	2
Accumulated	0	1	6	9	16	26	44	49	51
%	0	1.96	7.84	5.88	13.73	19.60	35.29	9.80	3.92
Accumulated	0	1.96	11.76	17.64	31.37	50.98	86.27	96.07	100

**Figure 29:** Distribution of RobS test scores.

In Figure 30 the histogram and normal distribution is presented. The graphic indicates a right-shift in the distribution of the data, having a longer tail on the left. This indicates that higher test scores are generated more often. This is also evident in the skewness, which is found to be $g^{139} = -0.62$ ($SE = 0.33$). The z-standardized skewness can be used to investigate the significance of the skewness (A. Field, 2009, p. 139). It is calculated as

$$Z_{skewness} = \left| \frac{g - 0}{SE_{skewness}} \right| = \left| \frac{-0.62 - 0}{.33} \right| = 1.88.$$

Thus, the left skew is not statistically significant, i.e., < 1.96 , as is the cutoff value for the respective sample size (A. Field, 2009, p. 139). The graphic indicates that the peak and the tails of the data are more in line with an assumed normal

¹³⁹ According to Dormann (2017, p. 9), versions of SPSS usually provide the second correction of skewness.

distribution. The kurtosis value of the data is $k = 0.06$ ($SE = .66$) and also not significant (z-standardized kurtosis = 0.09). Concluding, the graphic indicates a left skew in the data which is however not significant, while kurtosis is more in line with an assumed normal distribution.

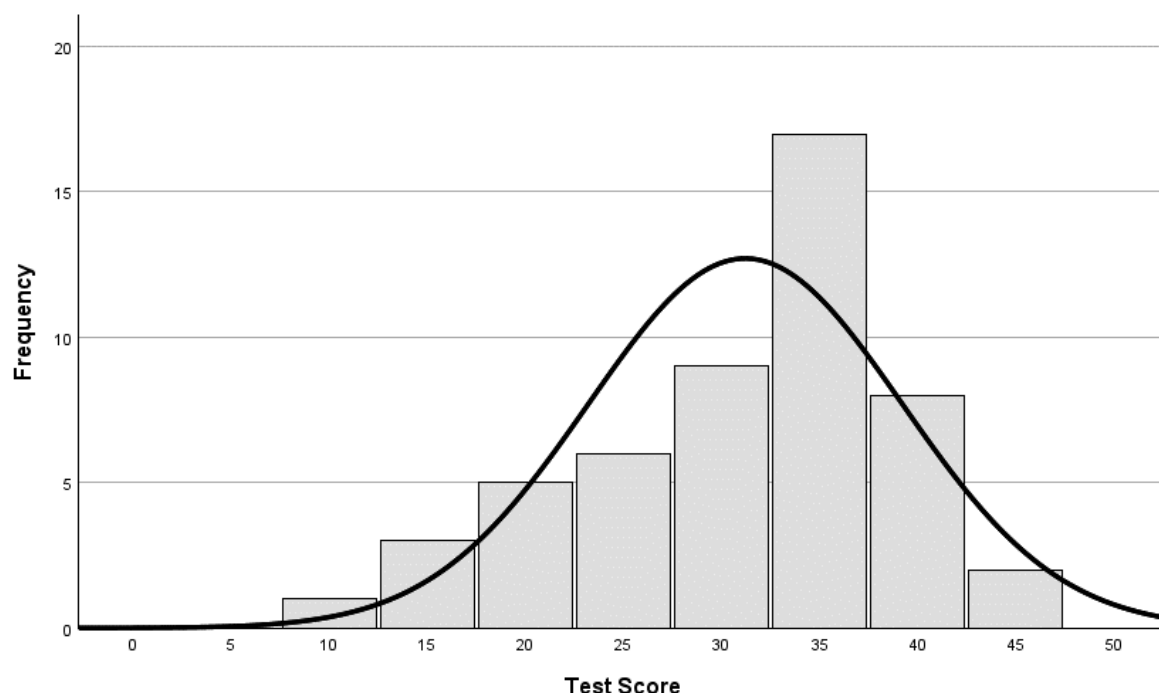


Figure 30: Histogram of the test scores in the full sample.

In addition to the graphic discussion of normal distribution, including skewness and kurtosis, numeric tests are another relevant source. In order to test for normal distribution, the Shapiro-Wilk-Test is chosen. It has proven to be the most powerful test and is often cited as being preferred for analysis with smaller sample sizes (Razali & Wah, 2011). The test shows that the data is distributed normally, $W = 0.96$, $p = .079$. An overview of statistics on normal distribution is provided in Table 55.

Table 55: Statistics on normal distribution of RobS test score.

	<i>n</i>	Skewness			Kurtosis			Shapiro-Wilk-Test	
		<i>g</i>	<i>SE</i>	<i>z</i>	<i>k</i>	<i>SE</i>	<i>z</i>	<i>W</i>	<i>p</i>
Rater 1	51	-0.62	-0.33	1.88	-0.60	0.66	0.09	0.96	.079

15.4.2.2 Inter-Rater Reliability on the Test Score Level

Using the aggregated item scores from the inter-rating procedure described in detail in Chapter 15.4.1.2.1, the inter-rater reliability of the total test score will be focused on. Identical to the approach taken to discuss this issue on the item level, the reliability will be explored for both rating I and II. First, the

methodological approach to measuring the inter-rater reliability will be presented before the results are displayed.

15.4.2.2.1 Methods

To discuss the reliability with which the test scores are generated, the data can primarily be viewed as interval scaled, as the aggregation of the item scores leads to equidistant intervals in the final scale. For the investigation of agreement between the three raters, the intra-class correlation coefficient (ICC) can be calculated. It allows for the investigation of consistency as well as direct agreement between three raters (Döring, 2023, p. 559). In order to determine the ICC, the data should be distributed normally for each rater and the variance between raters should be equal (Wirtz & Caspar, 2002, p. 162). While Döring refers to the similar thresholds for the ICC as previously outlined for Spearman's ρ and κ , Wirtz & Caspar describe that ICC scores $> .70$ are often used as indicators for sufficient reliability (2002, p. 160) and Cicchetti adds that a value between $.60$ and $.74$ is 'good' while anything above $.75$ is 'excellent' (Cicchetti, 1994, p. 286). Although these scores provide orientation in the discussion of ICC, the context of the data generation and analysis should, however, always be reflected upon when these scores are discussed. Additionally, pairwise comparisons between the raters are also useful in discussing if one rater deviates from the other two. For that purpose, Pearson's r provides the correlation of the test score as interval scaled while, again, Spearman's ρ does so for an ordinal perspective. The threshold level for both can be identified as $.60$ (Döring, 2023, p. 559).

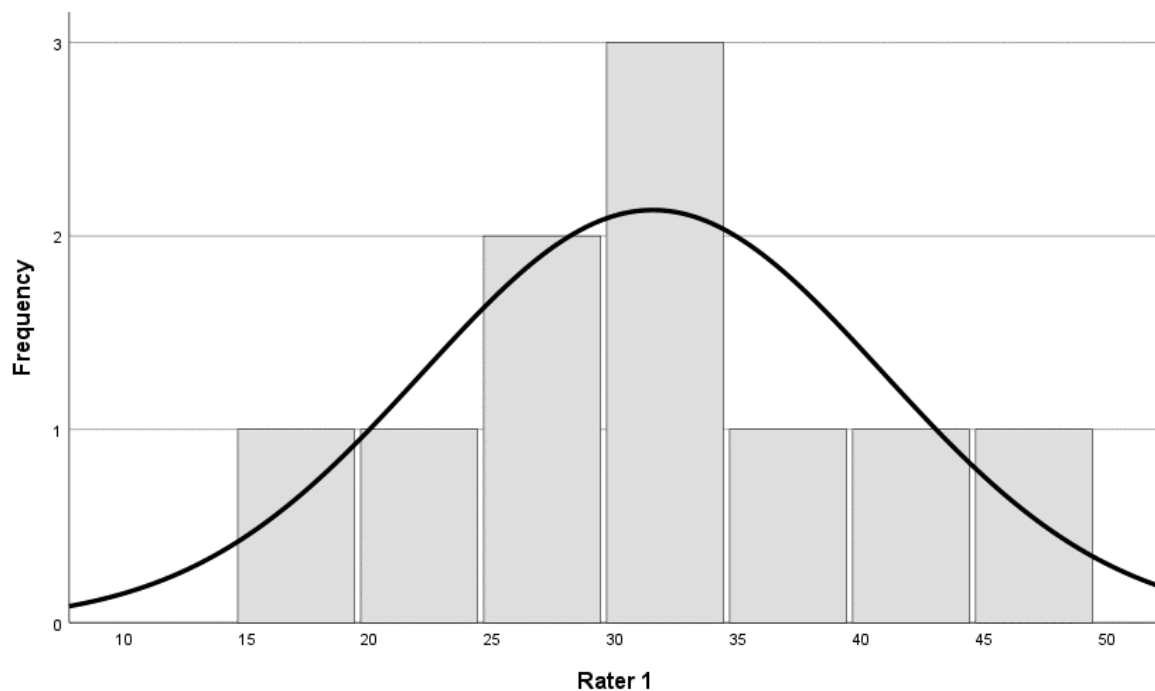
15.4.2.2.2 Results of Rating I on the Test Score Level

First, the generated test scores are described for the subsample for each rater. Table 56 displays descriptive statistics, i.e., mean, SE , SD , 95 % CI, median, minima, and maxima for the subsample that has been focused in the inter-rating of $n = 10$ participants. Rater 1 has a higher mean ($M_1 = 31.50$) than both rater 2 and 3, who have a similar mean ($M_2 = 28.70$; $M_3 = 28.30$). This is also reflected in a higher median, minimum and maximum for rater 1. All three raters have approximately equidistant differences for minima and maxima.

Table 56: Descriptive statistics of the RobS test scores of the three raters in rating I.

	<i>n</i>	<i>M</i>	<i>SE</i>	<i>SD</i>	95 % CI		Median	<i>Min</i>	<i>Max</i>
					lower	upper			
Rater 1	10	31.50	2.96	9.35	24.81	38.19	31	17	47
Rater 2	10	28.70	3.21	10.16	21.44	35.96	29	14	43
Rater 3	10	28.30	3.19	10.08	21.09	35.51	26.50	11	45

In order to explore potential significant differences between the raters, a one-way ANOVA is performed. The sample size of $n = 10$ performances is acceptable, as long as the requirements for variance-analyses are not violated (Bortz & Schuster, 2010, p. 214). Levene's test of homogeneity indicates that equal variances in the data can be assumed ($p = .888$, based on median). Looking at the histograms, however, a normal distribution of the data cannot be directly identified (see Figure 31 - Figure 33).

**Figure 31:** Histogram of the inter-rating subsample rater 1 - rating I.

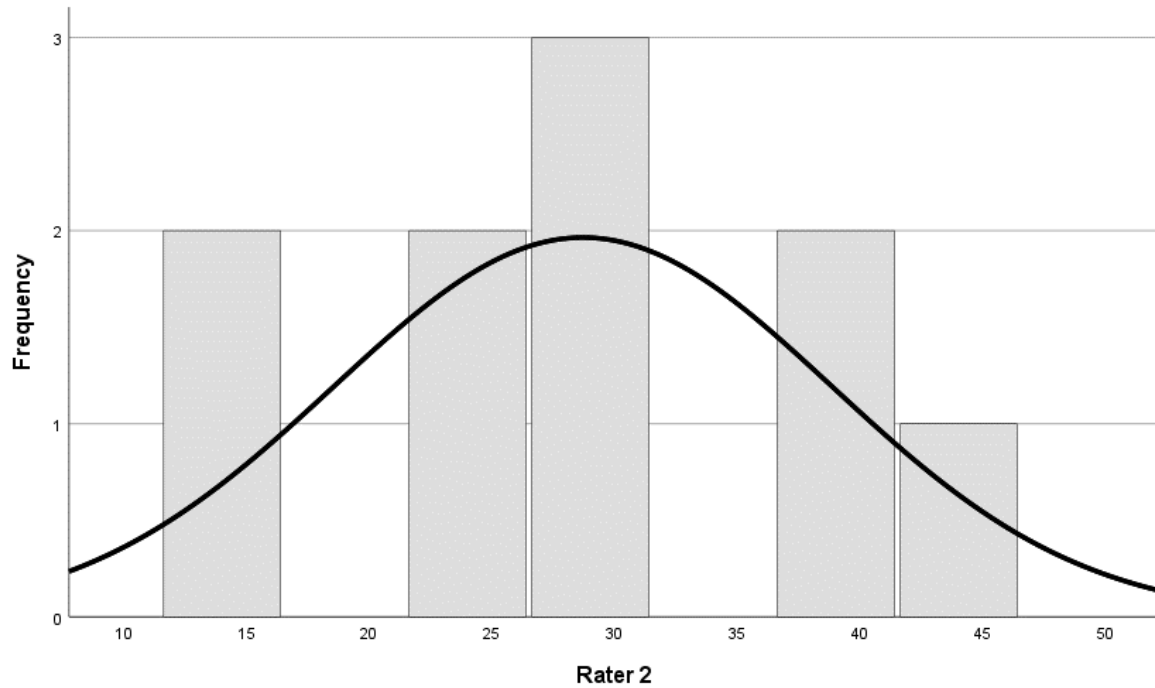


Figure 32: Histogram of the inter-rating subsample rater 2 - rating 1.

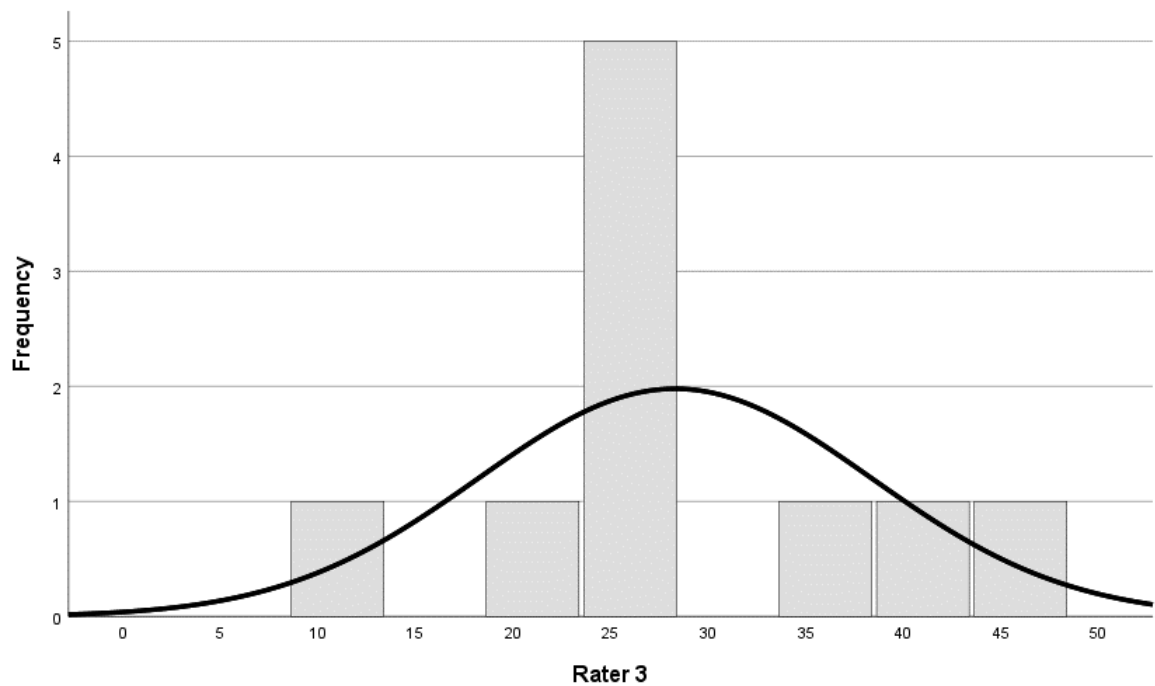


Figure 33: Histogram of the inter-rating subsample rater 3 - rating 1.

Table 57 shows the statistics on normal distribution of rating 1 for the three raters. The kurtosis value for rater 2 indicates a right shift and a longer left-leaning tail. The z-standardized score indicates that this shift is not statistically significant (i.e., < 1.96). Results of the Shapiro-Wilk-Test for normal distribution

indicate that the data for all three raters does not deviate from normal distribution ($p > .05$). Results of the ANOVA do not provide evidence that the raters differ statistically significantly from each other with regard to the mean, $F(2, 27) = 0.31$, $p = .734$. Considering the graphic analysis, the Kruskal-Wallis-Test is also performed as it does not require normally distributed data. Results do not yield a different result, as the null hypothesis is retained and the results from the ANOVA are corroborated, $H(2) = 0.68$, $p_{asympt} = .713$.

Table 57: Statistics on the normal distribution of rating I of all three raters.

	<i>n</i>	Skewness			Kurtosis			Shapiro-Wilk-Test	
		<i>g</i>	<i>SE</i>	<i>z</i>	<i>k</i>	<i>SE</i>	<i>z</i>	<i>W</i>	<i>p</i>
Rater 1	10	0.15	0.69	0.22	-0.60	1.33	0.45	0.99	.989
Rater 2	10	0.02	0.69	0.03	-1.26	1.33	0.95	0.94	.551
Rater 3	10	0.15	0.69	0.22	-0.01	1.33	0.01	0.95	.700

In order to investigate inter-rater reliability of the test score between the three raters, correlation coefficients can be used that require parametric data. In Figure 34, the test scores of rater 1 (X-axis) are plotted against those of both raters 2 and 3 (Y-axis). The diagram suggests that there is a positive relationship between the test scores of rater 1, and raters 2 and 3. In Figure 35, the test scores of raters 2 and 3 are plotted against each other, with a similar tendency visible.

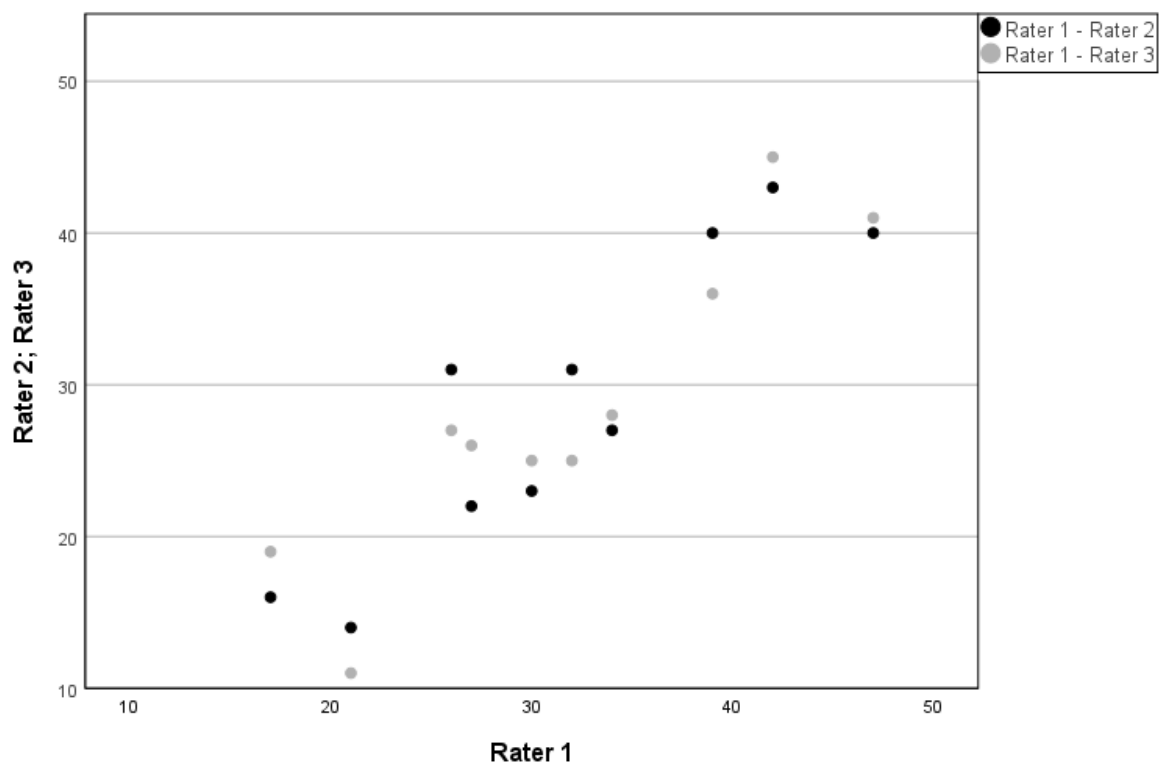


Figure 34: Scatter plot of rating I of rater 1 and raters 2 / 3.

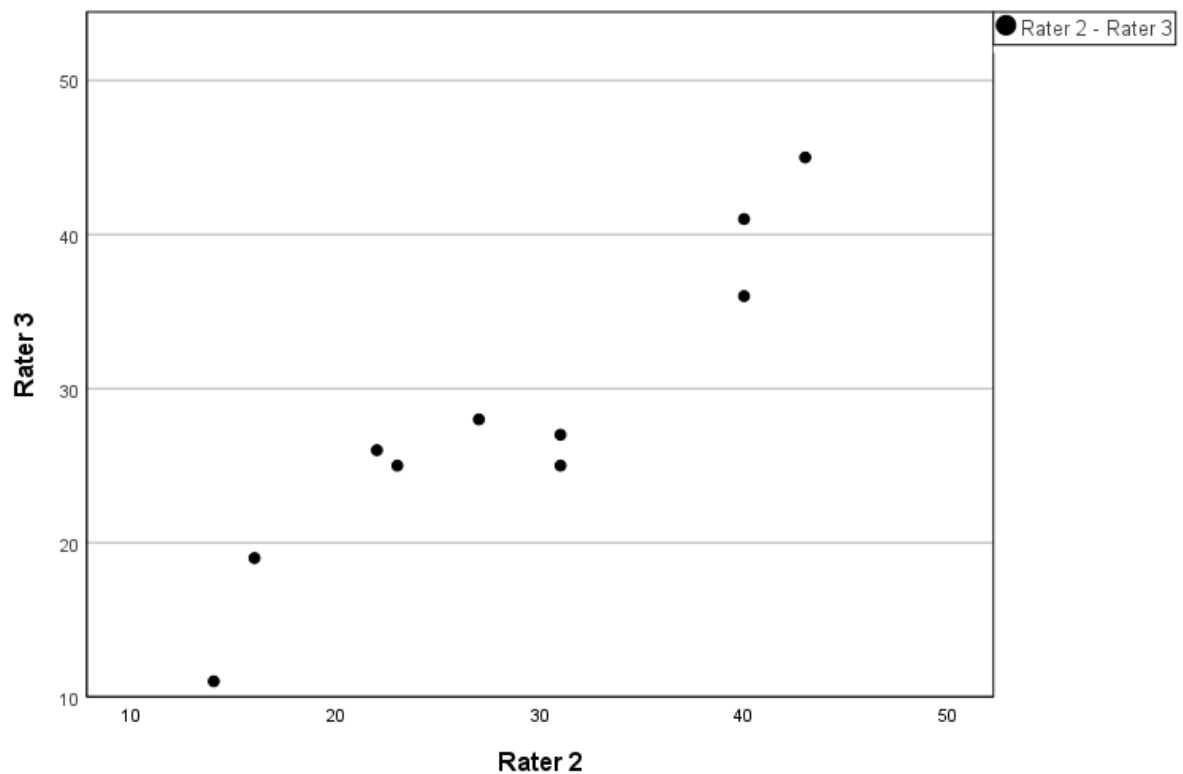


Figure 35: Scatter plot of rating 1 of rater 2 and 3.

Looking at the pairwise comparisons of the three raters in Table 58 and Table 59 below, Pearson correlation coefficients range from $r = .90$ to $r = .94$, $p < .001$, which indicates a good reliability (Döring, 2023, p. 559). Similarly, a rank-oriented analysis with the means of Spearman's ρ corroborates this interpretation, as the correlation coefficients are within the range of $\rho = .85$, $p = .002$, to $\rho = .89$, $p < .001$ and thus also good (Döring, 2023, p. 559).

Table 58: Pearson's r of the correlation of the RobS test scores between the three raters in rating 1.

Pearson's r		Rater 1	Rater 2	Rater 3
Rater 2	r	.90		
	Sig. (2-tailed)	< .001		
Rater 3	r	.90	.94	
	Sig. (2-tailed)	< .001	< .001	

Table 59: Spearman's ρ of the correlation of the RobS test scores between the three raters in rating 1.

Spearman's ρ		Rater 1	Rater 2	Rater 3
Rater 2	ρ	.85		
	Sig. (2-tailed)	.002		
Rater 3	ρ	.86	.89	
	Sig. (2-tailed)	.001	< .001	

The intra-class correlation coefficient (ICC) is a measure which indicates the degree to which multiple independent raters agree on the same metric variable (Wirtz & Caspar, 2002, p. 157). For the calculation of ICC, it is advised that the data is distributed normally and the variance is equal. Both requirements are tested previously and can be considered fulfilled for the sample. In order to test for between-rater effects, Tukey's test for non-additivity is performed (Wirtz & Caspar, 2002, p. 182). Results indicate that the null hypothesis cannot be rejected and thus no interaction between raters takes place ($p = .470$). Additionally, corrected item-total correlations are calculated. Raters' evaluations should not be systematically different, as this would complicate the interpretation of the ICC. The reported correlations in Table 60 indicate that the values are rather homogenous.

Table 60: *Corrected item-total correlations in rating I of raters 1, 2 and 3.*

	Corrected item-total correlation
Rater 1	.99
Rater 2	.96
Rater 3	.96

As all three raters rated the same $n = 10$ performances, and all raters have been trained, i.e., they are not randomly chosen, the two-way-mixed model for single measure¹⁴⁰ is used with both absolute agreement (just) and consistency (unjust) reported (Koo & Li, 2016): $ICC_{just}(2,1) = .89$, 95 % CI [.72, .97], $p < .001$; $ICC_{unjust}(3,1) = .91$, 95 % CI [.77, .98], $p < .001$. As the ICC score is above the threshold of .70, it can be considered at least good (Wirtz & Caspar, 2002, p. 160), while a score of above .75 is considered "excellent" by Cicchetti (1994, p. 286).

15.4.2.2.3 Results of Rating II on the Test Score Level

Table 61 displays the descriptive data for rating II. As the scoring of rater 1 did not change, it is not further elaborated on. Values for mean, *SE*, *SD*, 95 % CI and median of rater 2 and 3 are closer to those of rater 1. The minimal value is identical for all three raters, while the maximum has increased by one point for raters 2 and 3. Evidently, the ratings of all three raters converged in rating II.

¹⁴⁰ The two-way random model does not yield different results.

Table 61: Descriptive statistics of the test scores of all three raters in rating II.

	<i>n</i>	<i>M</i>	<i>SE</i>	<i>SD</i>	95 % CI		Median	Min	Max
					lower	upper			
Rater 1	10	31.50	2.96	9.35	24.81	38.19	31	17	47
Rater 2	10	30.30	2.97	9.41	23.57	37.03	28.50	17	44
Rater 3	10	31.40	3.13	9.89	24.32	38.48	32.50	17	46

An ANOVA was performed to test for potential differences in the rating. Homogeneity of variance was confirmed by Levene's test ($p = .991$, based on median). The data can be assumed to be distributed normally for each rater, as indicated by z-scores of skewness and kurtosis and results of the Shapiro-Wilk-Test ($p > .05$) (see Table 62). Looking at the histograms below for rater 2 (see Figure 36) and 3 (see Figure 37), normal distribution does not appear to be violated. Results of the ANOVA confirm that no significant differences exist between the three raters, $F = (2, 27) = 0.05$, $p = .953$. In order to compare the two scoring results, the Kruskal-Wallis test is also performed, corroborating the findings of the ANOVA, $H(2) = 0.14$, $p_{asympt} = .932$. Test scores of the three raters are plotted in Figure 38 and Figure 39¹⁴¹. The plots indicate a continued agreement of the test scores between the three raters in pairwise comparison. Agreement is further supported by the correlation coefficients of the pairwise comparison, as both the parametric perspective (r), as well as the ordinal perspective (ρ), provide consistent correlation scores that are significantly good as shown in Table 63 and Table 64 (Döring, 2023, p. 559)

Table 62: Statistics on the normal distribution of rating II of all three rater in rating II.

	<i>n</i>	Skewness			Kurtosis			Shapiro-Wilk-Test	
		<i>g</i>	<i>SE</i>	<i>z</i>	<i>k</i>	<i>SE</i>	<i>z</i>	<i>W</i>	<i>p</i>
Rater 1	10	0.15	0.69	0.22	-0.60	1.33	0.45	0.99	.989
Rater 2	10	0.25	0.69	0.36	-1.20	1.33	0.90	0.93	.468
Rater 3	10	-0.02	0.69	0.01	-0.90	1.33	0.68	0.95	.635

¹⁴¹ There is one exact agreement in the test score with 34 points for GJS22, which leads to an overlap in the graphic.

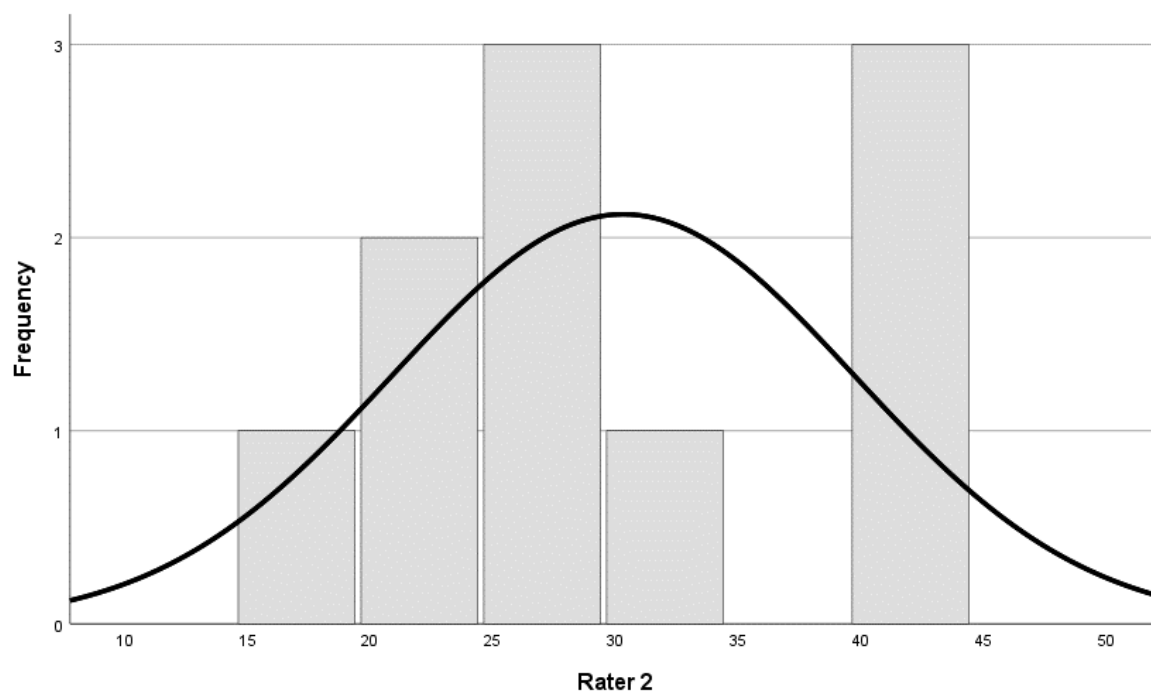


Figure 36: Histogram of the inter-rating subsample of rater 2 - rating II.

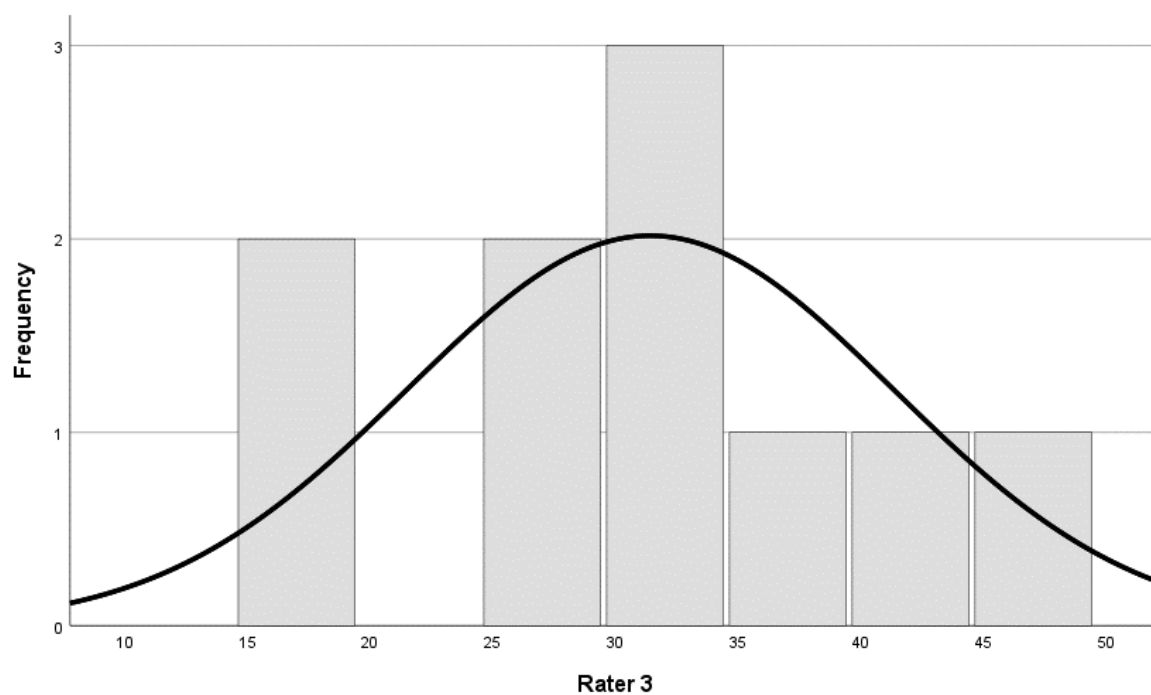


Figure 37: Histogram of the inter-rating subsample of rater 3- rating II.

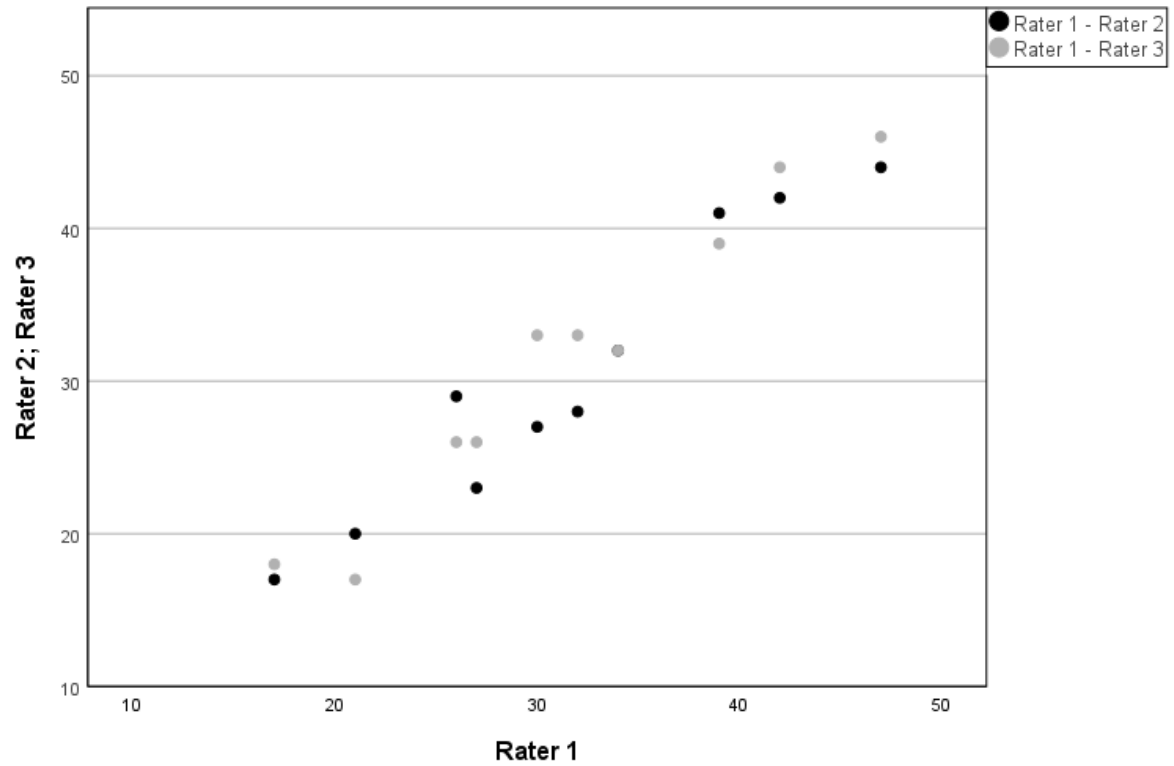


Figure 38: Scatter plot of rating II of rater 1 and raters 2 / 3.

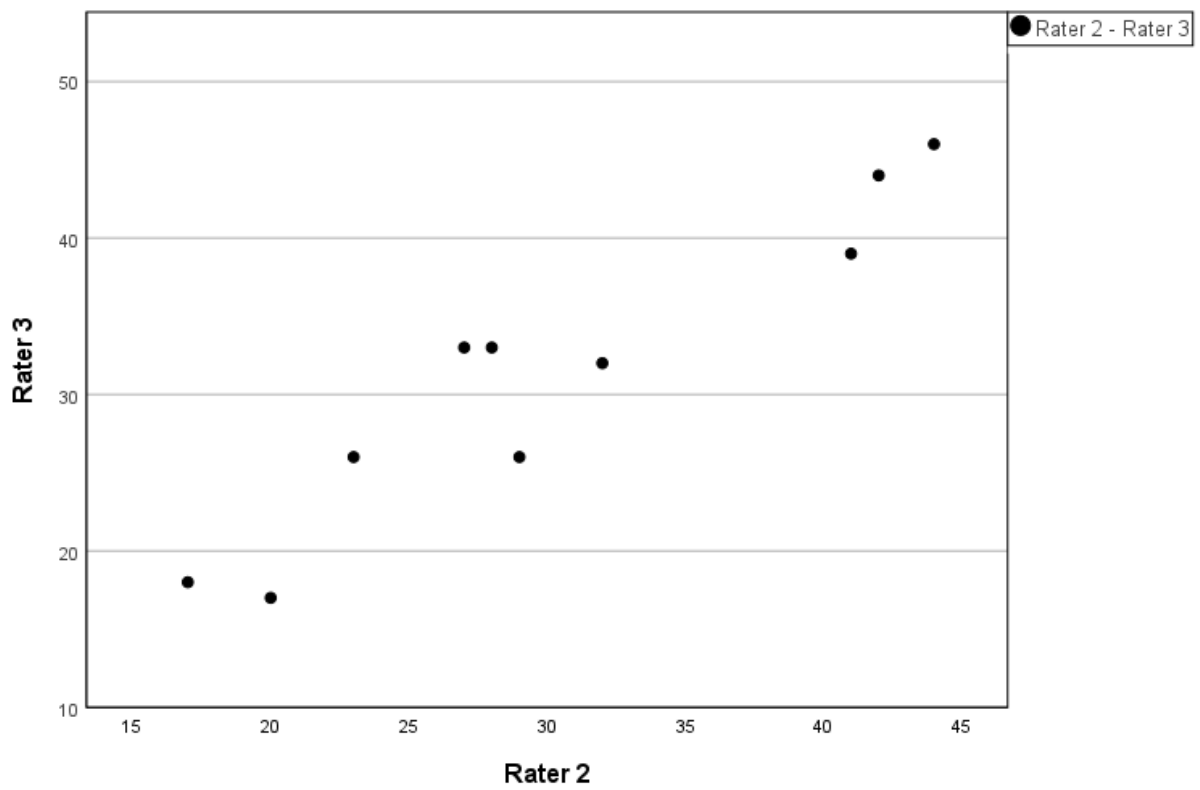


Figure 39: Scatter plot of rating II of rater 2 and 3.

Table 63: Pearson's r of the correlation of the RobS test scores between the three raters in rating II.

Pearson's r		Rater 1	Rater 2	Rater 3
Rater 2	r	.97		
	Sig. (2-tailed)	< .001		
Rater 3	r	.98	.95	
	Sig. (2-tailed)	< .001	< .001	

Table 64: Spearman's ρ of the correlation of the RobS test scores between the three raters in rating II.

Spearman's ρ		Rater 1	Rater 2	Rater 3
Rater 2	ρ	.93		
	Sig. (2-tailed)	< .001		
Rater 3	ρ	.95	.87	
	Sig. (2-tailed)	< .001	.001	

ICC is determined, as described in the previous chapter. The data's normal distribution was confirmed previously (Shapiro-Wilk test: $p > .05$). Tukey's test for non-additivity confirms that still no interaction between raters is evident ($p = .74$) and corrected inter-item correlations further support the homogeneity of the ratings (see Table 65).

Table 65: Corrected item-total correlations in rating II for raters 1, 2 and 3.

	Corrected item-total correlation
Rater 1	.99
Rater 2	.96
Rater 3	.97

The two-way-mixed model for single measure¹⁴² is used to calculate the ICC with both absolute agreement (just) and consistency (unjust) reported (Koo & Li, 2016): ICC_{just} (2,1) = .96, 95 % CI [.90, .99], $p < .001$; ICC_{unjust} (3,1) = .96, 95 % CI [.90, .99], $p < .001$. ICC scores are identical for the absolute, as well as the consistency model. Compared to rating I, the values have increased and can be considered good (Wirtz & Caspar, 2002, p. 160) or "excellent" (Cicchetti, 1994, p. 286).

¹⁴² The two-way random model did not yield different results.

15.4.2.3 Discussion of the Test Score-Focused Reliability

The results provide two perspectives into the reliability of the generation of test scores. First, the overall ability of the scoring instrument to generate test scores that differentiate between performance levels. Second, the reliability with which these test scores are generated by different raters.

The analysis shows that the scoring instrument is able to generate test scores that effectively differentiate between different performance levels. The lowest score in the sample is ten points and the highest score is 47 points, which is one-point shy of achieving the maximum score. The range of 37 points for lowest and highest score indicates that the scoring instrument is able to generate a test score that covers a large spectrum of the potential points. Nonetheless, scores in the range of 0-9 points are not generated in this sample. This notion is also reflected in the right shift of the data distribution (see Figure 30). This phenomenon might have different reasons. First, it might be due to the holistic and broad perspective taken in the underlying model. Naturally, a scoring instrument that considers a multitude of different facets gives raters more categories for which to allocate points and, thus, it might be less likely that fewer points are distributed overall. Second, some items are scored frequently with two points in the sample, but are considered relevant for the overall construct, as discussed previously, and thus not excluded (see Chapter 0). As a large number of participants receives points for these items, the accumulation of these item scores for the individual performances logically leads to fewer participants receiving fewer points. A third potential aspect is the general nature of the test design. Knowledge-focused assessment procedures generate lower test scores more easily as a lack of knowledge leads to a lack of points by simply not responding to items. The RobS, however, has a different sense of urgency as the communicative situation in the role-play demands action. Thus, participants might be more likely to receive points as they have to act in some capacity. Lastly, the nature of the sample might play a role. As participation is voluntary, the sample might be positively selected. Voluntary participation in such a research project might attract people that are generally more extrovert, openminded and communicative. As the communicative nature of the RobS is self-evident, this might also contribute to the right shift in the test scores.

In the sample $n = 9$ or 17.64 % of participants do not achieve more than 50 % (i.e., 24 points) in the RobS, as indicated in Table 54. The 50 % mark is relevant as it is often used as a benchmark to determine a pass or fail in exams in higher education (Sadler, 2005, p. 182). Consequently, about 18 % of the participants would have failed if this benchmark were applied here. Finding appropriate comparisons for this result has proven complicated. Looking at the second state examination in North-Rhine Westphalia as the exam with the highest stake

in teacher education, a failure rate of 8.67 - 10.22 % is reported in 2022 for the school type of *Gymnasien/Gesamtschulen* (Landesprüfungsamt für Lehrämter an Schulen NRW, 2022). It can be assumed that the extremely high stakes in the second state examination, though more similar in concept to the RobS than a written exam, do not allow for an adequate comparison. In one study, a rate of failure of 17.4 – 18.2 % for a written exam in the subject of social psychology and pedagogical diagnostics, which was part of the respective teacher education program, is reported (Perleth & Neumann, 2011). For OSCEs with multiple stations in medical education, reported failure rates range from 12.8 % (Homer, 2022, p. 466) to 33.7 % or 4.7 %, depending on the procedure of scoring (Kramer et al., 2003, p. 137). An OSCE, which is designed for the approbation of psychotherapists in Germany, has reported a very high failure rate of 59 %, although the administrators also considered 60 % as the benchmark for passing (Nelles et al., 2024). Evidently, failure rates vary, sometimes for the same OSCE and sample, depending on the scoring procedure. The potential failure rate of 18 % here is, thus, not outside the norm, considering other exams in teacher education, as well as other performance-based forms of assessment, have similar rates. The underlying question that remains is a discussion regarding the degree of difficulty of the test. By adjusting the minimum points for passing, the difficulty can be modified. This measure could also be used to counteract or balance the right shift in point distribution, which is previously discussed.

The reliability with which the test scores are generated is notably high between the three raters for both scorings. Both the pairwise comparisons between the three raters, as well as the results of the ICC, provide evidence that the test scores are generated in a consistent manner between the three raters. Compared to the ICC values that are reported in similar studies using simulations in Chapter 4.4.3, the results for this RobS indicate a higher agreement. For both ratings, the means between the three raters do not differ significantly, as corroborated by the results of the ANOVAs. The scores for the intra-class correlation are high concerning both the absolute agreement as well as consistency. Perhaps, while the agreement of coding of individual items is more challenging, the overall high number of items mitigates these discrepancies, leading to greater consistency in the test scores. This consistency is crucial for a prospective implementation of the RobS in the context of higher education, as the test scores not only determine pass or fail outcomes but are often converted into final grades, making reliability essential for fairness and accuracy.

15.4.3 Limitations and Implications from the Reliability Investigations

This chapter of the reliability study seeks to bring together the two main threads of the investigation, the item-focused and test-score-focused reliability. First, the limitations of the study will be outlined, before implications are explained.

First and foremost, the ratings are based on video recordings of the performances. This is beneficial from a research perspective, as the performances can be documented, transcribed and analyzed more thoroughly. Most importantly, this allows evaluators to modify the speed of the video and to rewind and watch specific events multiple times. Potentially, this video-based approach could also be taken if the RobS were to be implemented as an assessment format for teacher education at university. In medical education, however, these performance-based assessment formats are usually assessed *in situ*, in that the scoring is done during or immediately after the performance. Naturally, this bears a greater potential for errors in the rating process and complicates potential comparisons. Therefore, it is a question for future research to investigate and compare whether the reliability scores can be recreated if the scoring instrument is applied without the benefits that video-recorded performances offer.

Two further limitations are of methodological concern. First, the subsample that is used for the inter-rating procedure consists of $n = 10$ videos. This number represents about 20 % of the whole sample, and is comparable to other similar studies, as discussed. Regarding the qualitative investigation and revision of the rating manual, the number of inter-rated videos can be considered sufficient for the consultations. However, the statistical power of the calculations is restricted and puts greater weight on the individual rating (Döring, 2023, p. 791). This impact is, for example, evident in the discussion of item-focused reliability as a combination of a prevalence in scoring, which is also influenced by fewer overall data points, and single discrepancies between the raters can lead to low agreement values, as seen in the item *Use of material*. As a consequence, while the reliability coefficients for the items, as well as the test scores provide clues with regard to a reliable use of the scoring instrument by different raters, they have to be interpreted with caution. Second, the selection of the ten videos should be discussed. While the selection criteria and process are implemented to represent the whole range of performances of the full sample in the subsample, the results cannot be applied to the whole sample unequivocally. Moreover, despite following the self-imposed selection guidelines, the selection might have been unintentionally biased and thus have influenced the results of reliability coefficients. Another consequence of both the number of performances as well as the selection process might be an artificially enhanced correlation of the test scores. As the performances are selected to depict a variety of test scores, the differences between the performances might be more pronounced than in a different,

perhaps random, sample of videos or within the full sample. This could have led to the high ICC scores that are reported. Nonetheless, the results also provide evidence that different raters do find these differences in the performances, which is a relevant finding as it shows that the scoring instrument is sensitive to these differences. Considering these limitations, a future reevaluation of the full sample could provide further information that addresses these restrictions and allows for a broader perspective on the reliability of the scoring instrument.

Another limitation arises from the inter-rating procedure itself. First, the rater training might have had an impact. While the two raters received the identical material and the training was structured identically, they were trained separately due to unforeseen reasons, and this might impact the raters and their rating. Also, the training itself was quite extensive and took several hours, which might have impeded raters' receptive capacity. Second, and perhaps more importantly, several consultations are conducted between the raters, resulting in modifications to the rating manual, as well as the individual ratings. While the consultations were planned to discuss adjustments to the rating manual, it became evident that the complexity of the videos and scoring instrument led to instances in which an event was simply not identified by the raters in the rating process, but clearly coded as such when shown. It seems that the reliable scoring of the individual items requires at least two raters so as to simply not miss events in the videos. The rating of the whole sample based on the rating of rater 1 should be seen in this context; potentially, some events are not identified, which could have been, had a second rater also rated the videos. This was, however, beyond the scope of this project. Third, rater 1 might have influenced the other raters during the inter-rating process, despite trying to use the rating manual as an objective measure for the discussion throughout the consultations.

The results and outlined limitations can carry some implications. First, and continuing the last aspect of the limitations, multiple raters should jointly rate the performance. This could increase the likelihood that events are identified, especially in the case that the RobS is evaluated *in situ*, and not based on a video-recorded performance. Especially if the implementation of the RobS were to include detailed feedback for the individual facets of the model, this is advised as the item-specific reliability seems to benefit from multiple raters and consultations between them. If, however, a sole summative purpose of the RobS is the aim, the results of the test score reliability indicate that single raters are able to come to similar test scores without a second rater or consultation. Nonetheless, the question remains to what extent these results are replicable if the performance was rated *in situ*. Thus, if it is not possible to have multiple raters for a single performance, the use of videos could potentially be considered to allow raters to review scenes before coming to a final assessment.

15.4.4 Scoring Independence from Actors

A unique element of the RobS are the actors, who are trained to behave in a standardized manner so as to enable each participant to show their capabilities equally. As an integral part of the RobS, the role of the actors must be reflected upon from the perspective of reliability to examine whether measurement of errors can be partially attributed to them. In order to do so, an ANOVA will be performed to investigate if the test scores are statistically significantly different between the three raters.

15.4.4.1 Results

Descriptive statistics for each actor's subsample can be found in Table 66. Evidently, actor 1 is involved in only $n = 9$ simulations, while actor 2 ($n = 22$) and actor 3 ($n = 20$) participate in more and approximately the same number of simulations. With regard to the phase of study, the sample of actor 1 skews towards students in the Bachelor (BA) phase ($n = 8$), while this distribution is more balanced for actor 2 and 3. RobS with actor 1 achieve on average $M = 21.33$ points ($SD = 6.58$), while RobS with actor 2 achieve $M = 34.05$ points ($SD = 5.31$) and with actor 3, $M = 32.55$ points ($SD = 7.86$). As indicated in Table 86 in Appendix H.5, the data within the actor-specific samples does not seem to violate normal distribution (Shapiro-Wilk test: $p_{\text{actor}_1} = .983$; $p_{\text{actor}_2} = .508$; $p_{\text{actor}_3} = .473$) and Levene's test of homogeneity of variances is also not significant based on the median ($p = .316$).

Table 66: Descriptive statistics of the RobS test score of all three actors based on rater 1.

	<i>n</i>			<i>M</i>	<i>SE</i>	<i>SD</i>	95 % CI		Median	Min	Max
	Total	BA	MA				lower	upper			
Actor 1	9	8	1	21.33	2.19	6.58	16.28	26.39	21	10	33
Actor 2	22	12	10	34.05	1.13	5.31	31.69	36.40	34	21	45
Actor 3	20	11	9	32.55	1.76	7.86	28.87	36.23	34.50	15	47

The ANOVA shows that there is a significant difference between the groups, $F(2, 48) = 12.34$, $p < .001$. In order to identify which actor leads to the significant difference, the Bonferroni-post-hoc test is performed. Results indicate that results of actor 1 are statistically significantly different from actor 2, with a mean difference of -12.71 points ($SE = 2.63$; $p < .001$; 95 % CI [-19.34, -6.08], and actor 3 with a mean difference of -11.22 points ($SE = 2.66$; $p < .001$; 95 % CI [-17.82, -4.61]. There is no statistically significant difference between actor 2 and 3, $p = 1.00$.

15.4.4.2 Discussion

Considering the influence of the actor, it must be stated that participants that had simulations with actor 1 perform statistically significantly worse than those with actor 2 and 3. However, these results must be considered against the backdrop of the specific sample of actor 1. First, the sample is substantially smaller than those of actors 2 and 3. This is the result of test organizational issues, in that actor 1 was only available for a limited amount of time during the data generation due to the obligatory SIS they had to participate in. Second, their sample consists mostly of students in their Bachelor phase. As the results in Chapter 15.5 will show, participants in their Master phase achieve significantly more points than those in the Bachelor phase. Taking into account the limitations of the sample of actor 1, and considering that the much more comparable subsamples of actors 2 and 3 do not show a significant difference between the groups, results do not fully indicate that the actors has a clear impact on the test scores.

15.4.5 Summary

The aim of this chapter is to provide information on the reliability of the scoring instrument. All $N = 51$ performances are scored and the reasoning for the individual item score documented. These results are the foundation of the discussion of the frequency of items and distribution of test scores, as well as the internal validity. A subsample of $n = 10$ performances is used for inter-rating investigations with two further raters. The ten videos are selected to represent a wide range of possible level of performances, as well as to approximate other aspects of the sample like gender, phase of study and actor. The raters received extensive training. The rating process included two consultation sessions, the first after the rating of four videos, the second after the first full rating (Rating I). The goal of the second consultation was to discuss discrepancies of the rating and improve the rating manual. The raters were also able to modify their rating after the second consultation (Rating II).

The results are separated into an item-focused and a test-score-focused section. This is necessary because the items that are included in the final scale need to be sufficiently sound before they are used and aggregated into test scores. The initial set of 27 items was reduced to 24 with *Deep: Process, Adaptation of language and Use of material* being excluded. The item *Deep: Process* lacks any additional value as it is considered to be a subset of the item *Process. Adaptation on language and Use of material* are excluded based on their lacking internal consistency. Both items show negative corrected item-correlation scores and can, therefore not justifiably be included in the scale. Other items that do not achieve satisfying results are included, as they are considered to be relevant for

the construct. As the construct is modelled rather broadly, lower scores can also be expected and do not immediately require exclusion (Döring, 2023, p. 473). With regard to the agreement of the coding between the three raters, the overall agreement is $\kappa = .62$ in rating I, and, thus, "substantial" (Landis & Koch, 1977, p. 165), although individual items indicate results that are not satisfying. Discrepancies were discussed in the consultation and modifications are made to the rating manual. Rating II shows an improvement in the agreement, with $\kappa = .83$. Most individual items show agreement above the threshold of $\kappa = .60$, except for *Quality of Feed Forward* and *Amount and balance of information*. Those items are nonetheless included, due to their significance to the construct. After exclusion of the three items, the internal consistency of the scale is $\alpha = .81$ and the agreement is $\kappa = .82$.

Based on the finalized set of items, the test scores are focused on. Similarly, first the descriptive statistics based on rater 1, and then the results of the inter-rater reliability investigations are presented. The scores in the sample can be described as normally distributed, as indicated by discussions of skewness, kurtosis and the results of the Shapiro-Wilk-Test, although a right shift is evident in the data. The right shift is caused by the lack of scores in the range of 0- 9 points, or 0 -19 %, and might be caused by the frequently coded items that are not excluded, or a result of the performance-based test design. A potential threshold for failure of 50 % indicates that $n = 9$ participants (17.64 %) would fail the test as scored with the presented scoring instrument, which does not seem to be outside the norm, as far as available comparisons indicate. The inter-rater reliability of the generated test scores is examined by means of intra-class correlation. The results for both scorings indicate that the three raters have a high degree of consistency in the generated test scores. Potentially, the higher number of items mitigates potential reliability issues for single items so that the test scores still have a high correlation. However, the selection of performances might have emphasized differences between the individual participants, which might have led to artificially enhanced ICC scores.

Looking at the presented data as a whole, the final set of 24 items has a sufficient internal consistency and overall agreement between raters. The scoring instrument is able to cover a wide range of potential points, although not the whole range. The reliability of the generation of test scores, although potentially influenced by the selection of performances, is quite high. Further investigation into the validity will be based on the reported test scores.

15.5 Study 4: The Relationship between Test Score and Other Variables

As a last aspect for the validity argument, external criteria will be focused on to provide evidence for the discussion of the latent and interpreted trait. Hence,

several external variables will be put into relation to the test score of the RobS. First, the relevant criteria introduced in the IUA are outlined and the respective hypotheses stated. Second, the rationale behind the means of measurement of these external variables is discussed. Third, results will be displayed for each of the stated hypotheses. At the end, results will be discussed, limitations considered and a conclusion drawn. The findings of this study can be used to discuss the following claims and research questions:

Claim 6: Latent trait

The external variables that represent other facets of the target domain (i.e., EFL teacher competence) should correlate positively with the RobS score.

RQ13: To what extent does the RobS score correlate with other facets of the target domain?

Claim 7: Interpreted Trait

The collected data support the hypotheses that are made *a priori*, based on logical assumptions.

RQ14: To what extent are the hypotheses that are made *a priori* confirmed?

Due to the nature of this study, more detailed hypotheses for the claims and RQs will be presented in Chapter 15.5.2. The hypothesis will also be the focus of the subsequent discussions and referred to in the following instead of the RQs.

15.5.1 Relevant Criteria

In order to make interpretations of the results of the RobS, they should be put into context to other external variables. This is reflected in Claims 6 and 7 of the IUA. In Claim 6, it is argued that the RobS should take place in a situation which resembles the target domain. Thus, the RobS score should, ideally, align with test results from other tests that are part of the target domain, i.e. EFL teacher competence. In Claim 7, it is stated that the results can be used to interpret the participants' competence levels. This refers to logical *a priori* assumptions, such as the expectation that extended experience in the relevant field and academic progress lead to an increase of performance, resulting in a better test score, and that, experts' ratings of performance align with the RobS test score. For this study, four potential areas are thus of interest and will be focused on for the two claims: practical experience in school, academic success and progress, target domain-related knowledge and expert ratings. The order in which the aspects and results are presented in this section does not align with the IUA, but follows an inherent logic as naturally, academic progress and school-related experience might also influence other knowledge-related aspects of the target domain.

Thus, they will be explored first to identify potential covariates which must be controlled for.

A reasonable assumption can be made that participants who have had more experience in school-related settings do perform better as, the RobS seeks to measure the performance-oriented facet of competence in a situation that approximates practice. Thus, participants who have more experience, might perform better and achieve higher scores. School-related experience might occur either within the structure of the study program, e.g., as one of the three curricular internships in the Bachelor and/or Master phase (see Chapter 2.2), or extra-curricular. The latter might include additional internships, the work as a substitute teacher or other relevant activities.

Regarding more academically-related variables, it can be presumed that the academic progress correlates with the RobS score. Potentially, the number of semesters can be used as an indicator for study progress. However, the number of semesters does not necessarily indicate that the level of competence has increased, as length of study might be influenced by other factors such as semesters of absence or repetition of courses that have been failed. Thus, the distinction between participants who are in the Bachelor program and those who are in the Master program should also be made. The latter group is likely to have had more learning opportunities and the completion of the Bachelor program, which includes two curricular internships, can be taken as an indication that these learning opportunities have been successful, at least to a certain degree. Hence, it can be presumed that participants who are in the Master program have a higher RobS score than participants who are in the Bachelor program. Moreover, there are findings that indicate that the final graduation grade (e.g., *Abi-turnote*) is a suitable predictor for academic success (Blömeke, 2009; Trapmann et al., 2007). Better graduation grades are generally achieved by students who have higher cognitive abilities and are more willing to learn, which are attributes that are valuable in higher education as well (Meyer et al., 2023, p. 181). Thus, the graduation grade is often also used as a proxy for cognitive abilities, despite the rightful criticism on the appropriateness of grades as a reflection of achievement (see, e.g., Ingenkamp, 1977). Moreover, the final graduation grade is also found to be predictive for professional knowledge of EFL teachers in various phases of their studies (Blömeke, Buchholtz, & Bremerich-Vos, 2013, p. 179; Kirchhoff, 2017, p. 140; König et al., 2016, p. 330). Hence, it can be assumed that the graduation grade and the RobS score are negatively¹⁴³ correlated. Additionally, the graduation grade could also be a decisive covariate for knowledge-

¹⁴³ In the German grading system, 1 is the best grade. Thus, if better grades correlate with better test scores, the correlation is negative.

related aspects which should thus also be accounted for in other investigations. Another aspect that should be considered is the participants' university. As both universities represented in the sample should follow the same curricular guidelines stated by the KMK, there should not be a significant difference between the groups. This is necessary to argue that the test score describes the participants' competence independent from their place of study. As the two universities included in the sample have different entry requirements in the form of graduation grades, this aspect should also be controlled for.

Concerning the target domain, it could be assumed that there are correlations between its specific facets and the RobS score. In order to decide on the different aspects of the target domain, the (EFL) teacher competence models previously outlined offer a possible resource. Using the model by Blömeke et al., the competence facet of cognitive dispositions is a potential foundation for situation-specific skills and actual performance (2015). In the model by Baumert & Kunter, the cognitive facet of professional knowledge has also been highlighted as important for teachers' competence (2006). Shulman's taxonomy allows a distinction between at least three different facets of knowledge, PK, PCK and CK (1987, see also Chapter 3.2.1). It thus makes sense to focus on these three knowledge facets in order to estimate the extent to which the results of the RobS relate to the target domain. As previously described, these three facets have often been investigated and correlations between the three facets found (see Chapter 3.4). Considering the broadness of PK and the EFL-specific focus of the RobS, it can be considered sufficient to concentrate on specific pedagogical knowledge (SPK) which is mostly related to the RobS, i.e., feedback (Baumert & Kunter, 2013, p. 35). As CK is equally broad for the field of English, the focus is put on lexical knowledge, using a test whose results might also indicate language proficiency (LP) (see Chapter 15.5.3.2.3). Naturally, EFL-specific PCK should also be accounted for. In order to more specifically assess whether the results of the knowledge tests do correlate with the results in the RobS, the aspect of *phase of study* and *graduation grade* should be controlled for. Both aspects might be covariates, as they might correlate with both the RobS score and the knowledge facets.

Experts are a typical source that is used to collect data in validation studies (M. T. Kane, 2006, p. 28, 2013a, p. 449). Their evaluation of performance and the scoring instrument has been discussed in Chapter 15.3, focusing on the qualitative data to gain insights into the criteria the experts use in their assessment of the performances and their evaluation of the scoring instrument. Another relevant perspective is that of their actual quantifiable rating and ranking of the performance. This quantitative perspective can be used to set a benchmark for the scoring instrument, as reported by (Kulgemeyer & Tomczyszyn, 2015,

p. 121). However, this presupposes that the experts show a high degree of agreement (Kulgemeyer & Tomczyszyn, 2015, p. 122). Regardless, the extent to which the RobS score-based ranking aligns with experts and correlates with experts individually provides relevant information that can be used to discuss whether the scoring instrument is able to describe the level of performance adequately.

15.5.2 Hypotheses

As this study and section of the thesis is strongly focused on the quantitative relationship that exists between the different facets of the RobS, hypotheses will be stated and investigated. The overarching research aims to be explored in this study are detailed in the IUA. H3 relates to RQ13, while H1, H2 and H4 relate to RQ14. The following hypotheses represent the alternative hypotheses that are accepted if the p -value, which can be defined as “the probability that the chosen test statistic would have been at least as large as its observed value if every model assumption were correct, including the test hypothesis” (Greenland et al., 2016, p. 339), is $p \leq .050$. Null hypotheses are not stated explicitly but would indicate that there is no significant difference between groups or correlation with other scores or variables.

H1: The RobS score correlates significantly positively with the amount of school-related experience.

H1a: The RobS score is significantly higher for participants who have completed the School Internship Semester in the Master of Education program, compared to those who have not.

H1b: The RobS score is significantly higher for participants in the Bachelor program who have completed one of the curricular internships, compared to those who have not.

H1c: The RobS score is significantly higher for participants who have extracurricular school-related experience outside of the curricular internships, compared to those who have not.

H1d: The RobS score correlates significantly positively with the number of months participants report to have had as extracurricular school-related experience.

H2: The RobS score correlates significantly positively with the academic progress and success of the participants.

H2a: The RobS score is significantly higher for participants in the Master program, compared to participants in the Bachelor program.

H2b: The RobS score correlates significantly positively with the number of reported semesters of study.

H2c: The RobS score correlates significantly negatively with the reported final graduation grade.

H2d: The RobS score is significantly different for participants from different universities.

H3: The RobS score correlates significantly positively with other facets of the target domain of EFL teacher competence.

H3a: The RobS score correlates significantly positively with the score of the SPK test.

H3b: The RobS score correlates significantly positively with the score of the PCK test.

H3c: The RobS score correlates significantly positively with the score of the CK / LP test.

H3d: The scores of the SPK, PCK and CK / LP tests significantly correlate with each other.

H4: The RobS score correlates significantly positively with the expert ratings.

H4a: Experts' rankings show significant positive agreement amongst each other.

H4b: Experts' rankings and the RobS score-based ranking show significant positive agreement.

H4c: Experts' rankings correlate significantly positively with the RobS score-based ranking.

15.5.3 Methods

In this chapter, the methodological background for the investigation of the hypotheses is presented. Firstly, the collected personal and academic data and approach to analysis is outlined in relation to the hypotheses. Secondly, the selection process for the chosen tests to collect data for H3 is described in detail, including the rating process and psychometrical statistics, as well as the procedure to test H3a-H3d. Thirdly, the collection and analysis of the quantitative expert ratings is described.

15.5.3.1 Personal and Academic Data

Participants reported information on several potentially relevant criteria in the demographic section of the online survey, including their final graduation grade (e.g., *(Fach-)Abiturnote*) (see Appendix C). With regard to their

professionalization process, information was reported on phase of study, e.g., whether students are enrolled in a Bachelor or a Master program and the current number of semesters. Moreover, participants were asked to provide information on their school-related practical experience. It was enquired as to which extent the participants had completed the three curricular internships in the study program (i.e., AOI, OI and SIS, see Chapter 2.2). They were also questioned as to whether they had further school-related practical experience outside of the curricular internships and, if so, asked to provide the number of months. Information on the participants' university is evident by the research process and was not collected by the participants' self-report.

In order to test for H1 and H2, which are based on the demographic and biographical information, different statistical calculations are performed. For H1a-H1c, H2a and H2d, comparisons between groups are made. For H2d, an ANCOVA is additionally performed to account for the graduation grade as a covariate. For parametric data, the *t*-test is chosen. For non-parametric data, Welch's *t*-test and the Mann-Whitney *U* test are performed. For H1d, H2b, and H2c, bivariate correlations are performed. For parametric data, Pearson's *r*, and for non-parametric data, Spearman's ρ is presented.

15.5.3.2 Areas of the target domain

In order to test hypotheses H3a-H3d, bivariate correlations are calculated between the respective score of the SPK, PCK and CK / LP test and the RobS score. Additionally, the correlations among the three tests are also reported as an indication for the extent to which the three tests are related to one another and represent the target domain. Both the results for Pearson's *r* and Spearman's ρ are reported to allow for the comparison of parametric and non-parametric data. In order to account for the potential covariates, partial correlations are also reported that are controlled for phase of study and graduation grade.

In the following, the test instruments used for the measurement of the knowledge facets are presented. The choice of test for the respective aspect of the target domain will be discussed and the psychometric measures presented. This is necessary in order to provide information on the validity of the tests themselves before they are used to provide evidence on discussing the test in question.

15.5.3.2.1 Specific PK: Declarative Feedback Knowledge

In order to measure PK, a test on declarative feedback knowledge was used. As the RobS focuses specifically on feedback performance, it can be assumed that this specific knowledge facet of PK is more likely to correlate with the RobS score than general pedagogical knowledge (GPK) (Baumert & Kunter, 2013, p. 35).

Thus, using a more general test focusing on GPK is not considered and a test focusing on specific pedagogical knowledge (SPK) is favored. Hess et al. (2017) propose a test which consists of two open questions, asking participants to define feedback and name criteria for 'good feedback', which are then coded (2017, p. 18). While the test instrument shows sufficient inter-rater reliability ($\kappa = .86 - \kappa = .97$), it is not considered practical for the present study due to the required extensive rating process of the open items (Hess et al., 2017, p. 21). Thus, another test by Schütze et al. (2017) is considered and ultimately used¹⁴⁴. The test is designed in the context of an intervention study that is focused on measuring the effects of training on formative feedback practice of mathematics teachers in Germany. The items are self-developed, based on the input teachers received in the training intervention. The test is chosen for this study as it also includes aspects used in the feedback framework, especially the model of feedback by Hattie and Timperley (2007) (Besser et al., 2015) (see Chapter 7.3). Furthermore, it was tested with $N = 67$ teachers, has acceptable values of internal consistency (Cronbach's $\alpha = .84$), inter-rater reliability (average Cohen's $\kappa = .78$; range from $\kappa = .56$ to $\kappa = 1.00$), and has an appropriate length and duration for the overall context of the procedure (Schütze et al., 2017, p. 480).

The provided test includes eight items with multiple choice as well open questions. An example item is: "Which three central questions should be answered when providing feedback?" (Schütze et al., 2017, p. 480; translated by TJ)¹⁴⁵. One item that includes examples situated in mathematics instruction is adapted for the EFL context. Two researchers in the area of TEFL were asked to assign the adapted EFL examples to the most appropriate original example to ensure they allowed for the same solution. As TEFL instructors assigned the example as intended, the adapted item is used¹⁴⁶. The test data is evaluated according to the rating manual provided by the authors, with the maximum number of points being 37. The open questions in the test are all rated by three raters (the researcher and two student assistants, who are enrolled in a Master of Education program). The two student raters received an introduction to the rating material and were able to discuss rating based on results from the pilot survey. The overall training took about 90 minutes. To ensure inter-rater reliability, an overall value of Fleiss' κ was determined across all items with $\kappa = .76$ ($p = <.001$),

¹⁴⁴ I would like to thank the authors, especially Birgit Schütze and Michael Besser, for the access and the permission to use the test and the rating manual for this project.

¹⁴⁵ In the study, the items are provided to the participants in the original German version.

¹⁴⁶ The test items and the rating manual were provided under the agreement that they will not be shared. The information presented here is thus restricted to the available publication (Schütze et al., 2017).

and for each of the individual items that had a range from $\kappa = .61$ to $\kappa = .96$ ¹⁴⁷. The values are consistent with the values reported by the authors (Schütze et al., 2017, p. 480), indicating a sufficiently reliable adaptation of the scoring rules and inter-rating reliability. For the final score used for further investigations, the item score attributed by the majority of the three raters is adopted, in case full agreement is not evident, and items scores accumulated for the test score. The internal consistency of the scale is $\alpha = .55$ for the present sample and the provided items. This finding is substantially lower than the reported value of .84 by Schütze et al. (2017, p. 480). Potentially, this discrepancy is the result of the different research contexts, as the test was originally designed for the measurement of an intervention effect for mathematics teachers and used here for PSETs. As the specific knowledge is of a formative nature, and the items indeed cover a variety of different aspects of feedback and formative assessment, a lower alpha value might be acceptable (M. Stadler et al., 2021). Results of this test are thus used, but have to be interpreted with caution.

15.5.3.2.2 EFL-specific PCK: FALKO-E

In order to test for PCK in the domain of TEFL, the PCK-specific part of the FALKO-E test (FALKO-E PCK) is used (Kirchhoff, 2017)¹⁴⁸. The test covers several aspects of the target domain and can be used to differentiate participants' professional knowledge. The decision to use this test is the result of a discussion of three potential tests which all claim to measure TEFL-related PCK in TEFL: TEDS-LT (Blömeke, 2011; Blömeke, Bremerich-Vos, et al., 2013), PKE (König et al., 2016; Roters et al., 2013; Roters et al., 2015; Tachtsoglou & König, 2018) and FALKO-E (Kirchhoff, 2017)^{149,150}.

Tests on EFL-specific PCK: TEDS-LT, PKE and FALKO-E

To provide reason for the chosen test, the three test that were considered are discussed critically. The aim of TEDS-LT is to investigate the professional knowledge of teachers in low-structured domains, such as English, in a (quasi-) longitudinal approach (Blömeke, 2011, p. 7). The researchers developed a pen-

¹⁴⁷ Individual item reliability scores are not reported as this is also not done in the original publication and, thus, no comparison is possible.

¹⁴⁸ I would like to thank Petra Kirchhoff for the permission to use the test and for providing the rating manual for this project.

¹⁴⁹ A comprehensive overview and discussion of these three tests can be found in Marks (2023, pp. 73–92).

¹⁵⁰ Results of the studies in which these tests were used are discussed in Chapter 3.4.

and-paper test which tests CK and PCK. The latter part includes subdimensions that focus on

1. Methods/concepts/analyses
2. Processes/strategies/language acquisition
3. Teaching literature/Interculturality (Jansing et al., 2013, p. 80).

Items are based on the guidelines by the KMK, university curricula, and are validated in expert interviews (Roters et al., 2011, pp. 80–82). The test includes both closed and semi-open items and aims at covering levels of reproduction, analysis and evaluation/creation. A sample PCK item of TEDS-LT is: “Explain the difference between early and late immersion. Give a short answer” (Jansing et al., 2013, p. 89). The PCK part of the test consists of 20 items which are drawn from an item pool of 65 items. Within that pool, 38 items are closed and 27 semi-open. In the study, the semi-open items are double-coded and are reported to have acceptable inter-coder reliability values as Cohen’s κ is reported to have a range from $\kappa = .74$ to $\kappa = .95$ (Jansing et al., 2013, p. 84). The reliability of the PCK dimension is, however, considered insufficient (Jansing et al., 2013, pp. 93–94).

PKE is a continuation of TEDS-LT, but 33 new items for PCK are developed and validated due to the results in TEDS-LT (König et al., 2016, p. 327). The new set includes both closed and semi-open items which can be put into “three content dimensions: knowledge of curriculum, teaching strategies, and students” (König et al., 2016, p. 327). A sample item is: “Provide an example of a lexical item that shows that the term ‘word’ is problematic when talking about vocabulary learning” (König et al., 2016, p. 332). The items do not cover the aspects of teaching literature and culture, although this aspect of the EFL classroom is covered in TEDS-LT (Marks, 2023, p. 81). Cronbach’s α is reported as .72 (König et al., 2016, p. 328) and for inter-rater reliability, the average Cohen’s κ for the double-coded (20 % of the data) semi-open PCK items is reported as “ $M = .75$, $SD = .21$ ” (König et al., 2016, p. 327). Both values indicate sufficient reliability of the test.

The FALKO-E test consists of two parts, the first part focuses on CK and tests knowledge of literature, anglophone culture and linguistics. The second part covers TEFL-related PCK and consists of nine items. Of those nine items, two include both a closed and an open question, while the rest is entirely open. From a content perspective, the items cover EFL aspects of teaching literature and culture, but mostly language as they are designed for teachers of lower secondary classes (Kirchhoff, 2017, p. 122). The FALKO group also categorizes the PCK items according to three knowledge areas: Explain & represent, learner conceptions and the EFL-specific potential of tasks and texts (Kirchhoff, 2017, pp. 119–

120). Many items are designed to approximate practice and thus use brief vignettes, teaching material or learner products, which require the application of knowledge in a written format. Items focus, e.g., on the corrections of lexical items in a vocabulary book or task formulation for text production. A sample item (which does not require extensive context) is the following: “Name as many possibilities as possible to counteract final-obstruent devoicing, a phenomenon which is typical for German L1 learners of English” (Kirchhoff, 2017, p. 129; translated by TJ)¹⁵¹. The initial test and rating manual is reported to have acceptable reliability values as an average of Spearman’s ρ for all items is reported with $\rho = .80$, with a range for the open items from $\rho = .46$ to $\rho = .93$ (Kirchhoff, 2017, p. 151). Only one item is beneath a cited cut-off value by Wirtz and Caspar (2002) for at least moderate agreement of $\rho = .50$, but kept due to reasons of content (Kirchhoff, 2017, p. 132). Döring (2023, p. 559), for example, presents .60 as a benchmark for moderate reliability, which would indicate that one other item of the test has only weak reliability. The internal consistency of the PCK items is stated as $\alpha = .62$, which the authors consider sufficient regarding the broad construct (Kirchhoff, 2017, pp. 132–133).

In the end, the decision for FALKO-E PCK is made on based on two factors. First, the open items allow participants to indicate how they would act in a certain situation and are more competence-oriented than the other two tests, which more frequently use closed items. Second, the items for FALKO-E are publicly available. This facilitates the discussion of the results, and allows future researchers to better replicate the results of this study. These factors outweigh the disadvantages, which are primarily the required resources of time. FALKO-E PCK, due to the open items, requires more time for participants to formulate answers within the procedure. Additionally, the evaluation requires more time as the coding and rating of open items is more complex than that of closed and semi-open ones. Another potential challenge is the reliability of the scoring, as the original FALKO-E PCK reliability coefficients are not as convincing as in TEDS-LT or PKE. This is a limitation that needs to be accepted if the instrument does not rely on closed or semi-open items only. In order to account for the last aspect, extensive steps have been taken to ensure reliability, as is discussed subsequently.

Psychometric Statistics for FALKO-E PCK

With the intention to secure the quality of the rating for the project and ensure inter-rater reliability, the rating manual is adapted and specified. This step is taken to ensure that the reliability scores from the original FALKO-E rating procedure could, at least, be replicated. Also, some aspects in the coding manual

¹⁵¹ Items were provided to the participants in the original German version.

require further clarification. Thus, modifications and specifications are considered for the rating manual. In order to do so, the author (rater 1) and another researcher (rater 2), who also uses the test for validation purposes, cooperated¹⁵². Each researcher rated a subset of $n = 10$ cases of their own, and their cooperation partner's sample. Based on these ratings, ambiguities that the raters had were discussed and an agreement on coding modifications and specifications for the existing manual agreed upon, and documented in an addendum¹⁵³. One important result of the cooperation process which needs to be addressed is a joint decision on the scoring process. The rating manual provides possible item scores of zero, one and two, which leads to a maximum of 22 points. All item scores, with the exception of the closed items, are allocated based on the quantity of correctly named or identified aspects in the respective task, e.g., zero to one correct answers lead to zero points, two to three correct answers lead to one point, and more than three correct answers lead to two points. In Kirchhoff (2017), an "empirical maximum" (p. 135) of 47 points is described, as individual items have varying maximal scores and each individual correctly-named aspect is accumulated for the item score (Kirchhoff, 2017, p. 129). The presented sample item, for example, has a proposed maximum score of "n" (Kirchhoff, 2017, p. 151). The cooperating partners decided to score each item with a maximum of two points, as described in the rating manual. This is done in order to have a strict limit of maximum points and not to overestimate the value of individual items. As the used scores are still based on the sum of correct answers in each item, both potential scores are potentially available for further investigations.

The test is then rated, based on the official rating manual and the specifications laid out in the addendum, by two additional raters (rater 3 and rater 4). These raters are two students in their Master of Education program, with rater 4 not having English as a subject. The raters received a training, which lasted about 120 minutes. In that training, the prospective raters received a printed version of the rating manual and the addendum. Each item was first discussed theoretically, based on the two documents, and subsequently examples from the pilot survey were used to test ratings. After each item was discussed, further instructions were given, which included guidelines such as that the rating should be done item-wise and that after an item was finished, it was to be checked whether the rating rules were applied consistently.

Each rater rated ca. 60 % of the cases (Rater 3: $n = 31$; Rater 4: $n = 30$), with both raters rating the same sample of $n = 10$ cases that is also discussed in the

¹⁵² The researcher agreed to confidentiality and the data was anonymized before transmission.

¹⁵³ As the rating manual for the test is only available upon request, the addendum will also be provided upon request, but not included in this thesis.

previously outlined cooperation, to calculate inter-rater reliability between the four raters. Calculating Spearman's ρ of the average of all items between rater 3 and 4, a value of $\rho = .57$, which is below the benchmark stated by Döring (2023, p. 559) and well below the values reported by Kirchhoff (Kirchhoff, 2017, p. 151) is found. Additionally, Spearman's ρ of the final test scores for the quadruple-rated cases is calculated as an indication for overall reliability between each of the four raters (see Table 67).

Table 67: *Spearman's ρ of the correlation of FALKO-E PCK scores between four raters.*

Spearman's ρ		Rater 1	Rater 2	Rater 3	Rater 4
Rater 2	ρ	.70			
	Sig. (2-tailed)	.026			
Rater 3	ρ	.92	.52		
	Sig. (2-tailed)	< .001	.127		
Rater 4	ρ	.46	.52	.52	
	Sig. (2-tailed)	.178	.130	.125	

While the results indicate sufficient reliability between raters 1 and 2 ($\rho = .70$, $p = .026$) as well as raters 1 and 3 ($\rho = 0.92$, $p < .001$) using Döring's benchmarks (2023, p. 559), the results of rater 4 show low, insignificant coefficient scores. In a consultation, rater 4 stated that a lack of subject-specific knowledge in the field of EFL was challenging in the rating process and likely the reason for the deviating results. As a consequence, the results of rater 4 are dismissed and rater 3 rated the remaining $n = 20$ cases. As a substantial amount of time had passed between the initial training and the second rating, rater 1 rated another five cases of that sample. Spearman's ρ is calculated for the $n = 15$ cases that are double-rated by rater 1 and rater 3. The item-specific reliability coefficients of this sample for raters 1 and 3 are described in Table 68. Two items do not meet the aspired benchmarks, while the overall average of Spearman's ρ which is calculated in tandem with Kirchhoff, shows similar reliability values (FALKO-E: $\rho = .70$, PERFORM-LA: $\rho = .67$), which indicates a sufficing reliability. The correlation between the two raters with regard to the final test score is $\rho = .83$, $p < .001$, which is still acceptable, though lower than the previous value for the ten cases of $\rho = .92$, $p < .001$. As the final test score is the value that is needed for further investigation, the evidence suggests that the final result is generated sufficiently reliably. The results of rater 3 are then used for further investigations. The internal consistency of the scale for the present sample is $\alpha = .65$, which is above the value of $\alpha = .62$ in the original study (Kirchhoff, 2017, p. 132).

Considering the broad construct and the formative nature of domain knowledge, this is an acceptable value (M. Stadler et al., 2021).

Table 68: Spearman's ρ of individual items in FALKO-E PCK between rater 1 and rater 3.

Item ¹⁵⁴	Item name ¹⁵⁵	Spearman's ρ
Falko 1b	Erklärung Lehrbuch Begründung	.46 ($p = .086$)
Falko 2	Interlanguage	n.a. ¹⁵⁶
Falko 3	Vokabelheft	.78 ($p < .001$)
Falko 4	Auslautverhärtung	.47 ($p = .079$)
Falko 5	Textproduktion	.71 ($p = .003$)
Falko 6b	Mediation: Begründung	1.00 ($p = < .001$)
Falko 7	Drama	.66 ($p = .008$)
Falko 8a	Literatur Vorwissen	.65 ($p = .009$)
Falko 8b	Literatur Interkulturelles Lernen	.65 ($p = .009$)
Falko 9	Hörverstehenstest	.61 ($p = .016$)
Average		.67

15.5.3.2.3 CK / LP: The LexTale Test

For the measurement of CK, a test on lexical knowledge that can supposedly also be used to approximate English language proficiency is used: The LexTALE test conceived by Lemhöfer and Broersma (2012). The test is developed as a "5-min vocabulary test" (Lemhöfer & Broersma, 2012, p. 325) in order to allow research to measure the lexical knowledge of advanced learners of English (i.e., adult learners who started learning English in school, but use it regularly) quickly and reliably. The test consists of 60 items, 40 words and 20 nonwords, and participants have to decide whether the item is a proper word from the English language or not. One example of a word is 'denial', while a nonword is, for example 'platery' (Lemhöfer & Broersma, 2012, p. 340). The LexTALE score is then calculated as the average % of correctly identified items:

¹⁵⁴ Closed items 1a and 6a were not double-rated.

¹⁵⁵ Original item names taken from Kirchhoff (2017, p. 151).

¹⁵⁶ Rater 3 rated all cases with zero points which prohibits the calculation of the coefficient in SPSS. Unfortunately, the rating manual is ambiguous for this task regarding a requirement and whether it should be fulfilled *only* to achieve two points or *also* in order to achieve one point. A clear decision was not made in the addendum, as this issue did not arise. This aspect was also not addressed during the coding process. As rater 3 still applied the rule consistently to all cases and the overall correlation of the test scores, which are needed for further investigations, between raters was sufficient, no further recoding was done.

$$\text{LexTALE score} = \frac{\left(\frac{\text{number of words correct}}{40} \times 100 \right) + \left(\frac{\text{number of nonwords correct}}{20} \times 100 \right)}{2}$$

While the test was initially developed for testing lexical knowledge, the authors, in their validation process, used the *Quick Placement Test* (QPT), a language proficiency test, as a means for external validation (Lemhöfer & Broersma, 2012, p. 331). They found that for Dutch learners, there is a substantial correlation between the LexTALE score and the QPT of $r = .63$, $p < .01$. More recent studies trying to replicate the results of the initial validation study, indicate that the extent of the correlation for L1s such as Spanish ($r = .26$, $p < .001$), or Chinese ($r = .36$, $p < .001$) is substantially lower than that of the L1 Dutch (Puig-Mayenco et al., 2023). Puig-Mayenco et al. argue that, potentially, the typological relation of the West Germanic languages of English and Dutch contributes to this high correlation and advise against the use of LexTALE for comparisons of general language proficiency (2023). While the L1 of the participants has not been enquired in this survey, it can be assumed that the participants of the present study are fluent in German as the sample stems from German universities. Hence, as German is also typologically related to English and Dutch, this might indicate that the LexTALE scores might be a suitable approximation for general language proficiency in this study. Additionally, as is shown in the context of the pilot survey with $n = 10$ participants, the LexTALE scores do seem to correlate positively with results of the C-test ($p = .64$, $p = .046$) and C-tests have widely been relied on to account for general language proficiency (Grotjahn, 2002b) (see Chapter 14.2.4). Thus, while the relations are to be viewed with caution, some inference about the language proficiency can be made from the LexTALE test, at least for typologically related languages. Nevertheless, this test does not provide explicit information on participants' language proficiency in a holistic sense, primarily because it only focuses on lexical knowledge and excludes other relevant aspects like proficiency in reading, writing, speaking, listening or even grammar. Such a holistic approach would immensely add to the validity of a score that seeks to make inferences about language proficiency.

The LexTALE test is chosen despite this clear limitation. Measuring language proficiency holistically would also require immense resources in obtaining and scoring the data. The duration of such a test would be substantial and the procedure and the evaluation of open questions and items would likely require additional resources regarding raters and their reliability. The LexTALE test, however, requires only five minutes and, because it is dichotomous in nature, can be directly calculated with the respective syntax in SPSS. Additionally, the LexTALE test is publicly available for free via www.lextale.com, while other, potentially valid language proficiency tests often require additional fees. Furthermore, the

broader context of the study needs to be considered. The CK / language facet is one of three tests, and potentially the furthest removed from the RobS. As the other two tests that focus on SPK and PCK are potentially more relevant in providing evidence for the respective validity argument, they must be prioritized in terms of resources spent.

While lexical knowledge can very well be regarded as an aspect of EFL-related CK, other aspects are not included like linguistic, cultural and literary knowledge. The three tests described in the context of PCK actually have sections focusing on these content-related aspects. In the end, the decision not to include these tests is based on two arguments. First, these studies indicate that there is a substantial significant positive correlation between CK and PCK (Jansing et al., 2013, p. 95; Kirchhoff, 2017, p. 135; König et al., 2016, p. 328) (see also Chapter 3.4). It can thus be assumed that these constructs are likely related. Second, these tests, if used, would again require immense resources with regard to the collection of data during the procedure, as well as in coding. Thus, they are not included in the present study. As a consequence of this discussion, and despite the outlined limitations, the LexTALE test is chosen for the study as one aspect of CK which can potentially be used to make some inferences towards language proficiency. Thus, it is referred to as potentially accounting for both aspects as CK /LP.

15.5.3.3 Expert Ratings

Expert ratings of the performances were collected as part of the expert interviews, as outlined in Chapter 15.3. Experts were asked to rate four selected performances on a global scale with regard to the quality of the feedback conversation, using either the German or US grading scheme (see Table 69).

Table 69: *Grading system for experts' global rating.*

Points	10	9	8	7	6	5	4	3	2	1	0
German Higher Education Grades	1.0	1.3	1.7	2.0	2.3	2.7	3.0	3.3	3.7	4.0	5.0
US Grades	A	A-	B+	B	B-	C+	C	C-	D+	D	F

The videos are always shown in the same order to standardize potential order effects. For their expert study to validate similar role-play-based tests, Kulgemeyer and Tomczyszyn (2015) use a variation of a comparative judgment procedure and have experts watch pairs of videos and decide for the better, as they argue that experts are not able to rate or rank performances (Kulgemeyer & Tomczyszyn, 2015, p. 121). The approach taken in this study is different, in that experts rate each performance individually and have the opportunity to adjust their assessment after watching all videos. As the experts in this case are chosen specifically for their expertise in the RobS-related areas, it can be assumed that

experts are capable of rating the performances appropriately. Also, the number of videos is substantially lower with four, compared to 16 in the study by Kulgemeyer and Tomczyszyn (2015).

Expert ratings are used to investigate the agreement amongst the experts themselves as well as with the RobS score. In the following, the grades experts state in their final evaluation after watching all four videos are used for calculations. Experts' proposed grades are transformed into points to allow for a more apt comparison, as experts used grades either based on the German higher education system or the US grading system. The expert rating is regarded to be ordinality scaled for statistical analyses because the smaller sample size requires non-parametric procedures. The appropriate coefficient for that purpose is Kendall's concordance coefficient W , which can be used to estimate the agreement of ordinal data between multiple raters (A. Field, 2009, p. 576; Wirtz & Caspar, 2002, p. 133). Kendall's W can take values between 0 (no agreement) and 1 (total agreement) (A. Field, 2009, p. 576). Spearman's ρ and Kendall's Tau-b (τ -b) are presented for pairwise rank comparisons. The former considers the ranks as equidistant and takes into account the rank differences, while the latter only uses the ordinal information (Bortz & Lienert, 2008, p. 290). Kendall's τ -b is considered the more conservative coefficient and it also allows for the correction of ties within rankings, while Spearman's ρ considers ties as identical, and is more robust considering potential outliers (Bortz & Lienert, 2008, p. 301). For smaller samples sizes that include ties, Kendall's τ -b is the preferred coefficient (Wirtz & Caspar, 2002, p. 137). Both coefficients will be presented for a holistic picture of the data.

15.5.4 Results

In the following, the results will be presented in the order of the previously outlined hypotheses. If not otherwise indicated, means and standard deviations refer to the RobS score of the respective group.

15.5.4.1 H1: School-related Experience

The descriptive data of the curricular internships is presented in Table 70 as a reference for the discussion of hypotheses H1a and H1b.

Table 70: Participants self-reported data on curricular internships.

	Aptitude and Orientation Internship (AOI)	Occupational Internship (OI)	School Internship Semester (SIS)
Completed	38	36	18
In the phase	0	1	1
Not yet started	12	13	26
No answer	1	1	4
Inconclusive	0	0	2

H1a: The RobS score is significantly higher for participants who have completed the School Internship Semester in the Master of Education program, compared to those who have not.

The data of $n = 45$ participants is included. All participants in the Master program, except one, have finished the SIS and one person indicated they were still in the SIS, although the actual school-phase had finished and they were in the process of writing the report. Thus, the groups that completed the SIS and those who are in it have been combined for statistical calculations as 'In or after the SIS'. Two participants provided inconsistent answers, stating that they had completed the SIS, although reported to be enrolled in the Bachelor program. Thus, these two participants are excluded from calculations that involve the SIS and are labelled 'inconclusive'. The group 'In or after SIS' then, consists of $n = 19$ and achieves an average RobS score of $M = 36.79$ ($SD = 4.77$). The subsample of participants who have not yet started the SIS consists of $n = 26$ PSETs and achieves on average $M = 27.96$ points ($SD = 7.65$). The data is distributed normally for both groups (Shapiro-Wilk test: $p_{\text{No SIS}} = .495$, $p_{\text{In or after SIS}} = .855$) but variances are not equal (Levene's test: $p = .022$). Thus, the more robust Welch's t -test and Mann-Whitney U test are performed to account for non-parametric data. Results of Welch's t -test indicate that the two groups differ statistically significantly, $t_{\text{Welch}}(42.11) = -4.76$, $p < .001$. This is supported by the Mann-Whitney U test, $U = 84.50$, $Z = -3.74$, $r = -.56$, $p < .001$. H1a is thus accepted.

H1b: The RobS score is significantly higher for participants in the Bachelor program who have completed one of the curricular internships, compared to those who have not.

Looking at the Bachelor group only, $n = 21$ participants have done at least one of the two internships ($M = 28.05$, $SD = 8.19$), while $n = 10$ have not ($M = 27.30$, $SD = 7.42$). Variances are equal (Levene's test: $p = .890$) and the data is approximately normally distributed (Shapiro-Wilk test: $p_{\text{no internship}} = .212$, $p_{\text{at least one}}$

internship = .063). Results of the *t*-test do not show a statistically significant difference between the two groups $t(29) = -.24, p = .809$. H1b is thus not accepted.

H1c: The RobS score is significantly higher for participants who have extracurricular school-related experience outside of the curricular internships, compared to those who have not.

With regard to extracurricular school-related experience, $n = 33$ participants state that they do not have any ($M = 30.88, SD = 6.96$), while $n = 18$ do ($M = 31.83, SD = 9.82$). Variances are equal (Levene's test: $p = .350$) and approximately normally distributed (Shapiro-Wilk test: $p_{\text{no extracurricular}} = .291, p_{\text{extracurricular}} = .070$). Results of the *t*-test do not show a statistically significant difference between the two groups $t(49) = -.40, p = .688$. H1c is thus not accepted.

H1d: The RobS score correlates significantly positively with the number of months participants report to have had as extracurricular school-related experience.

The hypothesis is tested for the subsample of participants with experience first, then for the whole sample. Of the 18 participants with experience, one person did not specify the number of months. The remaining $n = 17$ participants have an average of $M = 10.65$ months ($SD = 9.18$) of experience, with the minimum being one month and the maximum being 36 months. The data is not distributed normally regarding the number of months with experience (Shapiro-Wilk test: $p = .023$) and therefore the non-parametric rank-oriented bivariate correlation coefficient is calculated. Results indicate that there is no correlation between the two variables, $\rho = .003, p = .991$. For the whole sample of $n = 50$, excluding the one person with experience who did not answer, the average is $M = 3.63$ months of experience ($SD = 7.32$). The data is not distributed normally regarding the number of months of experience (Shapiro-Wilk test: $p < .001$) and therefore also the rank-correlation coefficient is calculated. Results indicate that there is also no correlation between the two variables, $\rho = .15, p = .285$. H1d is thus not accepted.

15.5.4.2 H2: Academic Progress and Success

H2a: The RobS score is significantly higher for participants in the Master program, compared to participants in the Bachelor program.

Comparing the means descriptively, students in the Bachelor program ($n = 30$) achieve on average $M = 27.82$ points ($SD = 7.82$), students enrolled in the Master program ($n = 21$) achieve on average $M = 36.50$ points ($SD = 4.81$). Within the two groups, the RobS score is distributed normally (Shapiro-Wilk test:

$p_{Bachelor} = .133$, $p_{Master} = .666$) but the variances are not equal (Levene's test: $p = .008$). Hence, the Welch's t -test and the Mann-Whitney U test are performed to account for group differences. Results show that there is a statistically significant difference between the groups. Welch's t -test indicates that, on average, participants in the Master phase achieved 8.69 more points than those in the Bachelor group, 95 % CI [-12.26, -5.13], which leads to a large effect size, indicated by Hedges g , $t_{Welch}(48.93) = -4.90$, $p < .001$, $g = 6.93$. The Mann-Whitney U test corroborates these findings, $U = 119.50$, $Z = -3.68$, $r = -.52$, $p < .001$. The difference between the groups is visually plotted in Figure 40.

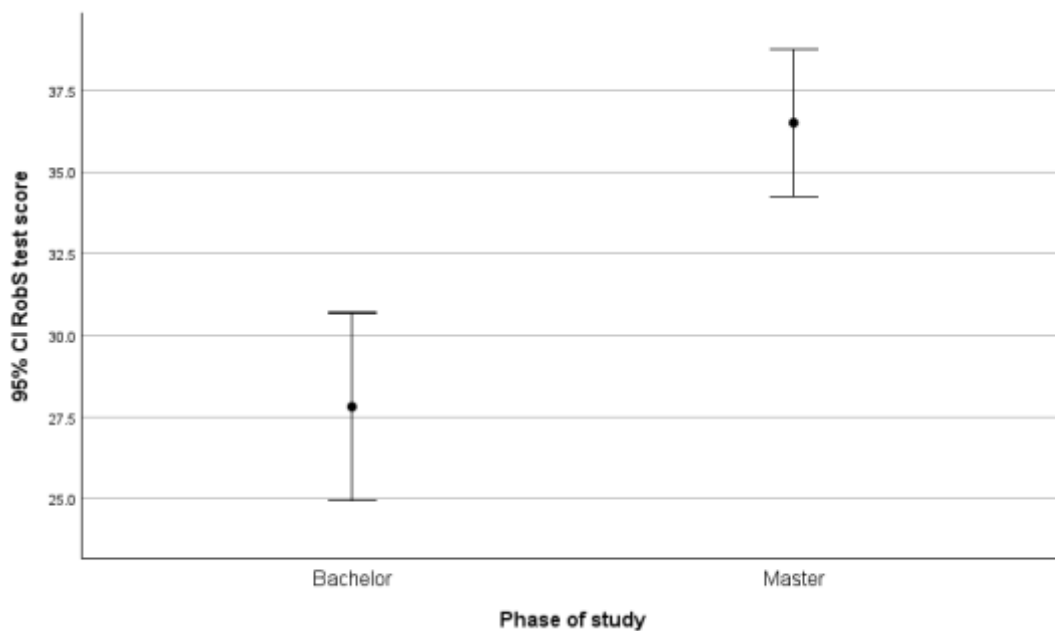


Figure 40: RobS score difference with regard to phase of study.

In order to account for the graduation grade as a potential covariate in these subgroups, a separate analysis was performed that controlled for it. The mean graduation grade in the Bachelor group is $M = 2.55$ ($SD = 0.10$), the median grade is 2.70. In the Master group, the average grade is $M = 2.27$ ($SD = 0.11$), the median grade is 2.40. The data is distributed normally in both groups (Shapiro-Wilk test: $p_{Bachelor} = .08$, $p_{Master} = .235$) and variances are also equal (Levene's test: $p = .464$). The results of the t -test show that there is no statistically significant difference between the two phases of study with regard to their grades, thus it is not controlled for in a separate investigation, $t(49) = 1.87$, $p = .067$. H2a is thus accepted.

H2b: The RobS score correlates significantly positively with the number of reported semesters of study.

The average number of semesters is $M = 9.40$ ($SD = 4.25$), the minimum is one, the maximum number of semesters is 19. The data is distributed approximately normally (Shapiro-Wilk test: $p = .062$). The number of semesters correlates positively and significantly with the RobS score, $r = .47$, $p < .001$. H2b is thus accepted.

H2c: The RobS score correlates significantly negatively with the reported final graduation grade.

The average graduation grade is $M = 2.44$ ($SD = 0.54$), the median is 2.50, with the grades being better, the lower they are in the German grading system. The best grade in the sample is reported as 1.20, the lowest grade 3.70. The data for the graduation grades is distributed approximately normally (Shapiro-Wilk test $p = .093$). Results show that there is a significant negative correlation between graduation grade and the RobS score which is moderate, indicating that participants with better grades do also get more points in the RobS, $r = -.44$, $p = .001$. H2c is thus accepted.

H2d: The RobS score is significantly different for participants from different universities.

In the sample, the $n = 45$ participants from Paderborn University receive, on average, $M = 30.33$ points ($SD = 7.90$) and the $n = 6$ participants from the University of Münster receive $M = 37.82$ points ($SD = 5.60$). The data in both groups is distributed approximately normally (Shapiro-Wilk test: $p_{Paderborn} = .080$, $p_{Münster} = .778$) and the variances are equal (Levene's test: $p = .235$). Results of the t -test show that the groups differ significantly from each other, $t(49) = -2.24$, $p = .030$. As the two universities have different entry requirements in the form of the graduation grade, this is controlled for. Participants from Paderborn University have an average grade of $M = 2.52$ ($SD = .50$), the median is 2.50, those from the University of Münster have an average grade of $M = 1.80$ ($SD = 0.43$), the median is 1.70. The data is distributed approximately normally (Shapiro-Wilk test: $p_{Paderborn} = .062$, $p_{Münster} = .506$) and the variances are equal (Levene's test: $p = .430$). Results of the t -test show a statistically significant difference in the graduation grade of the two groups, $t(49) = 3.37$, $p = .001$, $d = .49$. Thus, a one-way ANCOVA is performed to control the differences between universities for the graduation grade as a covariate. Obviously, the assumption of independence of the covariate and group variables (university) is violated, as shown by the t -test. However, this assumption refers to experimental conditions in which the groups refer to different treatments (A. Field, 2009, p. 397). For the present study, the ANCOVA is performed to provide a clearer picture of the relationship

of the different universities and RobS score, knowing that the average graduation grade is different in these groups. The second assumption, the homogeneity of regression slopes, however, is relevant as it indicates that the relationship between the graduation grade and test score is similar in both groups. The interaction between covariate (graduation grade) and group variables (university) is not significant, indicating that the assumption of homogeneity of regression slopes is not violated, $F(2, 47) = 0.17, p = .680$. The one-way ANCOVA indicates that if the graduation grade is included, there is no statistically significant difference between the groups, $F(1, 48) = 1.00, p = .323, \eta^2 = .020$. A post-hoc power analysis with G*Power (Version 3.1.9.7) indicates sufficient power with the reported ANCOVA values (99.92 %). H2d is thus not accepted.

15.5.4.3 H3: Facets of EFL Teacher Competence

Due to the more extensive nature that the consideration of three different tests requires, the results will first be presented descriptively before being used to test the hypotheses.

15.5.4.3.1 Descriptive Statistics, Normal Distribution and General Inferential Statistics

In the following, descriptive results of the test results are displayed (see Table 71) and discussed in terms of normal distribution, as a prerequisite for parametric procedures. As the correlations of RobS with the external tests will also be controlled for phase of study and graduation grade, the respective group differences and bivariate correlations are also presented. Further tabular overviews on the discussed statistics can be found in Appendix H.6.

Table 71: Descriptive statistics of the external tests.

	<i>n</i>	<i>M</i>	<i>SE</i>	<i>SD</i>	95 % CI		Median	<i>Min</i>	<i>Max</i>
					lower	upper			
SPK	51	23.80	0.75	5.34	22.30	25.30	23	12	32
PCK	51	6.30	0.46	3.26	5.38	7.21	5	0	18
PCK (without outlier)	50	6.06	0.40	2.83	5.26	6.86	6	0	12
CK / LP	51	79.49	1.33	9.52	76.80	82.16	80	55.00	100.00

For SPK, declarative feedback knowledge, the overall arithmetic mean is $M = 23.80$ points ($SD = 5.34, Min = 12, Max = 32$). The data is distributed approximately normally (Shapiro-Wilk test: $p = .059$). Students that are enrolled in

a Bachelor program¹⁵⁷ achieve on average $M = 22.35$ points ($SD = 5.54$), students in the Master program $M = 26.05$ points ($SD = 4.24$). As the data in these groups is distributed normally (Shapiro-Wilk test $p_{Bachelor} = .176$, $p_{Master} = .333$) and variances are equal (Levene's test $p = .135$) a t -test can be performed. The difference between these groups is statistically significant and quite substantial, $t(49) = -2.54$, $p = .014$, $d = 5.08$. Looking at the correlation of SPK with graduation grade, a statistically significant correlation is evident, $r = -.31$, $p = .026$.

For PCK, measured using FALKO-E PCK, the mean is $M = 6.06$ ($SD = 2.83$, $Min = 0$ ¹⁵⁸, $Max = 18$). The data is not distributed normally (Shapiro-Wilk test: $p = .040$). Bachelor students achieve on average 5.45 points ($SD = 2.84$), Master students on average $M = 7.6$ points ($SD = 3.51$). Levene's test indicates that variances between the two groups are equal ($p = .566$) but not distributed normally (Shapiro-Wilk test: $p_{Bachelor} = .711$, $p_{Master} = .010$). Looking at the boxplot in Figure 41, an outlier can be identified as a participant achieved 18 points.

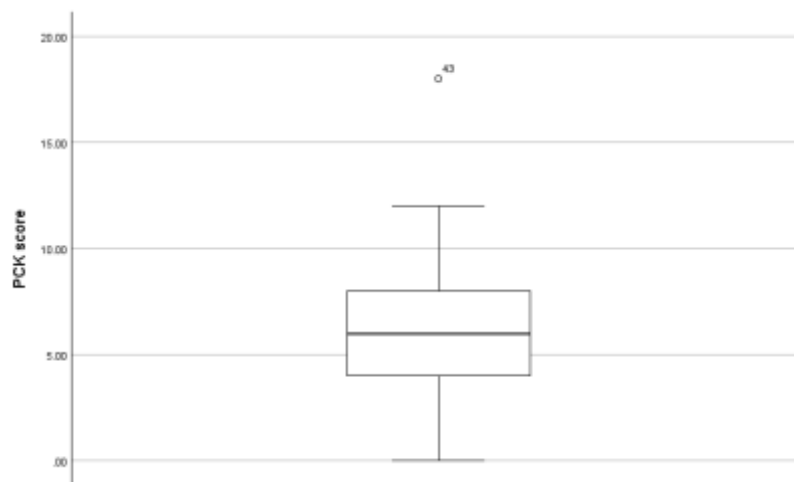


Figure 41: Boxplot of the PCK scores.

Exclusion of that outlier would result in approximately normally distributed data for the full sample ($p = .582$) as well as the subsample ($p_{Master} = .091$). For further calculation, the outlier is however not excluded, instead, both the Mann-Whitney U test for non-parametric data, as well the t -test for parametric data, are performed to compare PCK results between groups of phase of study. While the Mann-Whitney U test is the appropriate test as the data is not parametric, the t -test can be considered "robust" despite violations of normal distribution, as long as the variances between the groups are homogenous, which is

¹⁵⁷ As participants came from two different universities, both the Bachelor of Education as well as polyvalent Bachelor programs are subsumed.

¹⁵⁸ The participant with 0 points has processed the test but did not achieve points for their answers.

corroborated by the results of Levene's test (Bortz & Schuster, 2010, p. 122). The results of the Mann-Whitney U test do not show a statistically significant difference between the groups, $U = 211$, $Z = -1.92$, $r = .27$, $p = .055$. The t -test does show a significant difference between the groups $t(49) = -2.40$, $p = .020$, $d = 3.11$. Looking at the correlation with graduation grade the non-parametric rank-oriented coefficient shows a significant correlation, $\rho = -.31$, $p = .025$.

For CK / LP, the overall arithmetic mean is $M = 79.49\%$ ($SD = 9.52$, $Min = 55\%$, $Max = 100\%$). The data is distributed normally (Shapiro-Wilk test: $p = .948$). Students that are enrolled in a Bachelor program achieve on average $M = 78.79\%$, students in a Master program $M = 80.56\%$. The data is distributed normally (Shapiro-Wilk $p_{Bachelor} = .851$, $p_{Master} = .908$) and variances are equal (Levene's test: $p = .061$). The t -test shows no significant difference between the groups, $t(49) = -3.69$, $p = .522$. Looking at the correlation with graduation grade, no significant correlation is evident, $r = -.15$, $p = .295$.

15.5.4.3.2 Inferential Statistics

In order to test hypotheses H3a-H3d, bivariate correlations are calculated between the results of the knowledge tests and the RobS. Additionally, the correlations among the three tests are also reported as an indication for the extent to which the three tests are related to one another (H3d). Both the results for Pearson's r as well as Spearman's ρ are reported (see Table 72 and Table 73). As the data for the PCK test does not show normal distribution, the requirements for Pearson's r are violated and the rank coefficient ρ has to be used. In order to account for potential covariates, partial correlations are also reported that are controlled for phase of study (see Table 74 and Table 75) and graduation grade (Table 76 and Table 77). In order to provide context on the robustness of the results, a power analysis is performed with G*Power (Version 3.1.9.7). Using the established level of $\alpha = .05$ and $\beta = 20\%$ (Döring, 2023, p. 657), correlations can be assumed to be robust if r or $\rho > .38$.

H3a: The RobS score correlates significantly positively with the score of the SPK test.

The RobS score correlates significantly positively to a medium effect size with the SPK score, $r = .45$, $p < .001$ (J. Cohen, 1988, p. 80). When performing partial correlations and controlling for phase of study, SPK still correlates significantly positively with the RobS score, but less so, $r = .33$, $p = .018$. Similarly, SPK correlates significantly when controlled for graduation grade to a medium extent with the RobS score, $r = .36$, $p = .009$. H3a is thus accepted.

H3b: The RobS score correlates significantly positively with the score of the PCK test.

The RobS score correlates significantly positively to a medium effect size with the PCK score but not necessarily robust, $\rho = .37, p = .008$ (J. Cohen, 1988, p. 80). When performing partial correlations and controlling for phase of study, PCK does not correlate statistically significantly with the RobS score when considering the rank-oriented correlation, which is to be preferred mathematically, $\rho = .28, p = .052$. When controlled for the graduation grade, PCK remains statistically marginally significant, $\rho = .28, p = .050$, but not robust and only to a small extent (J. Cohen, 1988, p. 80). H3b is thus only partially accepted. It holds true for the uncontrolled correlation and when controlled for graduation grade, but not if controlled for phase of study.

H3c: The RobS score correlates significantly positively with the score of the CK / LP test.

There is no significant positive correlation between the RobS score and the CK / LP score, $r = .20, p = .153$. This fact is consistent when controlling for phase of study ($r = .18, p = .203$) and graduation grade ($r = .16, p = .282$). H3c is thus not accepted.

H3d: The scores of the SPK, PCK and CK / LP tests significantly correlate with each other.

A pairwise comparison of the three tests shows that they all correlate significantly. CK / LP correlates to a medium extent with both SPK, $r = .40, p = .004$, and PCK, $\rho = .39, p = .005$, while SPK and PCK correlate with each other to a large extent, $\rho = .71, p < .001$ (J. Cohen, 1988, p. 80). CK / LP, SPK and PCK continue to correlate similarly and significantly, even when controlled for phase of study or graduation grade (see Table 74 and Table 75). H3d is thus accepted.

Table 72: Pearson's r of the correlation between RobS, SPK, PCK and CK / LP.

Pearson's r		RobS	SPK	PCK	CK / LP
SPK	r	.45			
	Sig. (2-tailed)	< .001			
PCK	r	.41	.68		
	Sig. (2-tailed)	.003	< .001		
CK / LP	r	.20	.40	.40	
	Sig. (2-tailed)	.153	.004	.003	

Table 73: Spearman's ρ of the correlation between RobS, SPK, PCK and CK / LP.

Spearman's ρ		RobS	SPK	PCK	CK / LP
SPK	ρ	.42			
	Sig. (2-tailed)	.002			
PCK	ρ	.37	.71		
	Sig. (2-tailed)	.008	< .001		
CK / LP	ρ	.23	.38	.39	
	Sig. (2-tailed)	.11	.006	.005	

Table 74: Pearson's r controlled for phase of study of the correlation between RobS, SPK, PCK and CK / LP.

Controlled for phase of study	Pearson's r		RobS	SPK	PCK	CK / LP
	SPK	r	.33			
		Sig. (2-tailed)	.018			
	PCK	r	.29	.64		
		Sig. (2-tailed)	.040	< .001		
	CK / LP	r	.18	.40	.40	
		Sig. (2-tailed)	.203	.005	.004	

Table 75: Spearman's ρ controlled for phase of study of the correlation between RobS, SPK, PCK and CK / LP.

Controlled for phase of study	Spearman's ρ		RobS	SPK	PCK	CK / LP
	SPK	ρ	.31			
		Sig. (2-tailed)	.031			
	PCK	ρ	.28	.68		
		Sig. (2-tailed)	.052	< .001		
	CK / LP	ρ	.21	.37	.38	
		Sig. (2-tailed)	.152	.008	.007	

Table 76: Pearson's r controlled for graduation grade of the correlation between RobS, SPK, PCK and CK / LP.

Controlled for graduation grade	Pearson's r		RobS	SPK	PCK	CK / LP
	SPK	r	.36			
		Sig. (2-tailed)	.009			
	PCK	r	.30	.64		
		Sig. (2-tailed)	.034	< .001		
	CK / LP	r	.16	.38	.38	
		Sig. (2-tailed)	.282	.007	.007	

Table 77: Spearman's ρ controlled for graduation grade of the correlation between RobS, SPK, PCK and CK / LP.

Controlled for graduation grade	Spearman's ρ		RobS	SPK	PCK	CK / LP
	SPK	ρ	.34			
		Sig. (2-tailed)	.016			
	PCK	ρ	.28		.68	
		Sig. (2-tailed)	.050		< .001	
	CK / LP	ρ	.21	.37	.38	
		Sig. (2-tailed)	.154	.008	.007	

15.5.4.4 H4: Expert Ratings

Data from the expert ratings will be presented descriptively first. An overview of expert global rating according to grades, converted into a ten-point scheme is provided in Table 78. Ten points refers to the highest grade (1.0/A), zero points to the lowest grade, indicating a failed performance (5.0/F) and one point (4.0/D) is the lowest possible grade to pass. In the following, only point scores will be discussed. BNP09 has a point mean of 5.6 ($SD = 3.57$), achieving between 1 and 9 points. MMM07 has a point mean of 5.0 ($SD = 3.08$), achieving between zero and eight points. NDA16 has a point mean of 1.2 ($SD = 1.67$), achieving between zero and three points. CML16 has a point mean of 6.6 ($SD = 1.30$), achieving between four and eight points. BNP09 and MMM07 present a range of eight points, CML16 of four points and NDA16 of three points in the expert rating. NDA16 is ranked last by four experts, and second-to-last by expert 4, resulting in a mean rank of 1.2. MMM07 is the only participant that is placed at every rank at least once, and second-best twice, resulting in a mean rank of 2.7. With a slightly better mean rank of 2.8, BNP09 is rated best twice, and second-to-last three times. The highest mean rank is achieved by CML16 with 3.3. That performance is rated second-best four times and best once. The individual RobS scores all fall within the range of the experts' ratings. Comparing the RobS score ranking with the experts' mean ranked, there is only an agreement in that NDA16 is ranked as the lowest. There is no other agreement on ranking.

Table 78: Ratings of the performances based on grades.

Participant	Scoring	E1	E2	E3	E4	E5	Experts overall		RobS score
							Mean	Mean Ranked	
BNP09	Points	6	1	9	9	3	5.6 (<i>SD</i> = 3.57)	3	3
	Rank ¹⁵⁹	2	2	4	4	2	2.8		2
MMM07	Points	8	7	5	0	5	5.0 (<i>SD</i> = 3.08)	2	8
	Rank	3	4	2	1	3	2.7		4
NDA16	Points	0	0	2	3	1	1.2 (<i>SD</i> = 1.67)	1	0
	Rank	1	1	1	2	1	1.2		1
CML16	Points	8	4	8	7	6	6.6 (<i>SD</i> = 1.30)	4	4
	Rank	3	3	3	3	4	3.3		3

H4a: Experts' rankings show significant positive agreement amongst each other.

Results¹⁶⁰ indicate that the agreement between the experts is significant and moderate as $W = .502$, $\chi^2(3) = 7.531$, $p_{\text{exact}} = .047$ (R. C. Schmidt, 1997, p. 767). H4a is thus accepted.

H4b: Experts' rankings and the RobS score-based ranking show significant positive agreement.

Adding the RobS score-based points to the matrix to estimate Kendall's W yields slightly higher, but still moderate agreement values as $W = .517$, $\chi^2(3) = 9.305$, $p_{\text{exact}} = .017$ (R. C. Schmidt, 1997, p. 767). H4b is thus accepted.

H4c: Experts' rankings correlate significantly positively with the RobS score-based ranking.

An overview of the rank-oriented correlations is depicted in Table 79 and Table 80. The only significant correlation is evident between the RobS score and expert 2, which is also the only perfect rank agreement, $\rho = 1.00$, $p \leq .010$ ¹⁶¹; $\tau\text{-b} = 1.00$, $p \leq .010$. The identical correlations between expert 1 and the RobS score, as well as expert 1 and expert 5 are marginally insignificant with regard to Spearman, $\rho = .95$, $p = .051$, and less significant with the more conservative Kendall, $\tau\text{-b} = .91$, $p = .071$. The remaining correlations are all not statistically significant.

¹⁵⁹ Rank means according to SPSS for the calculation of Kendall's W . A higher rank means a better performance.

¹⁶⁰ For calculation in SPSS, participants are considered as variables, i.e., columns, contrary to the depiction in Table 78 (Wirtz & Caspar, 2002, p. 150).

¹⁶¹ The SPSS output does not provide the exact level of significance.

The rating of expert 3 correlates positively with all other rankings considering ρ , but does so to a smaller extent, with the exception of expert 4, where there is a higher correlation ($\rho = .80$, $p = .200$). These correlations are generally smaller considering Kendall's τ -b. The ranking of expert 4 only correlates positively with expert 3, negatively with expert 1, expert 2, and the RobS score, and not at all with expert 5. The ranking of expert 5 correlates more highly with the RobS score, expert 1, and expert 2, but less with expert 4 and not with expert 3. It can be observed that there are two camps of positive correlations, as experts 1, 2 and 5 as well the RobS score tend to correlate more highly on the one hand, while expert 3 and 4 do so on the other hand. Bivariate correlations of the experts' mean points and the RobS score are not significant, $\rho = .40$, $p = .600$; $\tau = .33$, $p = .497$.

Table 79: Spearman's rank-oriented correlations (ρ) between experts and the RobS score.

		RobS	E1	E2	E3	E4	E5
E1	ρ	.95					
	Sig. (2-tailed)	.051					
E2	ρ	1.00**	.95				
	Sig. (2-tailed)	$\leq .010$	.051				
E3	ρ	.20	.32	.20			
	Sig. (2-tailed)	.800	.684	.800			
E4	ρ	-.40	-.21	-.40	.80		
	Sig. (2-tailed)	.600	.789	.600	.200		
E5	ρ	.80	.95	.80	.40	.00	
	Sig. (2-tailed)	.200	.051	.200	.600	1.000	

Table 80: Kendall's τ -b rank-oriented correlations between experts and the RobS score.

		RobS	E1	E2	E3	E4	E5
E1	τ -b	.91					
	Sig. (2-tailed)	.071					
E2	τ -b	1.000**	.91				
	Sig. (2-tailed)	$\leq .010$	.071				
E3	τ -b	.00	.18	.00			
	Sig. (2-tailed)	1.000	.718	1.000			
E4	τ -b	-.33	-.18	-.33	.67		
	Sig. (2-tailed)	.497	.718	.497	.174		
E5	τ -b	.67	.91	.67	.33	.00	
	Sig. (2-tailed)	.174	.071	.174	.497	1.000	

H4c is thus only accepted for the correlation of expert 2 and the RobS score. There are no further significant correlations for experts individually as well as their mean and the RobS score.

15.5.5 Discussion

In the following, the results will be discussed as they relate to the hypotheses. The extent to which the results support the validity argument of the RobS is done separately in Chapter 17.

H1: The RobS score correlates significantly positively with the amount of school-related experience.

Looking at the RobS score in relation to curricular school-related experiences, only the differentiation regarding the completion of the SIS seems to lead to substantial differences in the RobS score, while the completion of the internships in the Bachelor do not. The AOI and OI, are in nature, less focused on gaining experience in teaching. The OI is itself more focused on providing students insights into areas of the profession that are not necessarily at school and the AOI, while taking place at a school, focuses on observations and not necessarily gaining extensive experience in teaching. Also, the internships are relatively short as they last about four to five weeks each. Possibly, the lack of direct interaction with learners and the short amount of time is a reason that there are no significant differences between the groups with regard to the RobS score. The SIS, in contrast to AOI and OI, requires more actual teaching and thus, interaction with learners, and is also much longer, lasting about five months. This might be one reason why the group that has completed the SIS performs significantly better than the group that has not.

Looking at the extent to which the extracurricular school-related experience correlates with the RobS score, no significant results are found. This finding is interesting considering that the group of participants that have completed the SIS seems to perform better. Potentially, the aspect of mentoring comes into play here as the extracurricular school-related experience, e.g., work as a substitute teacher, is not necessarily supervised and reflected upon, while this is the case in the SIS (S. Gerlach et al., 2023, p. 425). Arguably, appropriate mentoring and reflection on action is necessary to properly develop competences. Consequently, the underlying assumptions might have to be reconsidered as the quality of experience might be a factor as well. As no further description of the experience was asked for in the survey, it cannot be discussed in what kind of activities the participants engaged in and to what extent this might factor in. This must be taken into account for the context of these findings, as experience that is not directly related to instruction, e.g., in an afternoon-care context, is still

included here although it is not necessarily related to the focus of the RobS. Also, experience that is not school-related was not explicitly asked for in the survey, potentially excluding relevant types of experience such as tutoring or coaching.

H2: The RobS score correlates significantly positively with the academic progress and success of the participants

With regard to academically related factors, the RobS score is significantly higher for students in the Master than the Bachelor program and it also correlates positively with the overall number of semesters. Consequently, it can be assumed that in the course of the first phase of teacher education, which covers both Bachelor and Master phases, PSETs acquire the necessary capabilities to successfully master the RobS. This observation is, however, complicated by the fact that there is substantial overlap between the group of Master students and those who have completed the SIS: With the exception of one participant, all students in the Master group have also completed the SIS. As these two groups are almost identical, it is not possible to ascribe the significantly better scores to either the advanced phase of study or the SIS. It is quite likely that both facts contribute to the higher average test score, as students in the SIS might be able to apply the theoretical knowledge they have acquired at university in practice and reflect on the experience, leading to better performance. Future studies could potentially examine if the SIS does in fact have an effect on the performance in the RobS via pre-/post experimental designs in order to get more insight into the sources of these findings.

The final graduation grade, which can be used as a proxy for cognitive abilities and willingness to learn, also significantly correlates with the RobS score. This finding makes sense as the attributes associated with high graduation grades are also necessary for success in higher education. The graduation grade is, however, also a potential covariate, which might confound other results. As the different phases of study do not show a significant difference in their mean graduation grade it is not controlled for here. The university from which the participants stem does not seem to be a significant factor for RobS scores, as controlling for graduation grade indicates here that there are no significant differences between the groups. The ANCOVA which is performed to generate these findings does not, however, comply with the usual necessary assumptions. Especially the assumption of "independence of the covariate and the treatment effect" (A. Field, 2009, p. 397) is logical in the context of comparing experimental groups, where the assumption should be that the covariate is independent from the group. As the assumption in this study, however, is that covariate and group are not independent, this assumption should be disregarded. Accordingly, the

ANCOVA indicates that the graduation grade is the dominant variable and the university is not a reason for group differences.

H3: The RobS score correlates significantly positively with other facets of the target domain of EFL teacher competence.

The results of the three tests will first be explored individually, before a holistic perspective is taken. With regard to SPK as indicated by the score of the declarative feedback knowledge test, the results consistently show a significant correlation with the RobS score, even when controlled for phase of study or graduation grade. The extent of correlation is also the highest when compared to the correlation coefficients of the other tests. Potentially, the content of the test is most specific and also most closely related to the content of the RobS. This is further supported by the fact that several items of the test refer directly to the influential work of Hattie and Timperley (2007), which is an important pillar of the construct and scoring of the RobS.

The results for the PCK test do not allow for such a straightforward interpretation. The direct correlation leads to significant results, although smaller and less robust compared to SPK, controlling for covariates further complicates the picture. While the utilized test focuses on aspects of subject-specific pedagogical knowledge, it is potentially further removed from the specific construct that is tested in the RobS. Some items can be considered adjacent to the feedback scenario, for example the item focusing on corrections of lexical items, but all-in-all, the test does not include items that are directly focused on feedback or the specific writing process. As the items in FALKO-E PCK are further removed from the RobS, this might explain diminished correlation compared to the SPK test. When controlling for phase of study, using the more robust p coefficient due to violation of normal distribution of the PCK results, there is no significant correlation with the RobS score ($p = .052$). When comparing the two groups (Bachelor and Master) with regard to their PCK score, the results show that the groups are not significantly different, as indicated by the Mann-Whitney U test ($p = .055$). Hence, these results do not allow for any clear indication to be made regarding the relation of PCK, phase of study and RobS. Similarly, by controlling for graduation grade, results remain only marginally significant ($p = .050$). The direct correlation of PCK and graduation grade shows a significant correlation between the two test scores. Arguing more conservatively, it seems that the results are not robust enough to draw conclusions about whether the results of the PCK score and the RobS score actually do correlate.

The results of the CK test do not show any correlation with the RobS score. Arguably, the test is appropriately sensitive for the group as descriptive results ($M = 79.49\%$, $SD = 9.52$) align with other studies that have used the LexTALE

test for German L1 learners of a similar age group, e.g., $M = 71.3\%$ ($SD = 11.7$) (Hopp et al., 2020, p. 73) or $M = 78.0\%$ ($SD = 12.4$) (Jackson & Hopp, 2020, p. 898), and $M = 75.5\%$, ($SD = 12.5$) for Dutch L1 learners in the original publication by Lemhöfer & Broersma (2012). Hence, lack of correlation is most likely not based on the LexTALE test itself, but probably due to the fact that its content is the furthest removed from the focus of the RobS.

Looking at the three knowledge tests together, a consistent statistically significant correlation is evident for all tests among themselves. Effect sizes for correlations between CK / LP and PCK and CK and SPK are medium, while it is large for SPK and PCK (J. Cohen, 1988, p. 80). The effects persist even when controlled for phase of study and graduation grade. The results are in line with previous studies like TEDS-LT, PKE or FALKO, as discussed in Chapter 3.4 and corroborate the findings in that CK, SPK and PCK are presumably three facets of the EFL teacher's professional knowledge and part of the target domain, i.e. EFL teacher competence. Interestingly, the tests for SPK and PCK correlate to a much larger extent with each other than with CK / LP. This fact supports the notion that the further away the content of test is with regard to the performance in the pedagogical context, the less related the constructs are.

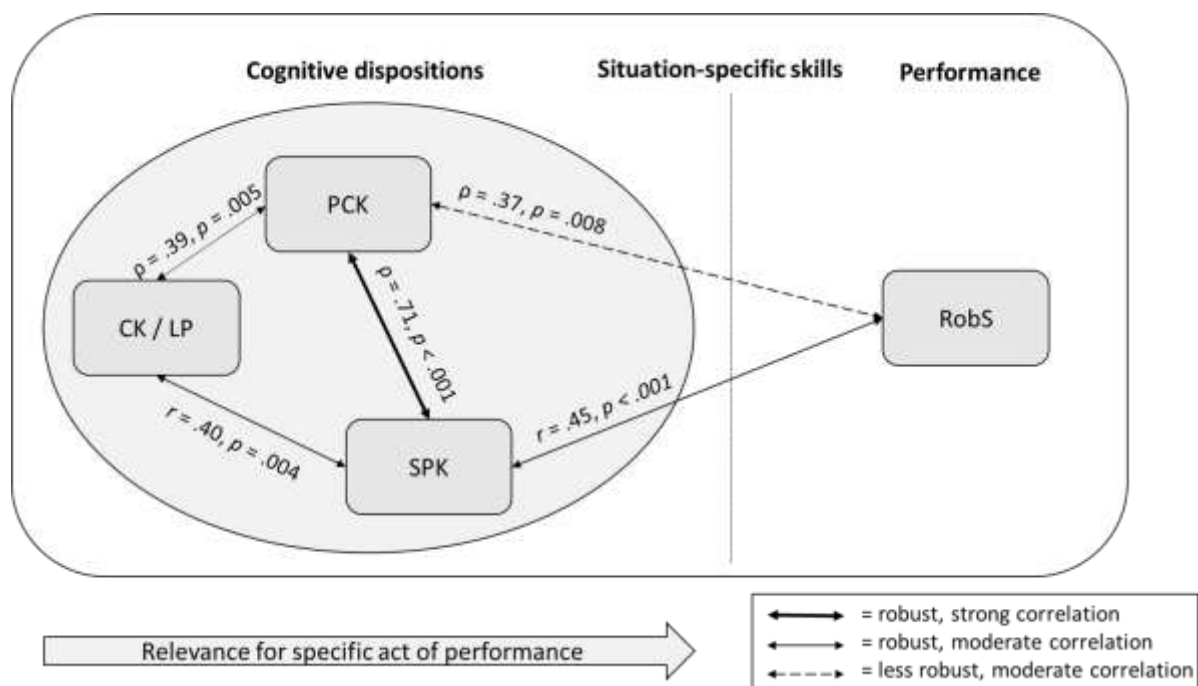


Figure 42: Visual representation of the correlation results for H3.

Looking at all the results, only the SPK test shows a robust correlation with the RobS score, while the results of PCK are less robust and the CK / LP test does not correlate with the RobS. Apparently, the extent to which the focus and content of the specific tests relate to the RobS is indicative of the extent of the

correlation. The SPK test directly focuses on feedback and shows a robust correlation with the RobS. The results of the PCK test, although tentatively correlated, is not specifically connected to the construct of the RobS and thus does not show the same robustness, especially when controlled for cognitive abilities and study progress. The CK / LP test is most removed from the RobS and thus shows no correlation. Comparing SPK, PCK and CK / LP, similar observations can be made as the two tests with a pedagogical component have a large correlation, while CK-related correlations are only medium. The correlation of CK / LP and PCK reflect the extent to which both dimensions correlate in the TEDS-LT study, in which there is a correlation of $r = .395$ between PCK and linguistic CK (Roters et al., 2011, p. 95). Figure 42 visualizes the correlations between the four tests and situates them within the model of competence described by Blömeke et al. (2015). Potentially, path models could offer interesting insights into the question of cause and effect, as CK, in the form of linguistic knowledge or language proficiency, might be a requirement for (S)PK and PCK, which, depending on the specificity of the knowledge, are conditions for successfully mastering specific acts of performance. However, such an analysis is beyond the scope of this thesis. What remains unclear, is the role of situation-specific skills which might have a mediating factor as well. However, no data was generated to account for that dimension of competence and thus no inferences can be made.

H4: The RobS score correlates significantly positively with the experts' ratings.

The agreement between experts can be classified as moderate, using Kendall's *W* (R. C. Schmidt, 1997, p. 767). The descriptive data, however, also reveals that there is a great extent of variability between expert ratings, as there is an eight-point difference in the ratings of BNP09 and MMM07. Expert 4, for instance, is the only one who rated MMM07 with zero points, i.e., let that person fail, while other experts rated that performance consistently higher. Looking at the correlation between experts, there are no significant correlations. However, fewer significant results are not an unusual circumstance, considering the relatively few cases of rated objects. While not statistically significant, the remaining results still allow for some indications, taking into account the other results of the expert interviews, as stated in Chapter 15.3.

The rating of expert 4 correlates negatively with that of experts 1, 2 and 5, showing only a positive correlation with the rating of expert 3. Most likely, this contrary ranking is based on the anomaly of the rating of MM07. Considering the evaluative judgements, expert 4 strongly focuses on a lack of specificity and the participant's inability to use appropriate meta-language as reasons for that grade:

E: I think I would fail that student. Because he does not have a meta-language to talk about text, and so, he just simply talks about THINGS or other students (laughs). The student is going away with absolutely no idea, in my view, of what exactly / at the end I think she says, and I should look and see exactly what to improve, yeah, she should find out exactly what to improve because he was not specific, and I find his explanations incomprehensible. (E4)

Arguably, the expert bases their assessment on a restricted set of criteria. Other experts did evaluate the participant more positively as they also consider other aspects, while still differentiating the participants lack of specificity, as expert 5 who rates the performance with five points:

E: He was much more at a higher level and reached more information and grammar. There was a touch more dialogue and checking in with the student. He had far less specific content than the first one but I think most of that specific content just went in and out of the students ears of the first one. And he gave also, so where to next, in terms of looking at the back and how you use it particularly about the irregular verbs (inaudible) much more specific about what she should do to improve where she went to next. He was clear about the ending was too short but the beginning was pretty good in the way she had good arguments moving into the ending. So, more specific, a much better balance between a higher levels and lower level of information. (E5)

As highlighted by the contrast of these two experts, and corroborated by the inferential results, there might not be sufficient consensus yet among experts that allows for the agreement to be strong enough to use it as a benchmark as proposed by Kulgemeyer and Tomczyszyn (2015, p. 122). This is supported by the findings in the analysis of the evaluative judgements, as a majority of criteria that is verbalized by the experts is used by a minority of experts (see Chapter 15.3.3.1). Similarly, Hennes et al. (2022) find that, while scientists are the most consistent group in their rating,

experts do not agree with each other in a significantly better way than novices or teachers regarding the relevance of the criteria, nor do they give text ratings with a significantly higher degree of consistency or that are significantly more critical. (2022, p. 290)

Apparently, there is still a substantial extent of disagreement with regard to the rating of texts. Arguably, this could extend to the rating of feedback on texts, explaining the rather moderate agreement among experts overall, but also the contrasting two groups of agreement between experts.

As the experts do not display strong agreement amongst themselves, comparing their results with the RobS score is complicated. Adding the RobS score-based ranking into the matrix estimating Kendall's W as an indication for agreement, yields a moderate and significant agreement, even a slight increase. As

agreement is not substantially altered, it can be argued that the RobS score fits into the existing ranking patterns. With regard to the individual correlations between experts and the RobS score, a significant correlation is only achieved once, in the case of a perfect ranking agreement with expert 2. While only singularly significant, experts 1, 2 and 5, as well as the RobS score, tend to correlate more positively on the one hand, while experts 3 and 4 do so on the other hand.

15.5.6 Limitations

Naturally, such investigations do not come without their limitations. It must be said upfront that of course, the sample size of this study is a limiting factor, especially when conducting inferential analyses of means or variances that seek to investigate differences within the full sample of $N = 51$. One example is the ANCOVA, which is performed in order to control for the graduation grade in the comparison of universities with only $n = 6$ participants in one group. This obvious limitation must be addressed, but should also be contextualized: It is the nature of the broader validation context in which these hypotheses have been stated and tested. They are, hence, just a piece of the puzzle which is the validity argument. They provide further pieces of evidence that are used to support the claims and are not singularly responsible for the validity of the test. In order to account for multiple different perspectives in the VA, some limitations have to be accepted.

With regard to the results and the discussion of H3, several limitations have to be noted. Generally speaking, and as extensively discussed in Part I of this thesis, the relationship between knowledge and action is a point of discussion in teacher education research. The hypotheses that are tested as H3a-H3c are postulated under the assumption that knowledge translates into performance, eventually. As these findings only allow a glimpse into this complex relationship, the results must be interpreted with caution. Results for H3a do not allow for causal assumptions to be made, i.e., that knowledge of feedback leads to an increased performance level in the RobS. Neither should results of H3b and H3c be used to argue that more knowledge in these facets does not contribute to improved scores. For these kinds of assumptions to be made, other procedures, like experimental designs with appropriate control groups, are necessary. These results only provide information on the extent to which there is correlation, which suffices for discussing the extent to which the RobS is situated in the same target domain. Moreover, due to the sample size, the statistical power is limited and not always above the level of .38, which has to be critically accounted for.

Further limitations concern the generation of the respective knowledge test scores. With regard to FALKO-E PCK, the scoring process faced challenges, as outlined in Chapter 15.5.3.2.2. The test scores that are generated for the test are

partially based on item scores that do not meet required benchmarks for inter-rater reliability and the scores themselves are based on additional coding rules, so that they are not necessarily comparable to the results of the original FALKO-E study. Thus, the interpretation of the FALKO-E PCK results requires some caution. Naturally, this extends to the use of the PCK score for correlation calculations, which also need to be reflected against these challenges. Similarly, the internal consistency of the SPK test is lower for the present sample than in the original study conducted by Schütze et al. (2017) and lower than the threshold of .80, requiring the results to be treated with caution as well. The LexTALE test, despite often being used as an indicator for language proficiency, is at heart, a test on lexical knowledge. In the context of this study, the test is used to collect evidence regarding extent to which the RobS score correlates with potentially relevant facets of the target domain. Either interpretation of the LexTALE score, as CK or LP, is potentially relevant for the target domain and thus no view is preferred here. Nonetheless, language proficiency certainly covers more facets than lexical knowledge. For instance, aspects of listening and speaking could potentially have a different relationship as they are more closely connected to the required linguistic performance of the RobS.

Limitations of the expert interviews themselves are addressed in Chapter 15.3.4. However, particular limitations that are specific to the rating are discussed here. First, it is relevant to reiterate the different modes in which the performances are presented and difficulties that experts noted with regard to the comprehensibility of the performances. Expert 1, 2 and 3 based their evaluation on a video- recording that included the audio, while expert 4 and 5 received only the video and subtitles. Potential discrepancies in the rating might also be a result of this mode. However, with regard to MMM07, for instance, expert 5 rates the performance identically to expert 3, while expert 4 differs from all experts. What is more, expert 2 and 3 noted comprehensibility issues with regard to the audio quality and expert 4 also stated that the subtitles were, at times, hard to follow and understand, as the translation was occasionally too complex. Hence, the rating data is not necessarily comparable and needs to be considered with caution, as experts were potentially not able to access all the information for their evaluation. Also, the ranking of the performances might be influenced by the order in which they were shown. CML16 achieves the highest mean rank, was, however, also always the last performance and was shown after the lowest ranked performance by NDA16 was watched. Arguably, the rating of CML16 has benefitted from the direct contrast, as well as from the experience that the experts gained after watching three performances. However, with the small sample size of experts, a variable order might have had different effects in ranking for each expert, which might have skewed the results unexpectedly. Using a

consistent order, the effects were at least identical for all experts. Lastly, expert ratings of the specific performances should be reflected upon in the light of the depicted context. In the RobS, a particular situation is approximated that is situated clearly in the context of EFL instruction in Germany. This is most clearly evident as German is spoken in the majority of the RobS, but also in the specific features evident in the text and the SL, who also inevitably represents a learner in the German context, as actors draw upon the role description as well as their personal history in the portrayal. For the evaluation of feedback that is situated in this scenario, knowledge about the particular German circumstances might be necessary, putting the international experts at more of a disadvantage than simply the issue of language.

15.5.7 Summary

The aim of this study is to provide information on the relationship that the RobS score has with external variables. The extent to which school-related experience, academic progress and success, target domain-related knowledge and expert ratings relate to the RobS score is investigated. For the penultimate aspect, the choice of tests is reasoned extensively and the scoring process, including the scores of the inter-rater reliability procedures, described in detail.

Results indicate that the group of Master students, of which all but one participant has finished the SIS, does have significantly better scores than the group of Bachelor students, of whom, naturally, none has finished the SIS. The number of semesters and the graduation grade also correlate significantly and as assumed with the RobS score. Differentiation in terms of other aspects, like internships that are part of the Bachelor program, or extracurricular school-related experience, does not seem to display significant differences of the RobS score. Of the knowledge facets, only the SPK score correlates significantly with the RobS score when controlled for phase of study or graduation grade, while the PCK score does not show a robust correlation, especially when controlled for these covariates. The CK / LP score does not correlate significantly with the RobS score. Expert ratings' only shows moderate agreement of experts amongst each other, with a substantial extent of variance between ratings. The RobS score correlates significantly only with one expert, but not with other experts or the experts' mean.

Reasons for the lack of correlation with regard to experience might be due to the fact that the internships in the Bachelor phase are of a shorter nature and do not focus on gaining teaching experience. School-related experience outside of the study program is probably lacking appropriate reflection, thus the participants might not be able to use the experience to improve their competence. With regard to the knowledge facets, the degree to which the focus of the tested

knowledge is related to the RobS seems to be indicative of the correlation. The SPK score, which directly measures participants' declarative knowledge of feedback, is the only facet that indicates a robust and significant correlation, while the slightly further removed score for PCK does not show the same robust correlation, especially when controlling for the phase of study, and the CK / LP score is least related to the RobS' content and thus shows no correlation. Comparison with expert judgement is complicated as there is no strong agreement among experts. The experts seem to be separated into two camps that show relatively strong correlations respectively, with the RobS score correlating more with the judgements of three experts, and less with the remaining two experts.

As a limitation, the sample size must be mentioned, but is also seen in the context of the purpose of the study as but one aspect of the validity argument. Also, the challenges in the scoring process of FALKO-E PCK might have impacted the PCK score. Additionally, the internal consistency of the SPK test is lower than required. Thus, results of the correlation analyses must be considered with caution. Furthermore, the score of the LexTALE test should be seen as one aspect of CK, which might be interpreted as an indication for general language proficiency. The expert judgment is limited by the difficulties that arose with regard to comprehension of the audio, the different formats and order in which the performances were presented as well as their knowledge about the specific German context.

16 Summary

In this chapter, the empirical findings of the research process are summarized. The research process is described in its two cycles. The first cycle includes the initial test development, focusing on the creation of the material, the actor training and validity investigations within the (pre-) pilot phase. In the second cycle, the RobS is conducted with the final set of material and with a larger sample to collect sufficient data for further investigations into the participants' perceptions of the test, as well as statistical analyses regarding the development of the scoring instrument and the exploration of the relationship to other external variables. Also, a detailed description of the main survey is provided.

In the pre-pilot phase, the material for the test is developed. Hybrid learner texts are created in multiple steps, drawing on empirical findings and authentic learner texts to ensure a satisfactory extent of authenticity while allowing participants to display the required competence facets. Also, the actors are trained to perform in a standardized way, drawing on the role description and specific VTs. In the pilot survey, the material is tested with the target group of PSETs and the research procedure is refined for the main survey. The research procedure and its components are described in detail in the phases of simulation procedure,

follow-up, and debriefing. A special focus in the pilot survey are the SRIs which are conducted to evaluate the extent to which the VTs elicit the desired competence facets. Using the video-recorded performance, $N = 10$ PSETs are asked to retrospectively verbalize their thought process after hearing a specific VT. The interview data is analyzed using the structuring qualitative content analysis (Kuckartz & Rädiker, 2022, p. 129, 2023, p. 99). Findings indicate that the intended facets are addressed satisfactorily. Thus, the VTs are used for the main survey. The remaining material is also discussed with stakeholders and participants. A final hybrid learner text is selected for the main survey. Concerning the research procedure, the final modifications that are supposed to reduce the overall duration of participation while still allowing for the collection of the required data are presented.

For the main survey, the slightly modified data collection process and the sample are described first, before the four individual studies are discussed. In study 1, data concerning the participants' perspective on the RobS is presented. In guideline interviews, participants are asked for their perspective on the authenticity and fairness of the RobS, the competences they believe are being tested in the RobS, and their opinion regarding a potential implementation of the RobS for summative assessment purposes. The data is analyzed according to the structuring approach to qualitative content analysis (Kuckartz & Rädiker, 2022, p. 129; 2023, p. 99). Overall, a majority of the participants perceive the RobS to be authentic concerning the actors as well as the situation. Possible limitations that are addressed refer to the age of the actor or specific aspects of their communicative behavior. When comparing the competences participants mention are being tested to the framework and scoring instrument, a large overlap can be identified. Some aspects that are mentioned do not align, but have mostly been discussed and consciously excluded from the framework. Concerning fairness, i.e., the ability to show one's capabilities in the context of summative assessment, a majority of the participants consider this positively. Limitations of fairness usually refer to the general skepticism of summative assessment. A possible implementation is viewed mostly positively, for teacher education in general, as well as specifically for the EFL-context. The actual standardization is, however, also addressed with skepticism. Limitations to these findings consider the method of interviews, in which social desirability is of concern. Moreover, this study only includes the perspective of PSETs, while especially for the latter aspect of implementation, further perspectives are of interest.

In study 2, $N = 5$ experts are asked to evaluate selected performances from the main survey, as well as the initial scoring instrument with the aim to 1) collect evidence that supports the appropriateness of the scoring instrument and 2) improve the scoring instrument according to the statements by making

modifications. In the first part of the interview, the experts watch four performances which they rate on a global rating scale. Additionally, the experts provide evaluative comments with reasons for their judgement. In the second part, the initial scoring instrument is presented and evaluated by the experts. The data is analyzed using the qualitative content analysis approach to Mayring (2022). Based on the qualitative data of the first part, a list of criteria the experts use for assessment is derived inductively. The results indicate that a majority of the experts agrees on eight criteria, of which all but one is related to the framework. Overall, however, there does not seem to be a large degree of agreement among the experts, as a number of the verbalized criteria are not mentioned by a majority of them. For the second part, the experts' statements are analyzed for each of the deductive categories of the framework. Additional inductive categories are added, if required. Findings indicate that overall, the scoring instrument and its items are appropriate. Some experts are, however, skeptical with regard to the comprehensiveness of the instrument at large as well as with specific item operationalizations. Results are used for modifications of the scoring instrument, which are described transparently. Limitations of the study relate to the singular perspective of experts on the scoring instrument and the technological difficulties that complicate the interpretation of the data collected in the first part, especially, as outlined.

In study 3, the reliability of the scoring instrument is investigated based on the scoring of the full sample of $N = 51$ participants, as well as a subset of $n = 10$ performances, which is focused on in the inter-rater reliability investigations. The aim is to explore the appropriateness of the scoring instrument from a quantitative perspective and to modify the set of items and the rating manual. The initial scoring instrument has 27 items, of which three are excluded as they do not provide additional value to the total score or correlate negatively with the corrected item-total scale. The internal consistency of the 24-item scale is $\alpha = .81$. Regarding the inter-rating, the chosen sample, the rater training, and the multiple rating and consultations steps are outlined. Based on the results of the consultations, modifications are made to the rating manual and described. Considering the inter-rating results of rating II, the items indicate an agreement range for all three raters of Fleiss' $\kappa = .54$ to $\kappa = 1.00$, and an overall agreement of $\kappa = .83$, $p < .001$; 95 % CI [.78, .88]. Considering the test score, the descriptive results show that within the sample, a range of ten to 47 points (Max = 48) is achieved, which is also reflected in a right-shifted distribution. The inter-rating results of rating II on the test score level indicate an "excellent" (Cicchetti, 1994, p. 286) level of agreement, $ICC_{just}(2,1) = .96$, 95 % CI [.90, .99], $p < .001$; $ICC_{unjust}(3,1) = .96$, 95 % CI [.90, .99], $p < .001$. Considering the actors as a potential source of measurement error, results show that the subsample that engaged

with actor 1 receives statistically significantly less points than the groups with actor 2 or 3. Potentially, the composition of the individual sample of actor 1, which mostly consists of students in the Bachelor phase, factors in here. Possible limitations in this study refer to the inter-rating study, e.g., the smaller subsample and the conscious decision to use performances that are likely to show the broad range of possible test score outcomes.

In study 4, the test scores of the RobS are used to test hypotheses regarding the relation the RobS has to other variables. For that purpose, data and results of the written survey (H1-H3) and the quantitative expert ratings (H4) are processed and drawn upon. H1 states that the RobS score correlates significantly positively with the amount of school-related experience. The results for the four sub-hypotheses show that for H1a, the participants that have completed of the SIS perform significantly better than those who have not. H1b, which focuses on the relationship between internships in the Bachelor phase with the RobS score, as well as H1c and H1d, which focus on the relationship and correlation with extracurricular experience in school, are not accepted. H1 is thus only partially accepted. Arguably, only the SIS offers a substantial and professionally supported experience that provides learning opportunities by reflection on action. Hence, the underlying assumptions may have to be reconsidered and adapted. H2 states that the RobS score correlates significantly positively with academic progress and academic success. The results for H2a show that students in the Master phase perform significantly better than those in the Bachelor program. Similarly, findings for H2b show that the number of semesters also correlates significantly positively with the RobS score. The results for H2c indicate that there is also a significant correlation between the final graduation grade and the RobS score. The results for H2d suggest that the university of the participants does not lead to different RobS scores. All hypotheses support the assumption and H2 is thus accepted. H3 states that the RobS score correlates significantly positively with other facets of the target domain. Results of H3a show that there is a robust significant correlation with moderate effect size between the RobS score and SPK, i.e., declarative feedback knowledge, even when controlled for final graduation grade or phase of study. Findings for H3b show that for EFL-specific PCK, there is a moderate significant positive correlation with the RobS score. However, it is less robust, especially when controlled for graduation grade, and insignificant, when controlled for phase of study. For H3c, it can be stated that there is no correlation between CK / LP, i.e., lexical knowledge, with the RobS score. Results for H3d indicate that the three facets SPK, PCK and CK / LP correlate significantly to a moderate to large extent, suggesting that these three tests are part of one target domain. The hypotheses for H3 are thus partially accepted. Potentially, the relevance of these knowledge facets for the situation

that is specifically depicted in the RobS explains the different correlation values of the three facets with the RobS score, as the more pedagogically relevant facets seem to be more correlated with the RobS score. H4 states that the RobS score correlates significantly positively with the expert ratings. Results for H4a suggest that there is significant moderate agreement among the experts regarding the ranking of the performances. Concerning H4b, findings indicate that the RobS score fits into the expert rating, leading to a small increase in agreement. Findings for H4c indicate that among the experts, there are likely two groups which show more agreement among each other, respectively. The RobS score shows a higher extent of correlation with a group of three experts and, within this group, a total agreement with one expert. The heterogeneous findings from the expert rating align with the qualitative investigations, which show that the experts use a variety of different categories for the evaluation of the performances. H4 is thus only partially accepted. Limitations of this study refer mostly to the scoring challenges of the PCK test, as well as the low internal consistency of the SPK test. For the expert rating, the technological difficulties, especially for the subtitled videos, might limit the quality of the rating. Also, knowledge of the German education system might be necessary for an appropriate contextualization and evaluation of the performance.

In the following and final part of this thesis, the data will be discussed with regard to the IUA, evaluating the warrants and claims for the VA.

Part V – Discussion: Evaluating the Validity of the RobS

*[The] justification and defense of measurement and its validity is
and may always be a rhetorical art.*

Samuel Messick (1988, p. 43)

In this final part of the thesis, the overarching research question is addressed and the findings are reflected on. Firstly, the claims made in the IUA are evaluated based on the collected evidence and a conclusion is drawn regarding the research question. Secondly, a comprehensive summary of the thesis is provided, limitations relating to the overall research process are drawn and implications made.

17 Evaluating the Claims: The Validity Argument

In this chapter, the validity argument will be made by discussing each claim made in the IUA, based on the evidence that has been presented in the previous parts of this thesis. At the end, a conclusion will be drawn regarding the overarching research question of this thesis (RQT):

RQT: To what extent can the RobS be considered valid?

In the sense of Messick, as stated above, the aim is consequently to rhetorically justify and defend the RobS as presented in this thesis.

17.1 Claim 1: Model

Claim 1: The theoretical model for feedback on learner texts depicts the required competence facets adequately.

Warrant: ... because it is based on relevant theoretical literature and sound empirical evidence. The model includes the relevant facets that are required for providing feedback on a learner text in a one-on-one situation and takes into account the specifications of the intention to evaluate performance-oriented aspects of competence. It refers to given normative settings, current empirical findings and sound theoretical guidelines in order to include multiple perspectives that are necessary to holistically describe the model (Schreiber & Gut, 2022, p. 131).

In order to discuss this claim, the theoretical design of the framework will be addressed first, followed by an evaluation of the experts' judgements on it.

The framework is described transparently and a relatively broad theoretical foundation is drawn upon, as evident in Chapter 7. The focus of the test is

reasoned extensively, drawing on empirical research as well as the relevant statutory guidelines, the latter supporting the aspect of curricular validity. Other studies are used for an orientation of relevant dimensions regarding feedback, so that the framework is compatible with existing research. This is especially clear concerning the inclusion of the feedback model that is proposed by Hattie and Timperley (2007), which is frequently referenced and is also based on empirical findings. The other aspects of the framework seek to add to that model in order to reflect the subject- and context-specific demands, i.e., focusing on the interactive and pedagogical strategies, as well as the actual content of the conversation. In that, supporting the development of the learner's writing competence is emphasized and empirical findings are drawn upon which show approaches, methods, and techniques that are likely to have a positive impact on learner achievement. A normative connotation is added to the framework as the requirements for EFL teachers, specifically the aspect of functional monolingualism, and the participants' accuracy regarding the information are included. Although quite broad in its foundation, there are also limits to the framework that are laid out, as aspects that are not directly relevant for the RobS are excluded from the framework consciously, while other, possibly relevant aspects might not be included unintentionally. The framework, thus, fulfills its purpose for the RobS, but could, and possibly should, be modified for other purposes.

The experts' evaluative comments of the performances and their evaluation of the scoring instrument offer an additional perspective on the model. Looking at the criteria that the experts mention when evaluating the performances as described in Chapter 15.3.3.1, it is striking that experts focus on different aspects of the performance. The experts make use of a diverse set of criteria to assess the performances, indicating that there is not necessarily agreement in the criteria that they verbalize in the evaluative comments. However, when comparing the list of mentioned criteria to the framework, there is still a substantial overlap, which is generally supportive of the warrant. Of the eight criteria that a majority of experts mention, seven are also found in the framework. Most of the remaining criteria can also be related to the framework to some extent. While there are also aspects the experts address that are not included in the framework, i.e., the structure of the conversation and history of the learner, there is a preceding discussion with regard to the framework as to why these notions are not considered for the context of this RobS (see Chapter 7.4). Results of the evaluation of the scoring instrument also indicate that overall, the experts consider the included items as appropriately derived from literature. With regard to the model, experts' evaluations lead to the inclusion of para-linguistic aspects for *Dialogue orientation*, i.e., *Body language*.

While expert evaluation is frequently cited as a potential source for an investigation of aspects concerning content validity in validation literature (see Chapter 10.2.1), their statements must be critically reflected on. When conducting expert interviews, the expectation is that the experts draw upon their implicit and explicit knowledge, which they have acquired during the course of their careers. However, experts are potentially prone to challenges when it comes to verbalizing their knowledge and often, their judgements are also influenced by intuition (Bartels, 2018, p. 160). These limiting factors must be considered, as they are reflected in the collected data. Consequently, while experts' judgements must be viewed critically, the evidence suggests that the majority of the aspects mentioned by the experts is also reflected in the framework, which rather supports the warrant.

In conclusion, the framework is based on adequate literature, i.e., logically derived theoretical deliberations and sound empirical evidence, that takes a relatively expansive and diverse perspective on feedback. The depiction of the framework is also reflective of its limitations, i.e., aspects and dimensions that have not been included. Moreover, the findings of the expert interviews further corroborate the warrant, as the criteria the experts verbalize most align with the framework. Based on these findings, the warrant is supported and the claim can be substantiated.

17.2 Claim 2: Test

Claim 2: The test material gives participants the opportunity to show their competence of providing feedback on a learner text in a one-on-one situation. The general format of the test focuses on the notion of "shows how" (Miller, 1990, p. 63) and the intention of measuring the performance-oriented aspect of competence (Blömeke et al., 2015).

Warrant: ... because the material, i.e., the set of documents including the learner text and the verbal triggers aligns with the model and is appropriate for the sample (Schreiber & Gut, 2022, p. 131).

In order to discuss this claim, the findings regarding the VTs, the appropriateness and perceived authenticity of the material and the actor, as well as the participants' perspective on the tested competences are drawn upon.

Concerning the VTs, SRIs are conducted to explore the extent to which the initiated cognitive processes align with the intended goal (see Chapter 14.2.2). For that purpose, participants of the pilot survey are asked to re-watch their performance as a stimulus and verbalize their thoughts after the respective VT was uttered. Overall, the results indicate that the VTs do elicit the desired reactions, corroborating the warrant. Due to the indexical nature of language, VTs

may, however, also evoke inadvertent reactions, which shows that a certain extent of uncertainty remains. Such an introspective approach to data collection has its own limitations, especially with regard to the accuracy of the retrospectively verbalized cognitive processes, but it still allows for an indication of the reaction that the VTs have on the participants. Consequently, the VTs can be considered to fulfill their purpose sufficiently and the findings support the warrant.

Regarding authenticity, a number of different explorations are made in the course of material development (see Chapter 14.1.1) and evaluation of the RobS (see Chapter 15.2). The authenticity of the material and its development is explored via multiple perspectives, focusing mostly on the learner text as the basis of the RobS. Two of the hybrid learner texts which are considered to be most authentic by the PSETs in the pre-pilot phase are used for the pilot survey. To provide evidence that the relevant text dimensions addressed in the framework are still found in the hybrid text, balancing the demands of authenticity and content validity, the learner texts are inter-coded. In the pilot survey, the utilized learner texts are considered sufficiently authentic by the participants. When specifically asked, some aspects of the texts are noted as not particularly authentic, but this does not seem to distract from the actual performance that is required in the RobS. Similarly, the other documents that are part of the material are considered appropriate by other stakeholders and sufficiently authentic by the participants of the pilot survey. Again, individual participants mention aspects that they perceive as inauthentic, but only rarely to the extent that it may impact the overall authenticity drastically, e.g., the amount of information on the task sheet (M2, Appendix A.3). Overall, the perception of authenticity and transparent description of material development support the warrant.

In the main survey, participants are questioned about the authenticity of the actor and the situation (see Chapter 15.2). The findings indicate that the vast majority of participants consider the actors to authentically portray their role, acting naturally and with consistency. Participants often highlight communicative behavior as one of the factors that contribute positively towards an authentic perception. One aspect that may limit authenticity is the age of the actor. However, the findings do not indicate that this was an issue for the participants' overall perception of the RobS. When asked specifically about the authenticity of the situation, the results are similar. Also, a majority of the participants consider the situation to be one that occurs in school to some extent. Considering that for content validity the test format should also have relevance for the target domain, this finding indicates that the participants perceive this to be the case with the RobS (Messick, 1995, p. 745). Thus, the objective of the RobS development, i.e., creating a simulation that is an authentic *approximation of practice*

(Grossman, Compton, et al., 2009, p. 2076), can be considered to be fulfilled, at least from the perspective of the PSETs. There is no backing in this study which allows for an indication of the perception of the RobS by other stakeholders like teachers, learners, or teacher educators. The existing findings, however strongly support the warrant, as the RobS is generally perceived as an authentic approximation of practice.

In this context, a relevant aspect to discuss is the extent to which participants perceive competences to be tested that actually are intended to be tested. Participants name a plethora of different competences they believe are being tested in the RobS in the follow-up interviews. Most of the mentioned aspects can be aligned with the general focus of the RobS, e.g., feedback and assessment, the subject-specific focus, i.e., the use of the English language, or the communicative format, e.g., teacher-student-interaction, as described in detail in Chapter 15.2.3. A frequently mentioned subcategory of teacher-student-interaction which is not tested, is that of emotional-motivational competences. In this case, the scoring instrument and the framework do not match with the participants' perception. Arguably, transparency about the focus of the assessment is important so that participants do not have false expectations or adapt the scoring instrument by adding this notion to the framework, otherwise, the perceived content validity might be limited. Consequently, the findings support the warrant to a certain extent, as a substantial portion of competences named by the participants is covered in the RobS.

In conclusion, it can be stated that the VTs elicit the cognitive processes that are intended and the competences that participants perceived to be tested do largely align with the framework, with the exception of emotional-motivational aspects. Several stakeholders perceive the test material, i.e., the learner text and the remaining documents to be sufficiently appropriate and authentic. A large majority of the participants state that they perceived the actors, as well as the situation to be authentic, albeit acknowledging potential shortcomings. A particular challenge for this claim relates to the appropriate balance between aligning the model with the test, while still creating a situation that approximates practice authentically. Considering these findings, not only can it be stated that a sufficient balance is found, but also that the evidence provides adequate backing, which substantiates the warrant as a whole.

17.3 Claim 3: Performance

Claim 3: The participants are enabled to display a performance in the test that represents the target domain.

Warrant: ... because the performances are generated in a standardized manner, participants are able to access and comprehend the material, show their actual

capabilities and the performance resembles aspects of the target domain (Schreiber & Gut, 2022, p. 132).

For the discussion of this claim, the description of the data collection process, the influence of the actor on the performance, and the participants' perception of fairness are drawn upon.

The data collection is transparently described for the pilot survey in Chapter 14.2.1 and for the main survey in Chapter 15.1. The standardized procedure is outlined and is identically conducted for all participants. The RobS itself is standardized due to the provision of an identical set of material in the preparation to all participants in the main survey. Participants did not broadly indicate issues in comprehension of the material when given the opportunity to ask clarifying questions. Moreover, the findings from the interviews in the pilot survey do not indicate that there were any comprehension issues, especially when asked about the language of the material. In the main survey, the same set of VTs was available to and used by the SL, depending on the participant's behavior. While the individual course of the RobS is dependent on the participant, the standardized preconditions of data collection give all participants the same opportunity to perform, supporting the warrant in this aspect.

One aspect of the RobS that is not identical to all is the actor. Through a joint training and identical set of VTs, the foundation is laid for a standardized portrayal of an SL, in an attempt to minimize the impact of the specific actor. Looking at the findings, however, the warrant cannot be unequivocally supported in that regard. Evidently, the test scores that are generated in RobS with actor 1 are significantly worse than those of actors 2 and 3, which do not differ from each other substantially (see Chapter 15.4.4). Explanations for this might be that the subsample of actor 1 is substantially smaller and consists mainly of Bachelor students, who generally also perform significantly worse than Master students, while actors 2 and 3 have more balanced subsamples concerning the phase of study. Although no further anomalies are observed in these performances, it cannot be fully ruled out that the choice of actors might also have an influence on performance to some extent and, consequently, this finding does not necessarily support the warrant.

In the follow-up interviews, participants are also asked whether they believe that the RobS allows participants to demonstrate their actual capabilities. While a majority of participants views this rather positively, limitations are also mentioned, suggesting that the general context of summative assessment might hinder participants in showing their capability, regardless of the assessment format. Additionally, the individual preference for assessment formats is also highlighted as a reason that the RobS is not necessarily fair. Arguably, these limitations are

not necessarily inherent in the RobS, but point to a larger discussion of summative assessment, which is seemingly perceived with skepticism by the participants. Thus, these results support the warrant rather tentatively.

Concerning the resemblance of the RobS performances to demands of the target domain, no specific additional data was collected. As discussed in the previous chapter, a majority of the participants believe that the situation depicted in the RobS is likely to occur at school, a finding that also supports this warrant. However, it remains unclear whether other stakeholders, like teachers and teacher educators, agree. Complicating this discussion are findings that indicate that such one-to-one feedback situations focusing on writing, rarely happen at school (Porsch, 2010, p. 66), which makes a direct comparison between the RobS and the real target domain difficult. Nonetheless, feedback situations generally occur at schools frequently and feedback, especially in the writing context, is essential for the learning process, as described at length in Chapter 6. Consequently, the focus of the RobS remains legitimate, despite the fact that no further evidence supporting this specific aspect is presented that may support this warrant.

In conclusion, the warrant for this claim can only be partly supported by the evidence. Identical documents and VTs, as well as an identical research procedure lay the foundation for standardization. However, it cannot be fully ruled out that the actor's behavior in the RobS influences the test score, despite logical explanations as to why the evidence suggests so. While most participants perceive the RobS to be a fair assessment allowing them to display their competence, general challenges regarding fairness are noted in the context of summative assessment. Besides the perspective of the participants, there is no evidence that allows for a direct evaluation of the extent to which the performances generated in the RobS resemble the target domain, albeit the general notion of providing feedback is certainly part of the target domain.

17.4 Claim 4: Item Score

Claim 4: The performance of the participants during the test can be assessed using the scoring instrument.

Warrant: ... because the items in the scoring instrument, which are based on the framework, reflect the relevant facets to measure the construct comprehensively. The individual items are operationalized appropriately to the test format and can be generated in a replicable manner (Schreiber & Gut, 2022, p. 132).

In order to discuss Claim 4, the expert interviews, item frequencies, and the item-specific reliability investigations are drawn upon.

In the second part of the expert interviews, the evaluation of the scoring instrument and its items is the focus of discussion. Overall, the scoring instrument is evaluated positively by the experts, which can be interpreted as general support for the warrant. However, the operationalization of individual items and the general complexity and comprehensiveness of the initial instrument are discussed critically. Thus, substantial changes are made and documented transparently. For instance, the item *Body language* is added to account for the paralinguistic dimension, which is considered to be missing in the preliminary version of the scoring instrument, and the scoring system is simplified by reducing the number of point-tiers. Modifying the instrument to more closely align with expert judgements strengthens the support for the warrant, as it can be assumed that the experts, who are all researchers in adjacent fields, possess the specific knowledge required to create a suitable scoring instrument. However, as explained in the discussion for Claim 1, expert evaluations must be reflected upon. Considering the diversity of opinion in the evaluation of the performances, it cannot be stated that experts necessarily agree on the constituents of a 'good' feedback performance. As a consequence, individual points of criticism are possibly not shared among the experts. The modifications, thus, bear the risk of favoring some perspectives and should ideally be discussed in a second cycle of expert interviews, which has however not been feasible within this project. Overall, it can nonetheless be stated that the findings contribute to the warrant's support.

For the items to be appropriate to assess the performances, they need to fulfill essential demands. Firstly, they should be necessary to describe the different performance levels, i.e., used in the rating process. Secondly, they need to be discriminatory, i.e., essentially measure different aspects. Looking only at the 24 items included in the final scale, all point-tiers for each item are used at least once, i.e., all point options are needed to describe the performances in the sample. Items that might be considered a subset of other items are excluded, based on the analysis of co-occurrences. These statistical descriptions support the warrant, as all item point-tiers are used and none of the items in the 24-item scale seems to be redundant.

A last aspect to consider is the consistency with which the scoring instrument and the items are applied by different raters. The inter-rating investigations are conducted to explore the reliability, as well as to improve the descriptions of the items for assessment purposes. Regarding the latter, the changes are described in detail in Chapter 15.4.1.2.3 to ensure transparency. Overall, the agreement of the three raters is satisfying for both rating cycles, considering Fleiss' κ for all items is $\kappa = .60$ in rating I (before consultations), and $\kappa = .82$ in rating II (after consultations). While in rating I, several items do not meet the thresholds values,

most do after consultations in rating II. Some items that do not meet the threshold of the inter-rater reliability value are still included as not to reduce construct validity. This is in line with Döring's opinion that exclusion of items should not only be made based on statistical values, but be reflected against the construct (2022, p. 562). This is essential so as not to adapt the items too closely to a specific sample, as other samples might generate different reliability values. However, it must be noted that on the item-level, reliability is substantially improved after raters engaged in consultations regarding the discrepancies, as the identification of events in the performances seems to be a potential challenge. Consequently, multiple raters seem to be required to achieve sufficient reliability for individual items. All in all, the findings from the inter-rating procedure rather support the warrant.

In conclusion, the warrant is sufficiently supported. Modifications to the scoring instrument are made to increase the quality with regard to expert judgments and inter-rater reliability. This relates to additions and exclusions of items and simplification of the scoring instrument, compared to its initial version (see Appendix D.5). Overall, experts evaluate the items in the scoring instrument positively regarding the relevance for measuring the performance. The statistical reliability values for rating II and the final set of items all indicate that the items are considered to be appropriate to measure the performance, especially when multiple raters are used.

17.5 Claim 5: Test Score

Claim 5: The sum of the item scores generates a test score that represents the participants' levels of performance.

Warrant: ... because the test scores are generated reliably by several raters and the range of generated test scores depicts a distribution across the spectrum, indicating that it differentiates sufficiently. The individual items that contribute to the test score are all relevant for the construct and are internally consistent (Schreiber & Gut, 2022, p. 132).

For the discussion of Claim 5, the descriptive data of the test scores, the internal consistency of the scale, and the inter-rater reliability on the test-score level is drawn upon.

The test score indicates variance across the possible spectrum of points. The test scores in the sample range from 10 to 47 points. Thus, while the range of 37 points indicates that a large span of the point range is covered and the upper range of possible scores has almost reached its maximum (48 points), the lower tier is not covered entirely. This fact is reflected in a right shift in the data distribution and the average score of $M = 31.22$ ($SD = 8.00$) points and median of 33

points. Several reasons might explain this distribution, as described in Chapter 15.4.2.3. Still, 80 % of the potential test score span is covered and there are no indications of ceiling or flooring effects in the data distribution. Concluding, the descriptive statistics of the test scores rather support the warrant.

The internal consistency of the final 24 items that are used to aggregate the test score is sufficient, with Cronbach's $\alpha = .81$. While some items have a lower correlation with the corrected item-total correlation, i.e., $p < .30$, no remaining item correlates negatively, which is a definitive criterion for exclusion. The low-correlating items are not excluded, as they are essential for the construct and their exclusion would severely limit the construct validity. What is more, Döring argues that for tests that have a broad construct to measure, such as the RobS, higher correlations might be problematic as this would indicate redundancy (2023, p. 472). Consequently, the internal consistency, as well as the corrected item-total correlations, can be interpreted as sufficient for the context and corroborate the warrant.

Lastly, the inter-rater reliability on the test-score level is focused on. Given that the test score is the foundation for further decisions, e.g., grades or certifications, reliability is of particular relevance for summative assessment. The results indicate that when aggregating the item scores, there is an overall high consistency between the raters. The ICC values for all three raters for both rating I and rating II are higher than .75 and can be considered "excellent" (Cicchetti, 1994, p. 286). Moreover, the pairwise correlation between raters also reveals a high extent of agreement. As a potential limitation, it must be added that the choice of performances for the inter-rating sample might have influenced these values, as a variety of different performance levels has been consciously chosen based on the scores of rater 1. This conscious range might artificially increase the ICC values. However, it is apparent that the differences in the test scores can be reproduced consistently, which can also be interpreted as a strengthening of the reliability inference. Hence, these findings support the warrant strongly.

In conclusion, the inferences made here, based on the evidence, support the warrant. While the whole range of potential test scores is not generated in the sample, a substantial span is covered. The reliability indices also support the merit of the test score, as the internal structure of the scale has a sufficient value of Cronbach's α and the inter-rater reliability scores on the test score level are also satisfactory, which is particularly relevant for the context of summative assessment.

17.6 Claim 6: Latent Trait

Claim 6: The individual test score can be used to describe the participants' levels of competence of providing feedback on learner texts in one-on-one settings.

Warrant: ... because the situation in which the performance takes place closely resembles the target domain (Schreiber & Gut, 2022, p. 132).

For the discussion of Claim 6, the results of the correlations with the knowledge tests are drawn upon. Correlations can be used as indicators to discuss the extent to which the RobS score is related to aspects of the target domain, i.e. EFL teacher competence, and appropriately measures the intended competence facet (see Chapter 15.5.4.3).

The results of the tests on SPK, PCK and CK / LP are used to investigate possible correlations with the RobS score. Among them, the three tests show a significant correlation of a medium to large effect size, indicating that they are likely to be representative of one domain. Considering SPK, there is a robust correlation of medium effect size between the SPK score and the RobS score ($r = .45$, $p < .001$). The effect is smaller but persistent when considering the potential covariates *phase of study* and *graduation grade*. Evidently, more declarative knowledge on feedback is associated with a better feedback performance, at least in this sample and this specific context. Moreover, the SPK test is most closely related to the RobS from a content and construct perspective, and, thus, the findings support the ability of the RobS to account for feedback performance and, thereby, the warrant. For PCK, the correlation is of a smaller, but still medium effect size ($p = .37$, $p = .008$). However, it is less robust, especially when taking the graduation grade as a covariate into account ($p = .28$, $p = .050$). There is no significant correlation between RobS and PCK, when controlling for phase of study ($p = .28$, $p = .052$). As a resource, PCK is seemingly less correlated with the RobS score than SPK. One possible explanation for this might be found in the construction of the PCK test, which does not necessarily include aspects that are relevant for the specific construct used in the RobS. Another possible reason might be the fact that PCK is less immediately relevant as a resource in the situation and that SPK, in this case, is more acutely necessary in the RobS. Still, there is sufficient indication that the RobS can also be situated within the field of EFL-specific PCK, tentatively supporting the warrant. There is no significant correlation for CK / LP with the RobS score, while this facet is arguably part of the target domain, as indicated in the moderate correlations to PCK and SPK. Apparently, a differentiation between facets of the target domain must be made concerning the extent of their pedagogical relevance. In that regard, it is not surprising that the knowledge facet that is least pedagogically relevant is not correlated with

the RobS score, which focuses rather on a pedagogical situation. All in all, the findings partially support the warrant.

In general, the underlying assumption of this claim needs to be reflected upon as there is an ongoing discussion regarding the relation of knowledge and performance of (pre-service) teachers. While the presented claim presumes that both aspects are part of one target domain, i.e., EFL teacher competence in the sense of Blömeke et al. (2015), there are heterogeneous findings on the relation of PK, PCK and CK on the one hand, and actual instructional quality and learner achievement on the other hand (see Krauss, Lindl, Schilcher, & Tepner, 2017, pp. 37–38). Therefore, it cannot necessarily be assumed that there should be a correlation between the knowledge facets and the performance aspects. However, the evidence generated in this study is still valuable. For one, it furthers the exploration of this relationship as the evidence suggests that the relationship between different knowledge facets and performance is quite variable. Moreover, these findings allow for an estimation to be made of which aspects of knowledge might correlate more closely with the RobS score, which in itself provides insightful information for the claim. Additionally, the limitations that arose in the generation of these external test scores must be taken into account when reflecting on the substance of the claim, as detailed in Chapter 15.5.6.

In conclusion, the evidence rather supports the warrant as the pedagogically relevant facets of the target domain correlate with the RobS score, though not necessarily comprehensively, considering the smaller and less robust correlation of PCK, especially when controlled for phase of study.

17.7 Claim 7: Interpreted Trait

Claim 7: The test score can be used to make interpretations about the participants' competences.

Warrant: ... because the collected empirical data aligns with assumptions that have been made *a priori* (Schreiber & Gut, 2022, p. 133).

In order to discuss the claim, the remaining findings on *a priori* hypotheses with regard to school experience, academic progress and success, and the expert ratings are drawn upon (see Chapter 15.5). The extent to which these hypotheses are accepted is discussed here specifically to the extent that the results support the claim.

Concerning school-related experience, the assumption is that the general extent of school-related experience correlates with a higher RobS score. However, only the differentiation between groups concerning the completion of SIS reveals significant differences. Neither the internships in the Bachelor program, nor extracurricular school experience seems to lead to significant differences in

the RobS score. Of the described types of experience, only the SIS spans over a consistent amount of time, demands a substantiated amount of instruction at school, and includes structured mentoring and reflection on performance, which might explain these findings. However, it must be noted that the SIS also takes place in the Master program. Hence, no clear differentiation can be made as to whether the general academic progress is a factor, or the SIS itself. Future studies could focus on this fact and examine the specific impact that the SIS has on RobS performance. The other practical phases are either much shorter and do not require substantial instructional experience, i.e., the AOI or OI, or do not necessarily have the appropriate support system for PSETs, e.g., in the job as substitute teachers. Thus, these kinds of school-related experiences likely do not sufficiently contribute to professional instructional development, which possibly explains the rejection of the assumptions. Potentially, these specific assumptions and hypotheses, though logically derived, should be modified to account for the quality and extent of the school-related experience. This would allow for an interpretation of the results that is supportive of the warrant as it indicates that actual learning needs to occur to perform well in the RobS. Experience, apparently, is not a means, in and of itself, leading to better performance. Only those PSETs who have gained a substantial amount of practical experience in which they actively engage in reflection on action perform better. This perspective would generally support the warrant, as this indicates that the RobS is sensitive to these professional developments. Consequently, the inferences made based on the findings could be interpreted to support the warrant, if assumptions are modified to account for the quality of experience.

Concerning the academic progress, participants in the Master program perform significantly better than those in the Bachelor program. Similarly, there is a significant correlation of medium effect size between RobS score and number of semesters ($r = .47, p < .001$). This finding aligns with the previous discussion on the SIS, as no clear differentiation can be made whether these significant differences or the correlation are based on general academic progression or the SIS. With regard to the warrant, however, both explanations are logical and supportive as, either way, the RobS score seems to be sensitive to academic professionalization processes. The final graduation grade correlates negatively with the RobS, as expected due to the inverse German grading system, $r = -.44, p = .001$. This variable is often used as an indicator for cognitive ability and predictor for success of study. Hence, the better performance of participants with better grades in test situations of any kind, including the RobS, is more likely. Lastly, participants' university is not a significant factor for success in the RobS, as the graduation grade is likely the more significant covariate when controlled for. This also supports the warrant, as the statutory guidelines prescribed by the KMK

are identical for both universities and, thus, no differences are to be expected in the RobS performance. Hence, the inferences made based on academic progress and success all support the warrant as they align with the assumptions.

Lastly, inferences are made based on the quantitative results of the expert interviews, i.e., their global rating. Presumably, the experts' rating is supposed to align, to a certain extent, with the test score. Statistically, experts agree to a moderate extent on the ranking of the four performances in the sample, $W = .502$, $\chi^2(3) = 7.531$, $p_{exact} = .047$. Adding the final test score of rater 1 into the rank matrix leads to a slightly better agreement, $W = .517$, $\chi^2(3) = 9.305$, $p_{exact} = .017$. Generally, this finding supports the warrant as competence levels are moderately agreed upon between experts and the test score. However, there still are caveats evident when taking the pairwise correlations between the RobS score and the individual experts, as well as their averaged ranked mean into account. The descriptive results and pairwise comparison show a more diverse ranking among the experts as, for instance, expert 4 deviates from the other experts, showing negative, though non-significant, correlations with experts 1 and 2. Arguably, there are two camps of agreement among the experts, with the RobS score exhibiting more positive and partially significant correlations with one group of experts compared to the other. The moderate agreement is also a reflection of the qualitative data which shows that a majority of experts agree on only about a third of criteria they use in the rating process. As discussed extensively in the limitations of the expert interviews, it therefore does not seem as if there is an overall high agreement concerning specific criteria to assess feedback performance, which then is reflected in the quantitative data. Consequently, the inferences made are of rather moderate support for the warrant.

In conclusion, the findings with regard to school-related experience and academic progress rather support the warrant, as the RobS score is significantly better for those participants that are further along in their studies and have had the opportunity for extended, guided school-related experience. The findings of the expert rating are also supportive as there is a significant agreement between experts and the test score, albeit to a lesser extent. The heterogeneous rating of the experts has to be taken into account here.

17.8 Claim 8: Use

Claim 8: The test can be used for the intended summative purpose.

Warrant: ... because the generated evidence supports the intended use of the test and stakeholders view the test positively (Schreiber & Gut, 2022, p. 133).

In order to discuss this claim, the findings on the participants' perspectives on a potential implementation are drawn upon (see Chapter 15.2.2.4).

Generally, the participants view a potential implementation positively, both for teacher education in general and for EFL teacher education specifically. They often highlight the possibility that the RobS might address a gap in teacher education and that it might contribute to them being better prepared for practice. For the EFL context specifically, participants emphasize the fact that the RobS covers the aspect of appropriately communicating with learners in an adaptive manner. However, individual participants also point to the challenges, e.g., skepticism concerning the standardization. Although generally positive, these findings only indicate the perception of a hypothetical scenario. Likely, the perception differs once the RobS is actually implemented as an assessment format and higher stakes are associated with it. Moreover, the findings represent only a singular perspective on potential use. Other stakeholders should be included in order to substantiate the discussion of this claim generally, as the singular perspective is not sufficient. Nonetheless, the findings that have been generated support the warrant.

The actual use of the test, including the perspectives on it, is not explored in this thesis. Further studies are necessary in order to provide an appropriate indication of the acceptance of the RobS as an assessment format when it has been implemented. This extends to the potential ramifications and consequences of use, especially for the participants who are the subjects of the assessment. However, this is also not claimed in the IUA, as this thesis is concerned with developing a prototype of a RobS that is still to be evaluated *in situ*.

In conclusion, there is no evidence that allows for a direct discussion of the use of the RobS as summative assessment. However, the participants' perspective on a hypothetical implementation lends some support to the idea that the RobS might be accepted by the participants in the future, when implemented.

17.9 Conclusion: To use, or not to use?

For the validity argument of the RobS, several claims are stated and evaluated with the collected evidence, with the overall aim to provide answers for the overarching research question of this thesis:

RQT: To what extent can the RobS be considered valid?

Considering the collected evidence that supports the warrant and claims, the findings that do not, as well as the limitations that are outlined in the validity argument, the following can be stated:

Overall, the validity argument indicates that the RobS can be used with sufficient confidence to describe and assess PSET's feedback competence on writing.

This conclusion is drawn based on the evidence that

- the **model** reflects the targeted construct appropriately;
- the **test material** is an appropriate and authentic operationalization of the model;
- the **performance** is sufficiently standardized and perceived as fair;
- the **items** used for scoring are appropriate for the model and generated reliably;
- the **test score** allows for appropriate differentiation and is generated reliably;
- the **latent trait** described with the test score is situated sufficiently in the target domain;
- the **interpretation** of the test scores aligns sufficiently with *a priori* hypotheses;
- the (potential) **use** is favored by the participants.

Naturally, limitations are considered in this conclusion. Critically, the potential influence of the actors, as well as the merely moderate agreement amongst experts must be considered as limiting factors. Moreover, the assumptions regarding school-related experience need to be modified. In the end, however, the evidence that supports the claims outweighs the limitations, not only quantitatively but also qualitatively. Especially, the essential claims that require the most backing as indicated in the IUA, can be considered sufficiently supported so that the RobS can be implemented to further explore those claims that require actual *use*.

18 Looking Back – Looking Forward

In this final chapter, a general review of the dissertation is provided. A summary of the previous chapters is given first. Then, limitations that relate to the general research process are drawn before implications for the future use of the RobS are presented.

18.1 Summary

The aim of this thesis is to develop and validate a RobS in the EFL context that can be used for the summative assessment of PSETs' feedback competence on writing. In the following, the thesis will be summarized based on its specific parts and chapters.

Part I provides the foundation for this thesis. In Chapter 2, the goal and the three-phasic structure of teacher education in Germany are outlined. In Chapter 3, the term competence is discussed and specific competence models for teacher education are presented. The statutory guidelines from the KMK are

outlined and empirical findings that seek to evaluate the extent to which a development of knowledge occurs in EFL teacher education are described. It is argued that PSTs do not feel adequately prepared for practice and face challenges in applying their knowledge in concrete situations. In Chapter 4, a performance-oriented approach to the first phase of teacher education is outlined, based on the framework of constructive alignment (Biggs, 1996) and the core practice approach (Grossman, 2018; Grossman & Fraefel, 2024). In that, simulations are presented as a suitable, authentic assessment format. Using Miller's framework for clinical assessment (1990), a gap in the assessment formats in teacher education can be identified, as the level of *shows how* is largely omitted in teacher education. In medical education, simulations are used as authentic, performance-oriented assessment formats, often with actors as standardized patients in role-play-based scenarios to address this level (Schnabel, 2018), but not necessarily in teacher education (J. Meier et al., 2023, p. 91). Thus, the aim of this thesis, to develop and validate such a performance-oriented, role-play-based simulation for summative assessment purposes in the EFL context is derived.

Part II focuses on the test that is developed in this thesis. In Chapter 6, the reasons for focusing on feedback on writing in the developed test are outlined. Feedback is defined, based on a socio-constructive and dialogic understanding (Schluer, 2022, p. 17). The justification for feedback on writing in the EFL classroom as the focus of the RobS is derived from the necessity described in the statutory guidelines, the theoretical models of feedback and writing, the empirical findings that emphasize the beneficial impact of feedback on writing, and the lack of actual implementation, suggesting that an overall stronger focus should be placed on feedback in EFL teacher education. In Chapter 7, the theoretical framework for the RobS development is described. Several perspectives are included to allow for a holistic view of the feedback conversation in the RobS. This includes that of general models of feedback, like that from Hattie and Timperley (2007) or Narciss (2006), communicative and adaptive aspects (Aljaafreh & Lantolf, 1994; Bleichenbacher et al., 2017; Carless, 2020; Kopp & Mandl, 2014a; Rosenberg & Holler, 2016), and specific aspects that relate to L2/FL writing (Becker-Mrotzek, 2014; Ferris, 2013; Hayes & Flowers, 1980; Hyland, 2019; Lee, 2017). In Chapter 8, the RobS is presented according to its three phases: *Preparation*, *Simulation*, and *Assessment*. For the preparation, the assessment task and the material are described. For the simulation, the means of standardization are outlined. Last, the scoring instrument is presented for the assessment process.

Part III provides the methodological backbone for this thesis. In Chapter 10, potential criteria for test evaluation are described. This includes the secondary criteria of practicality, fairness and authenticity, as well as the primary criteria of

objectivity, reliability and validity. Validity is focused on in depth, as a modern understanding of validity subsumes the other criteria. The argument-based approach to validation, which understands validity as an ongoing process and as the extent to which evidence supports claims that are made regarding a specific use and interpretation of a test and its score, is presented as it is adapted in this thesis. In Chapter 11, the argument-based approach to validation is operationalized for this thesis based on the framework by Schreiber and Gut (2022). The *interpretation/use argument* is described with eight claims and corresponding warrants, which indicate the kind of evidence that is required for appropriate support.

In **Part IV**, the empirical findings are presented and discussed. In Chapter 13, the research design is described along two research cycles. The first focuses on the development of the material, the actor training and validity investigations within the (pre-) pilot phase. In the second cycle, the material is used to collect sufficient data in the main survey to conduct further investigations. In Chapter 14, results from the pre-pilot phase and the pilot survey are presented. Regarding the pre-pilot phase, the multiple steps taken in the development of the material are described, focusing on the investigations of the learner text's authenticity. During the actor training, the actors are prepared to portray their roles in a standardized way, relying on the role description and the VTs. In the pilot survey, the material and research procedure are tested and investigated with a small sample of the target group. The data collection process is described, including details on the follow-up part. As part of the pilot survey, the appropriateness of the VTs is explored in SRIs. The data suggests that the VTs elicit the desired competence facets to a sufficient extent. In Chapter 15, the findings of the main survey with $N = 51$ participants are presented and discussed in four distinct studies. In study 1, the participants' perspective on the RobS is focused on. Summing up the findings from the follow-up interviews, the participants view the RobS as rather authentic and fair. The participants' assumptions regarding tested competences align largely with the framework, and they have a generally positive attitude towards a potential implementation of the RobS for assessment purposes. In study 2, the evaluative statements from experts regarding selected RobS performances and the scoring instrument are explored. Regarding the performances, the results indicate that while experts do not necessarily agree on most criteria, a large portion of the criteria that are mentioned can be related to the framework and scoring instrument. Concerning the scoring instrument, expert's judgements are generally positive and critical comments are used to modify and improve the scoring instrument. In study 3, the reliability of the scoring is investigated based on the scoring of the full sample of $N = 51$ participants, as well as a subset of $n = 10$ performances, which is in focus in the inter-rater

reliability investigations. After an evaluation of the preliminary version of the scoring instrument and rating manual, modifications are made and described transparently. With the final set of 24 items, sufficient reliability coefficients concerning the internal consistency and inter-rater agreement are achieved on an item-level. On a test score-level, a sufficient range of test scores and distribution is generated in the sample and a high level of agreement is achieved among the raters. Concerning the impact of the individual actor on the test score, a possible influence can be explained, but not ruled out. In study 4, different hypotheses are stated and tested that are made based on *a priori* assumptions, using the final test scores and the additional collected data. H1, which states that the RobS score correlates significantly positively with the amount of school-related experience, is partially confirmed. Only the participants that have completed the SIS perform significantly better than those who have not. Arguably, the other assumptions need reconsideration, including the quality of the experience as well. H2, which states that the RobS score correlates significantly positively with academic progress and academic success, is confirmed. H3, which focuses on the correlation of the RobS score with other facets of the target domain, is partially confirmed. Pedagogically relevant facets of the target domain, i.e., SPK and, to a lesser extent, PCK, indicate more robust correlations with the RobS score than CK / LP. H4, which states that the RobS score correlates significantly positively with the expert ratings, is partially confirmed. While overall, the test score shows moderate agreement with the experts' ratings, individual, pairwise comparisons indicate that there is also a substantial heterogeneity among the experts' ratings, reflecting the qualitative findings and complicating a robust conclusion.

In **Part V**, the findings of this are discussed on a general level. In Chapter 17, the validity argument is presented. For each claim of the IUA the extent to which the evidence supports the warrants is discussed. Overall, the evidence is considered sufficient for the claims, although limitations are also addressed. In the end, the conclusion is drawn that, based on the findings, the test can be used with sufficient confidence to assess PSETs feedback competence on writing. In the following chapters, limitations are discussed and implications drawn.

18.2 Limitations

In this chapter, the overall research process is reflected upon and general limitations are outlined. Limitations that relate to individual studies, methods, or results are not focused here as they are discussed in the chapters that correspond to the findings. Rather, a meta-perspective will be taken with regard to the general test development and validation process.

Regarding the test development phase, some early decisions did not work out, in hindsight, for example, the role of language in the RobS. In the EFL

classroom, language plays a central role as it is not only the content focus of instruction, but also the medium. EFL teachers need to be language role models and provide opportunities for the learner to produce written and oral language, following the principle of functional monolingualism. However, it is essential that the learner can actually use the information from the feedback conversation and comprehension is more likely to occur if the conversation takes place in the SL's L1, German.

For the development of the RobS, the question therefore was: In which language *should* the feedback conversation take place? This question is complicated by the fact that, in the fictional context, the conversation takes place during the break. Thus, the situation does not take place *in* the EFL classroom albeit being related to it, further obscuring any clear norm. If English was considered to be the language that is normatively set as the target, for instance by prescribing it in the task description for the participants, this would cause certain ripple effects. First, the actors had to be instructed to react appropriately in English. Considering the competence level of the SL, it would not be authentic to have them speak English at a higher level than the hybrid learner text. For an authentic performance, mistakes would then have to be included in the SL's oral statements to some extent. This would have added another layer to the performance for the actors, complicating the role substantially, and potentially impeding an authentic perception overall. Moreover, this leads to the question of whether the participants should provide oral corrective feedback and the extent to which this should be included in scoring. If German were to be prescribed, these challenges would not occur. However, this would be in opposition with the outlined demand for functional monolingualism, and the teacher's function as a language role model. In the end, the decision was made to construct the RobS in such a way that the ability to adapt to the learner's level of language is in focus. This was considered to be the most appropriate course of action, as it allows for the normative demands to be included, while also facilitating the role of the actor. Furthermore, this allows adaptive capabilities to be focused on in the RobS from various perspectives, including the focus of adapting the level of language complexity to a comprehensible level for the learner.

While the operationalization of this demand was quite successful and the language switch was consistently elicited through the language of the material and the specific VTs, in the end, this item is excluded from the final scale as it indicates a lack of consistency with the remaining items (see Chapter 15.4.1.3.1). Possibly, the participants that fulfilled the item's demands for the maximum of two points, e.g., started in English and switched to German, were too focused on their own language performance or on keeping the conversation in English. Hence, they were not able to fulfill the remaining aspects of the RobS and thus

received less points overall. Considering that the issue of language plays an important role in the RobS, often requiring several turns between the participants and the SL until a switch takes place, the initial decision should be critically reflected upon. In hindsight, a clear instruction to speak German would have simplified the RobS for the participants and the actors and might have been the more suitable approach. Consequently, for future, adapted use, the VTs and the material might need to be modified as well. Future research, however, is necessary to explore the extent to which this would be considered authentic and appropriate for EFL teacher education.

Another limitation is found within the concept of the RobS itself – the standardization of the learner. The actors received extensive training to depict the role with consistency, using the same VTs and mannerisms outlined in the role description. Moreover, the material depicts a set of errors and mistakes that are, generally speaking, typical features of interlanguage. While this standardization is necessary to allow for a comparison between the different performances, it also represents a challenge in that it might perpetuate stereotypes. Participants might transfer the actors' appearance and behavior and overgeneralize it, assuming that similar, real learners behave in a similar manner. This stands in opposition to the heterogeneity and diversity that teachers deal with in classrooms, in which adaptivity to the learners' individual needs is essential. However, the RobS that is presented here is only designed to depict *one* possible learner in a standardized way. As the RobS is an *approximation of practice* (Grossman, Compton, et al., 2009, p. 2076), such simplifications need to be taken into account. Importantly, a debriefing should occur after simulations, so that the participants can reflect on this notion. Ideally, participants would also not only engage in one RobS for assessment, but in several different simulations throughout their professionalization, hopefully encountering a diverse variety of actors portraying the SL or other roles.

Regarding the sample in the pilot and main survey, two limitations must be addressed. Firstly, it is a convenience sample. Participants were recruited voluntarily, based on a diverse promotional campaign. The sample is thus certainly not representative and certain groups of students might not have been reached. This might relate especially to students who are of a less extrovert nature. If the RobS were, however, to be implemented as an assessment format, these students would also be the subjects of assessment. Consequently, the performances and perspectives of this group might be underrepresented in this sample. Naturally, this might refer to the over- or underrepresentation of other groups as well. Secondly, some of the participants participated as a part of the credit for their seminars. Although participation was still voluntary, this might have affected the quality of the interview data, which might be more prone to

effects of social desirability than, for instance, the performance in the RobS. This has to be considered for this portion of the surveys.

A further potential limitation is the decision to rely on individual coefficients of reliability in this study. Alternatively, a potentially relevant approach with regard to reliability is that of Generalizability theory (G theory) (Webb & Shaughnessy, 2005, p. 717). G theory considers a generated test score as one potential score of a universe of scores, and each facet of measurement might lead to measurement error, e.g., different raters, or the actors (Webb & Shaughnessy, 2005, p. 717). The universe of test scores then reflects all possible combinations of how the facets could influence measurement. In a G study, then, facets are identified and their influence on variance estimated (Webb & Shaughnessy, 2005, p. 717). Such an approach does not rely on individual coefficients of reliability, like Fleiss' κ , to describe the agreement between different raters. Rather, it considers the raters as one facet, i.e., one source of variance, and allows for an estimation of the extent to which the raters contribute to error, compared to other facets, like the actors. It is thus more comprehensive than a contemplation of individual coefficients. However, the approach taken here is considered sufficient for the explorative validation of the RobS, as the mathematical procedures for G studies potentially exceed the scope of this project. Future research could use the data and reevaluate the findings from this perspective.

What is more, a decision was made not to propose a RobS score – grade transformation table. In the context of the test score distribution, a potential pass threshold of 50 % of the points is discussed (see Chapter 15.4.2.3). However, such a benchmark is arbitrary and not based on the actual competence level that is considered to be appropriate to pass. Such a benchmark, and accordingly that of succeeding grades, should be set within the specific educational context of instruction and assessment. For instance, different score ranges could be used for different phases of study, accounting for the respective experience. One approach, then, is to make decisions regarding the minimum passing score of each item and determine different threshold levels for the specific grades¹⁶². In order to do so, further expert studies could be conducted that specifically focus on the specific item score requirements and determine these levels.

In the end, the findings of this thesis should be contextualized in general. In the argument-based approach to validation, validity is not a property of a test but is always dependent on the extent to which evidence supports the intended use (AERA, APA, NCME, 2014, p. 11). The empirical evidence and theoretical

¹⁶² Expert 5 proposes a similar approach in the context of the interviews.

assumptions presented in the VA and in this thesis in general, do indicate sufficient support as discussed. However, this conclusion is of a preliminary nature. It is all but likely, that individual findings are refuted in the future and individual conclusions have to be reconsidered. This may relate to future studies that require a reevaluation of the RobS' authenticity, towards which participants were generally positive. However, it does not necessarily have to be to the detriment of the validity. For instance, other studies on RobS might indicate that the influence of the individual actors in them is smaller than partially suggested in this study, further supporting the discussed reasons for this finding. Thus, the results should not be taken as a final judgement of the validity of the RobS, but as a conclusion made for the time being that holds true, as long as the evidence does.

18.3 Implications and Future Perspectives

In this final chapter, some implications will be made that point to future perspectives of RobS. Considering the findings and the outlined limitations, the RobS that is developed and presented in this thesis is arguably a suitable test that could be used in teacher education. It allows for the measurement of the feedback competence of PSETs in a performance-oriented setting and, assuming that enough raters are used, it can also be used to provide detailed information for the PSET concerning specific aspects of their performance. Therefore, the RobS successfully addresses the identified assessment gap in teacher education and, possibly, the desire of many PSTs who do not feel appropriately prepared for actual practice.

Most importantly, the RobS should not be considered in isolation, but rather in the context of constructive alignment. If a stronger focus on the performative aspect of competence is aimed for in the first phase of teacher education, the current learning objectives as stated in the statutory guidelines by the KMK need reassessment (see also Marks, 2023). Similarly, university courses should adequately prepare PSETs for the RobS by including TLAs that are more action- and performance-oriented. As discussed, the core practice approach as well as the idea of SBL might be appropriate concepts that allow PSETs to develop competence facets that prepare them not only for assessment in the RobS, but also, hopefully, for actual practice. Consequently, all three dimensions must always be considered and included in future implementations. While currently several projects are exploring the identification and development of core practices (see, for instance, Nückles & Fraefel, 2024 for an overview), more research is required for all three dimensions of CA, especially assessment. The RobS that is developed and evaluated in this thesis focuses on one essential core practice of EFL teachers. However, there are other, equally relevant core practices that have not been

included here, e.g., oral corrective feedback or lesson planning. Based on the promising findings, especially considering the perspective of the participants, future research should seek to explore the extent to which the development of these core practices could be implemented and assessed in the first phase of teacher education. This thesis offers just one potential approach that can be followed and modified to develop further simulations that focus on different core practices.

In the German context, an implementation of RobS is likely most suitable during the first phase of teacher education, especially during the Master of Education. Considering Miller's framework for clinical assessment and the assessment gap in teacher education, RobS should be implemented between traditional assessment formats that focus on aspects of knowledge and formats that consider behavior under the full complexity of the profession (see Chapter 4.5). As in the second phase of teacher education PSETs gain a substantial amount of teaching experience at school, they can use this opportunity to develop their competence under the full complexity of the classroom. Thus, assessment by means of classroom observation, i.e., *action* in Miller's framework, fits well within the second phase. RobS should therefore precede the second phase and be implemented in the first phase. In their Bachelor studies, PSETs are equipped with the necessary theoretical knowledge, which can be assessed with traditional exams. Due to the fact that RobS do not necessarily seek to replace existing formats that are successful in measuring the desired facets of competence in different phases, they should not be implemented in Bachelor programs. Moreover, the internships in the Bachelor studies have more of an explorative character for the PSETs and assessment of action-oriented competences is not aimed for yet (see Chapter 2.2). The Master programs, however, are usually more focused on the practical elements of teaching. In most of Germany's federal states, Master of Education programs also include an SIS as an extensive period of practical experience. Thus, RobS align well with this stage's aims and could be used for assessment purposes here.

Within the Master programs, RobS could be introduced before or after the SIS¹⁶³. Before the SIS, RobS could function as a means to diagnose the extent to which PSETs can deal with learners before they do so more extensively at school, presuming previous TLAs align with the action-oriented focus of the RobS. Given that PSETs pass the RobS, it can be assumed that they can interact with learners appropriately. Alternatively, RobS could also be used as an assessment format of the SIS, in which the engagement with the class could be viewed as a TLA. At

¹⁶³ In federal states without an SIS, a comparable more extensive internship could be seen as a substitute.

the moment, the summative assessment of the SIS usually focuses on the aspect of action-research and less on the actual practice (Porsch & Gollub, 2018, p. 240). RobS could therefore be used to either replace the existing assessment formats of the SIS or as an additional component, so that multiple aspects of the profession are assessed. Nevertheless, a structured TLA in the form of seminars at university seems to be the most suitable option. At university, the learning environment is more controlled, whereas during the SIS, PSETs spend most of their time at school and lecturers have less influence in how the PSETs are provided with or can make use of learning opportunities. Consequently, PSETs can be prepared for the RobS more precisely at university in specifically designed seminars. Thus, an implementation of RobS before the SIS, aligned with suitable learning opportunities at the university is the most favorable option.

A successful implementation of RobS requires a discussion that includes stakeholders of the first and second phase of teacher education to ensure a seamless transition between the phases. Stakeholders of the first phase are those that would need to actually use the RobS as an assessment format and adapt appropriate TLAs. Traditionally, the institutions of the second phase focus more on performance-oriented aspects of the profession and continue the PSETs' competence development after university. Hence, it is important that a stronger focus on performance in the first phase is appropriately coordinated with the second phase, so that an overall alignment across all phases is guaranteed. Moreover, a broad discussion between the phases might also increase the overall acceptance of RobS as an assessment format, as the implementation would be supported by a multitude of stakeholders. Possibly, Schools of Education could function as intermediaries, as most of them they already have a cross-phase function in the German teacher education system.

Considering the rise of artificial intelligence (AI), a discussion on new assessment formats seems even more relevant. On the one hand, assessment formats that focus on the recall of information are becoming less relevant in a world in which AI can take on these functions (Bearman et al., 2020). On the other hand, several traditional assessment formats in the first phase of teacher education rely on written work, such as essays or term papers. However, cheating is facilitated by the use of AI large language models like ChatGPT, as students can easily use AI to solve the assignments. Consequently, these assessment formats might no longer be suitable to actually ensure that students have achieved the learning objectives (Dawson et al., 2024). RobS are one possible alternative approach to assessment, in which it is highly unlikely that students can cheat using AI or other means. RobS, therefore, might also contribute to an overall higher degree of fairness as cheating is likely to be reduced.

In the end, performance-oriented assessment formats like the RobS are of course not the ultimate solution to the complex challenges facing teacher education. If at all, they are but one tool that can be used in an overall revised approach to university-based teacher education, one that seeks to focus on performance and practice to a greater extent. Hopefully, this might increase teachers' impact and lead to better learner achievement in the end, but future research needs to explore the longer-lasting effects. Nevertheless, the collected evidence and the inferences made in this thesis have shown that within a performance-oriented constructive alignment, the RobS is a suitable format to assess feedback competence on writing in the EFL context. Hence, the RobS represents a possible addition to the existing assessment formats in teacher education, which enables future teachers of English to *show how* they provide feedback, rather than just *tell* lecturers about it.

19 References

- Ade-Ojo, G. O., Markowski, M., Essex, R., Stiell, M., & Jameson, J. (2022). A systematic scoping review and textual narrative synthesis of physical and mixed-reality simulation in pre-service teacher training. *Journal of Computer Assisted Learning*, 38(3), 861–874.
<https://doi.org/10.1111/jcal.12653>
- AERA, APA, NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Afzal, E., Fecke, J., Alboaie, L., & Braun, E. (2024). Training an Avatar-Based Conversational AI Chatbot and Investigating the Implementation of Communicative Action Principles. In *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning - CSCL 2024* (pp. 389–390). International Society of the Learning Sciences.
<https://doi.org/10.22318/cscl2024.839190>
- Aguado, K. (2018). Lautes Denken als Datenerhebungsverfahren in der empirischen Fremdsprachenforschung. In K. Aguado, C. Finkbeiner, & B. Tesch (Eds.), *Language Culture Literacy: Vol. 10. Lautes Denken, «Stimulated Recall» und Dokumentarische Methode: Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lernforschung* (pp. 9–25). Peter Lang.
- Aguado, K., Finkbeiner, C., & Tesch, B. (Eds.). (2018). *Language Culture Literacy: Vol. 10. Lautes Denken, «Stimulated Recall» und Dokumentarische Methode: Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lernforschung*. Peter Lang. <https://doi.org/10.3726/b13433>
- Ajjawi, R., & Boud, D. (2017). Researching feedback dialogue: an interactional analysis approach. *Assessment & Evaluation in Higher Education*, 42(2), 252–265. <https://doi.org/10.1080/02602938.2015.1102863>
- Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2024). From authentic assessment to authenticity in assessment: broadening perspectives. *Assessment & Evaluation in Higher Education*, 499–510.
<https://doi.org/10.1080/02602938.2023.2271193>
- Albu, C., & Lindmeier, A. (2023). Performance assessment in teacher education research—A scoping review of characteristics of assessment instruments in the DACH region. *Zeitschrift Für Erziehungswissenschaft*, 26(3), 751–778. <https://doi.org/10.1007/s11618-023-01167-7>
- Aljaafreh, A., & Lantolf, J. P. (1994). Negative Feedback as Regulation and Second Language Learning in the Zone of Proximal Development. *The Modern Language Journal*, 78(4), 465–483.

- Angelini, M. L. (2023). Simulation in Teacher Preparation. In M. L. Angelini & R. Muñiz (Eds.), *Springer Texts in Education. Simulation for Participatory Education* (pp. 43–51). Springer International Publishing.
https://doi.org/10.1007/978-3-031-21011-2_3
- Angelini, M. L., Diamanti, R., & Aguilar-Moya, R. (2024). Predictive Factors for Difficulty in Simulation Methodology in Teacher Education. *Simulation & Gaming*, 55(3), 502–526. <https://doi.org/10.1177/10468781241234812>
- Ashwell, T. (2000). Patterns of Teacher Response to Student Writing in a Multiple-Draft Composition Classroom: Is Content Feedback Followed by Form Feedback the Best Method? *Journal of Second Language Writing*, 9(3), 227–257. [https://doi.org/10.1016/S1060-3743\(00\)00027-8](https://doi.org/10.1016/S1060-3743(00)00027-8)
- Bajis, D., Chaar, B., Basheti, I. A., & Moles, R. (2019). Pharmacy students' medication history taking competency: Simulation and feedback learning intervention. *Currents in Pharmacy Teaching & Learning*, 11(10), 1002–1015. <https://doi.org/10.1016/j.cptl.2019.06.007>
- Bandura, A. (1997). *Self-efficacy: The exercise of control* (13th ed.). Freeman.
- Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, 61(2), 213–238. <https://doi.org/10.3102/00346543061002213>
- Barahona, M. (2020). The potential of translanguaging as a core teaching practice in an EFL context. *System*, 95, 102368. <https://doi.org/10.1016/j.system.2020.102368>
- Barras, M. (2018). Stimulated Recall in der Sprachtestforschung. Ein praktisches Beispiel aus der Erprobung eines computerbasierten Leseverstehens-tests. In K. Aguado, C. Finkbeiner, & B. Tesch (Eds.), *Language Culture Literacy: Vol. 10. Lautes Denken, «Stimulated Recall» und Dokumentarische Methode: Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lernforschung* (pp. 69–86). Peter Lang.
- Bartels, H. (2018). *Entwicklung und Bewertung eines performanznahen Videovignetentests zur Messung der Erklärfähigkeit von Physiklehrkräften*. Dissertation. *Studien zum Physik- und Chemielernen: Vol. 264*. Logos Verlag.
- Baumert, J. (Ed.). (2001). *OECD PISA. PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Leske + Budrich.
- Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift Für Erziehungswissenschaft*, 9(4), 469–520.
<https://doi.org/10.1007/s11618-006-0165-2>
- Baumert, J., & Kunter, M. (2013). The COACTIV Model of Teachers' Professional Competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers* (pp. 25–48). Springer US.

- Baumert, J., Kunter, M., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (2013). Professional Competence of Teachers, Cognitively Activating Instruction, and the Development of Students' Mathematical Literacy (CO-ACTIV): A Research Program. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers* (pp. 1–21). Springer US. https://doi.org/10.1007/978-1-4614-5149-5_1
- Baurmann, J. (2017). *Schreiben - Überarbeiten - Beurteilen: Ein Arbeitsbuch zur Schreibdidaktik* (5th ed.). *Praxis Deutsch*. Klett | Kallmeyer.
- Bearman, M., Dawson, P., Ajjawi, R., Tai, J., & Boud, D. (Eds.). (2020). *The Enabling Power of Assessment. Re-imagining University Assessment in a Digital World*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-41956-1>
- Becker-Mrotzek, M. (2014). Schreibleistungen bewerten und beurteilen. In H. Feilke & T. Pohl (Eds.), *Deutschunterricht in Theorie und Praxis: Vol. 4. Schriftlicher Sprachgebrauch - Texte verfassen* (pp. 501–513). Schneider-Verlag Hohengehren.
- Becker-Mrotzek, M., & Böttcher, I. (2020). *Schreibkompetenz entwickeln und beurteilen* (9th ed.). *Scriptor Praxis. Sekundarstufe I + II*. Cornelsen.
- Behney, J., & Marsden, E. (2020). Introduction to SLA. In N. Tracy-Ventura & M. Paquot (Eds.), *Routledge handbooks. The Routledge handbook of second language acquisition and corpora* (pp. 37–49). Routledge Taylor & Francis Group.
- Bereiter, C., & Scardamalia, M. (1987). *The Psychology of Written Composition. The psychology of education and instruction series*. Routledge.
- Bergmann, J. R., & Meyer, C. (2021). Reflexivity, Indexicality, Accountability. In J. R. Bergmann & C. Meyer (Eds.), *Media in Action. Ethnomethodologie re-loaded* (Vol. 1, pp. 37–54). Transcript Verlag. <https://doi.org/10.14361/9783839454381-003>
- Besser, M., Leiss, D., & Klieme, E. (2015). Wirkung von Lehrerfortbildungen auf Expertise von Lehrkräften zu formativem Assessment im kompetenzorientierten Mathematikunterricht. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, 47(2), 110–122. <https://doi.org/10.1026/0049-8637/a000128>
- Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *ETS Research Report Series*, i-99.
- Biesta, G. (2016). Who's Afraid of Teaching? Heidegger and the Question of Education ('Bildung'/'Erziehung'). *Educational Philosophy and Theory*, 48(8), 832–845. <https://doi.org/10.1080/00131857.2016.1165017>

- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347–364. <https://doi.org/10.1007/BF00138871>
- Biggs, J. (2003). *Aligning Teaching and Assessing to Course Objectives*, University of Aveiro.
- Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university: What the student does* (4th ed.). Open University Press.
- Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Bland, A. J., Topping, A., & Tobbell, J. (2014). Time to unravel the conceptual confusion of authenticity and fidelity and their contribution to learning within simulation-based nurse education. A discussion paper. *Nurse Education Today*, 34(7), 1112–1118. <https://doi.org/10.1016/j.nedt.2014.03.009>
- Bleichenbacher, L., Hilbe, R., Klee, P., Kuster, W., & Roderer, T. (2017). *Beurteilung berufsspezifischer Sprachkompetenzen von Lehrpersonen, die Fremdsprachen unterrichten: Projektresultate*. <https://www.phsg.ch/sites/default/files/cms/Forschung/Institute/Institut-Fachdidaktik-Sprachen/201711%20BSSKP%2015-17%20Produktbericht%20Webversion.pdf> [Last retrieved: January 02, 2025]
- Blömeke, S. (2006). KMK-Standards für die LehrerInnenbildung in Deutschland: Ein Kommentar. *Journal Für Lehrerinnen- Und Lehrerbildung*(1), 25–33.
- Blömeke, S. (2009). Ausbildungs- und Berufserfolg im Lehramtsstudium im Vergleich zum Diplom-Studium – Zur prognostischen Validität kognitiver und psycho-motivationaler Auswahlkriterien. *Zeitschrift Für Erziehungswissenschaft*, 12(1), 82–110. <https://doi.org/10.1007/s11618-008-0044-0>
- Blömeke, S. (2011). Teacher Education and Development Study: Learning to Teach (TEDS-LT) - Erfassung von Lehrerkompetenzen in gering strukturierten Domänen. In S. Blömeke, K. Schwippert, A. Bremerich-Vos, H. Haudeck, G. Kaiser, G. Nold, & H. Willenberg (Eds.), *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen: Erste Ergebnisse aus TEDS-LT* (pp. 7–24). Waxmann.
- Blömeke, S., Bremerich-Vos, A., Kaiser, G., Nold, G., Haudeck, H., Keßler, J.-U., & Schwippert, K. (Eds.). (2013). *Professionelle Kompetenzen im Studienverlauf: Weitere Ergebnisse zur Deutsch-, Englisch- und Mathematiklehrerbildung aus TEDS-LT*. Waxmann.

- Blömeke, S., Buchholtz, C., & Bremerich-Vos, A. (2013). Zusammenhang institutioneller Merkmale mit dem Wissenserwerb. In S. Blömeke, A. Bremerich-Vos, G. Kaiser, G. Nold, H. Haudeck, J.-U. Keßler, & K. Schwippert (Eds.), *Professionelle Kompetenzen im Studienverlauf: Weitere Ergebnisse zur Deutsch-, Englisch- und Mathematiklehrausbildung aus TEDS-LT* (pp. 167–188). Waxmann.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond Dichotomies: Competence Viewed as a Continuum. *Zeitschrift Für Psychologie*, 223(1), 3–13. <https://doi.org/10.1027/2151-2604/a000194>
- Blömeke, S., Schwippert, K., Bremerich-Vos, A., Haudeck, H., Kaiser, G., Nold, G., & Willenberg, H. (Eds.). (2011). *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen: Erste Ergebnisse aus TEDS-LT*. Waxmann.
- Bocklisch, F., Bocklisch, S. F., & Krems, J. F. (2012). Sometimes, often, and always: Exploring the vague meanings of frequency expressions. *Behavior Research Methods*, 44(1), 144–157. <https://doi.org/10.3758/s13428-011-0130-8>
- Bonnett, A. (2024). Towards the hidden side of the hidden side. Rekonstruktive Professionsforschung in der deutschsprachigen Fremdsprachendidaktik. In D. Gerlach (Ed.), *Wissen, Können und Handeln von Fremdsprachenlehrpersonen: Professions- und Wissensforschung in den Fremdsprachendidaktiken* (pp. 31–48). Peter Lang.
- Börner, W. (1987). Schreiben im Fremdsprachenunterricht: Überlegungen zu einem Modell. In W. Lörcher (Ed.), *Tübinger Beiträge zur Linguistik: Vol. 317. Perspectives on language in performance: Studies in linguistics, literary criticism and language teaching and learning; to honour Werner Hüllen on the occasion of his 60. birthday* (pp. 1336–1349). Narr.
- Börner, W. (1989). Planen und Problemlösen im fremdsprachlichen Schreibprozeß: Einige empirische Befunde. In U. Klenk (Ed.), *Variatio linguarum: Beiträge zu Sprachvergleich und Sprachentwicklung; Festschrift zum 60. Geburtstag von Gustav Ineichen* (pp. 43–62). Steiner-Verl.-Wiesbaden-GmbH.
- Bortz, J., & Lienert, G. A. (2008). *Kurzgefasste Statistik für die klinische Forschung: Leitfaden für die verteilungsfreie Analyse kleiner Stichproben; mit 13 Abbildungen und 97 Tabellen sowie zahlreichen Formeln* (3rd ed.). Springer-Lehrbuch Bachelor, Master. Springer.
- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7th ed.). Springer-Lehrbuch. Springer.
- Böttcher, I., & Becker-Mrotzek, M. (2009). *Texte bearbeiten, bewerten und benoten* (4th ed.). Lehrer-Bücherei Grundschule. Cornelsen Scriptor.

- Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698–712. <https://doi.org/10.1080/02602938.2012.691462>
- Braun, E. (2021). Performance-based assessment of students' communication skills. *International Journal of Chinese Education*, 12(1), 1-12. <https://doi.org/10.1177/22125868211006202>
- Braun, E., Athanassiou, G., & Gockel, Svea, Pollerhoff, Kethleen. (2016). KomPrü – Performance-based Assessment of Students' Communication Skills. In H. A. Pant, O. Zlatkin-Troitschanskaia, C. Lautenbach, M. Toepper, & D. Molerov (Eds.), *Modeling and Measuring Competencies in Higher Education: - Validation and Methodological Innovations (KoKoHs) Overview of the Research Projects* (pp. 11–14).
- Braun, E., Athanassiou, G., Pollerhof, K., & Schwabe, U. (2018). Wie lassen sich kommunikative Kompetenzen messen? Konzeption einer kompetenz-orientierten Prüfung kommunikativer Fähigkeiten von Studierenden. *Beiträge Zur Hochschulforschung*, 40(3), 34–55.
- Braun, E., Schwabe, U., & Klein, D. (2018). Performance-Based Tests: Using Role Plays to Assess Communication Skills. In S. McGrath, M. Mulder, J. Papier, & R. Stuart (Eds.), *Handbook of Vocational Education and Training* (pp. 1–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-49789-1_82-1
- Bresges, A., Dilger, B., Hennemann, T., König, J., Lindner, H., Rohde, A., & Schmeinck, D. (2014). Einleitung. In A. Bresges, B. Dilger, T. Hennemann, J. König, H. Lindner, & A. Rohde (Eds.), *Kompetenzen diskursiv: Terminologische, exemplarische und strukturelle Klärungen in der LehrerInnenbildung* (1st ed., pp. 9–14). Waxmann.
- Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F. E. Weinert (Ed.), *Enzyklopädie der Psychologie: Vol. 3. Enzyklopädie der Psychologie* (pp. 177–212). Hogrefe.
- Bromme, R. (2008). Lehrerexpertise: Eine psychologische Konzeption für die Entwicklung und Erforschung des Wissens und Könnens von Lehrern. In W. Schneider & M. Hasselhorn (Eds.), *Handbuch der Psychologie: Vol. 10. Handbuch der pädagogischen Psychologie* (159-167). Hogrefe.
- Brooks, C., Carroll, A., Gillies, R., & Hattie, J. (2019). A Matrix of Feedback. *Australian Journal of Teacher Education*, 44(4), 14–32. <https://doi.org/10.14221/ajte.2018v44n4.2>

- Buholzer, A., Baer, M., Zulliger, S., Torchetti, L., Ruelmann, M., Häfliger, A., & Lötscher, H. (2020). Formatives Assessment im alltäglichen Mathematikunterricht von Primarlehrpersonen: Häufigkeit, Dauer und Qualität. *Unterrichtswissenschaft*, 48(4), 629–661. <https://doi.org/10.1007/s42010-020-00083-7>
- Busse, V. (2019). Schriftliches Feedback in sprach- und geisteswissenschaftlichen Fächern: Die schriftsprachlichen Fähigkeiten fördern. *Friedrich Jahresheft*, 24–27.
- Busse, V. (2022). Schreiben fachübergreifend fördern: Einführung und theoretische Grundlagen. In V. Busse, N. Müller, & L. Siekmann (Eds.), *Schreiben fachübergreifend fördern: Grundlagen und Anregungen für Schule, Unterricht und Lehrkräftebildung* (1st ed., pp. 6–20). Klett | Kallmeyer.
- Cañadas, L. (2023). Contribution of formative assessment for developing teaching competences in teacher education. *European Journal of Teacher Education*, 46(3), 516–532. <https://doi.org/10.1080/02619768.2021.1950684>
- Carless, D. (2020). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education*, 23(2), 143–153. <https://doi.org/10.1177/1469787420945845>
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Carless, D., & Winstone, N. (2020). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, 28(1), 150–163. <https://doi.org/10.1080/13562517.2020.1782372>
- Carlson, H. L. (1999). From Practice to Theory: a social constructivist approach to teacher education. *Teachers and Teaching*, 5(2), 203–218. <https://doi.org/10.1080/1354060990050205>
- Caspari, D., & Grünewald, A. (2022). Auf dem Weg zur Bestimmung fachdidaktischen Wissens angehender Lehrkräfte - Ergebnisse einer Umfrage unter Lehrenden an Universitäten und Fachhochschulen. *Zeitschrift Für Fremdsprachenforschung*, 33(1), 97–121.
- Cavanaugh, S. (2022). Microteaching: Theoretical Origins and Practice. *Educational Practice and Theory*, 44(1), 23–40. <https://doi.org/10.7459/ept/44.1.03>
- Chan, K. K. H. (2023). Eliciting and working with student thinking: Preservice science teachers' enactment of core practices when orchestrating collaborative group work. *Journal of Research in Science Teaching*, 60(5), 1014–1052. <https://doi.org/10.1002/tea.21823>

- Chappelle, C. A., & Lee, H. (2021). Conceptions of validity. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 17–31). Routledge. <https://doi.org/10.4324/9781003220756-3>
- Chenoweth, A., & Hayes, J. R. (2001). Fluency in Writing: Generating Text in L1 and L2. *Written Communication*, 18(1), 80–98. <https://doi.org/10.1177/0741088301018001004>
- Chenoweth, A., & Hayes, J. R. (2003). The Inner Voice in Writing. *Written Communication*, 20(1), 99–118. <https://doi.org/10.1177/0741088303253572>
- Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., & Fischer, F. (2020). Facilitating Diagnostic Competences in Higher Education—a Meta-Analysis in Medical and Teacher Education. *Educational Psychology Review*, 32(1), 157–196. <https://doi.org/10.1007/s10648-019-09492-2>
- Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-Based Learning in Higher Education: A Meta-Analysis. *Review of Educational Research*, 90(4), 499–541. <https://doi.org/10.3102/0034654320933544>
- Chernikova, O., Holzberger, D., Heitzmann, N., Stadler, M., Seidel, T., & Fischer, F. (2024). Where salience goes beyond authenticity. *Zeitschrift Für Pädagogische Psychologie*, 38(1-2), 15–25. <https://doi.org/10.1024/1010-0652/a000357>
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>
- Codreanu, E., Huber, S., Reinhold, S., Sommerhoff, D., Neuhaus, B. J., Schmidmaier, R., Ufer, S., & Seidel, T. (2022). Diagnosing Mathematical Argumentation Skills: A Video-Based Simulation for Pre-Service Teachers. In F. Fischer & A. Opitz (Eds.), *Springer eBook Collection. Learning to Diagnose with Simulations: Examples from Teacher Education and Medical Education* (1st ed., pp. 33–47). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-89147-3_4
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- Cohen, S. A. (1987). Instructional Alignment: Searching for a Magic Bullet. *Educational Researcher*, 16(8), 16. <https://doi.org/10.2307/1175370>

- Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. *Academic Medicine: Journal of the Association of American Medical Colleges*, 88(6), 872–883. <https://doi.org/10.1097/ACM.0b013e31828ffdcf>
- COSIMA Research Unit. (n.d.). *Facilitating diagnostic competences in simulation-based learning environments in the university context*. <https://www.en.for2385.uni-muenchen.de/index.html> [Last retrieved: January 02, 2025]
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. <http://student.unifr.ch/pluriling/assets/files/Referenzrahmen2001.pdf>
- Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment; companion volume*. Council of Europe Publishing.
- Couper-Kuhle, E., Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzluft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., . . . Uhmann, S. (2011). A system for transcribing talk-in-interaction: GAT 2. *Gesprächsforschung*, 12, 1–51.
- Cramer, C., König, J., Rothland, M., & Blömeke, S. (Eds.). (2020). *Handbuch Lehrerinnen- und Lehrerbildung*. utb.
- Cronbach, L. J. (1971). Test validity. In R. L. Thorndike (Ed.), *Educational measurement* (2. ed., pp. 443–507). American Council on Education.
- Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 3–17). Lawrence Erlbaum Associates, Inc.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>
- Dahlstrom, N., Dekker, S., van Winsen, R., & Nyce, J. (2009). Fidelity and validity of simulator training. *Theoretical Issues in Ergonomics Science*, 10(4), 305–314. <https://doi.org/10.1080/14639220802368864>
- Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education*, 49(7), 1005–1016. <https://doi.org/10.1080/02602938.2024.2386662>
- De Florio-Hansen, I. (2018). *Teaching and learning English in the digital age*. Waxmann. <https://doi.org/10.36198/9783838549545>
- De Florio-Hansen, I. (2023). *From assessment to feedback: Applications in the second/foreign language classroom*. Cambridge University Press. <https://doi.org/10.1017/9781009218948>

- Deci, E. L., & Ryan, R. M. (2011). Self-Determination Theory. In P. A. M. van Lange (Ed.), *SAGE Social Psychology Program. Handbook of Theories of Social Psychology: Volume One* (pp. 416–437). SAGE Publications.
- Deibl, I., Zumbach, J., Geiger, V. M., & Neuner, C. M. (2018). Constructive Alignment in the Field of Educational Psychology: Development and Application of a Questionnaire for Assessing Constructive Alignment. *Psychology Learning & Teaching*, 17(3), 293–307.
<https://doi.org/10.1177/1475725718791050>
- Delanoy, W., Eisenmann, M., & Matz, F. (Eds.). (2015). *Anglo-American Studies: Vol. 49. Learning with literature in the EFL classroom*. Peter Lang Edition.
- Dewey, J. (1904). The relation of theory to practice in education. In C. A. McMurtry (Ed.), *The relation between theory and practice in the education of teachers: Third yearbook of the national society for the scientific study of education, Part 1* (pp. 9–30). The University of Chicago Press.
- Dewey, M. (2007). English as a lingua franca and globalization: an interconnected perspective. *International Journal of Applied Linguistics*, 17(3), 332–354. <https://doi.org/10.1111/j.1473-4192.2007.00177.x>
- Dieker, L. A., Hynes, M. C., Hughes, C. E., Hardin, S., & Becht, K. (2015). TLE TeachLivE™: Using Technology to Provide Quality Professional Development in Rural Schools. *Rural Special Education Quarterly*, 34(3), 11–16.
<https://doi.org/10.1177/875687051503400303>
- Dieker, L. A., Rodriguez, J. A., Lignugaris/Kraft, B., Hynes, M. C., & Hughes, C. E. (2014). The Potential of Simulated Environments in Teacher Education: Current and Future Possibilities. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, 37(1), 21–33. <https://doi.org/10.1177/0888406413512683>
- Diez, M. E. (2010). It Is Complicated: Unpacking the Flow of Teacher Education's Impact on Student Learning. *Journal of Teacher Education*, 61(5), 441–450. <https://doi.org/10.1177/0022487110372927>
- Doff, S. (2019). English Language (Teacher) Education in Germany after 1945. In R. Hickey (Ed.), *English in the German-Speaking World* (pp. 96–113). Cambridge University Press. <https://doi.org/10.1017/9781108768924.006>
- Döring, N. (2023). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (6th ed.). *Lehrbuch*. Springer. <http://www.springer.com/> <https://doi.org/10.1007/978-3-662-64762-2>
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5th ed.). *SpringerLink Bücher*. Springer. <http://link.springer.com/book/10.1007/978-3-642-41089-5>
<https://doi.org/10.1007/978-3-642-41089-5>

- Dormann, C. F. (2017). *Parametrische Statistik: Verteilungen, maximum likelihood und GLM in R* (2nd ed.). *Statistik und ihre Anwendungen*. Springer Spektrum. <https://doi.org/10.1007/978-3-662-54684-0>
- Dotger, B. H., Dotger, S. C., & Maher, M. J. (2010). From Medicine to Teaching: The Evolution of the Simulated Interaction Model. *Innovative Higher Education*, 35(3), 129–141. <https://doi.org/10.1007/s10755-009-9128-x>
- Dotger, B. H., Harris, S., & Hansel, A. (2008). Emerging authenticity: the crafting of simulated parent–teacher candidate conferences. *Teaching Education*, 19(4), 337–349. <https://doi.org/10.1080/10476210802438324>
- Dresing, T., & Pehl, T. (2018). *Praxisbuch Interview, Transkription & Analyse. Anleitungen und Regelsysteme für qualitativ Forschende*. (8. Aufl.). Dr. Dresing und Pehl GmbH.
- Ellis, R. (2009). Corrective Feedback and Teacher Development. *L2 Journal*, 1(1). <https://doi.org/10.5070/l2.v1i1.9054>
- Elton, L., & Laurillard, D. M. (1979). Trends in research on student learning. *Studies in Higher Education*, 4(1), 87–102. <https://doi.org/10.1080/03075077912331377131>
- Erlam, R., Ellis, R., & Batstone, R. (2013). Oral corrective feedback on L2 writing: Two approaches compared. *System*, 41(2), 257–268. <https://doi.org/10.1016/j.system.2013.03.004>
- Fecke, J., Afzal, E., & Braun, E. (2023). A Conceptual System Design for Teacher Education: Role-Play Simulations to Train Communicative Action with AI Agents. In *Proceedings of the 17th International Conference of the Learning Sciences - ICLS 2023* (pp. 2197–2200). International Society of the Learning Sciences. <https://doi.org/10.22318/icls2023.197799>
- Fecke, J., & Müller, L. Simulationen in virtuellen Lernumgebungen: Welche Vor- und Nachteile haben avatarbasierte und videokamerabasierte Formate bei der Durchführung von Rollenspielen? *Zeitschrift Für Hochschulentwicklung*, 17(1), 215–232. <https://doi.org/10.3217/zfhe-17-01/13>
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
- Fenstermacher, G. D. (1994). Chapter 1: The Knower and the Known: The Nature of Knowledge in Research on Teaching. *Review of Research in Education*, 20(1), 3–56. <https://doi.org/10.3102/0091732X020001003>
- Ferris, D. R. (1995). Student Reactions to Teacher Response in Multiple-Draft Composition Classrooms. *TESOL Quarterly*, 29(1), 33–35. <https://doi.org/10.2307/3587804>

- Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing*, 8(1), 1–11. [https://doi.org/10.1016/S1060-3743\(99\)80110-6](https://doi.org/10.1016/S1060-3743(99)80110-6)
- Ferris, D. R. (2010). Second Language Writing Research and Written Corrective Feedback in SLA. *Studies in Second Language Acquisition*, 32(2), 181–201. <https://doi.org/10.1017/S0272263109990490>
- Ferris, D. R. (2013). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing*, 19, 6–23. <https://doi.org/10.1016/j.asw.2013.09.004>
- Ferris, D. R., Liu, H [Hsiang], Sinha, A., & Senna, M. (2013). Written corrective feedback for individual L2 writers. *Journal of Second Language Writing*, 22(3), 307–329. <https://doi.org/10.1016/j.jslw.2012.09.009>
- Field, A. (2009). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)* (3rd ed.). Sage Publ.
- Field, K. (2004). *Issues in Modern Foreign Languages Teaching* (1st ed.). Routledge. <https://doi.org/10.4324/9780203005682>
- Fink, M. C., Robinson, S. A., & Ertl, B. (2024). AI-based avatars are changing the way we learn and teach: benefits and challenges. *Frontiers in Education*, 9, Article 1416307. <https://doi.org/10.3389/feduc.2024.1416307>
- Finkbeiner, C., & Schluer, J. (2018). Stimulated Recall als Zugang zu mentalen Prozessen in der Fremdsprachenforschung. In K. Aguado, C. Finkbeiner, & B. Tesch (Eds.), *Language Culture Literacy: Vol. 10. Lautes Denken, «Stimulated Recall» und Dokumentarische Methode: Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lernforschung* (pp. 27–50). Peter Lang.
- Fischer, J [Juan], Bearman, M., Boud, D., & Tai, J. (2024). How does assessment drive learning? A focus on students' development of evaluative judgement. *Assessment & Evaluation in Higher Education*, 49(2), 233–245. <https://doi.org/10.1080/02602938.2023.2206986>
- Fischer, J [Julian], Machts, N., Bruckermann, T., Möller, J., & Harms, U. (2022). The Simulated Classroom Biology—A simulated classroom environment for capturing the action-oriented professional knowledge of pre-service teachers about evolution. *Journal of Computer Assisted Learning*, 38(6), 1765–1778. <https://doi.org/10.1111/jcal.12718>
- Fischer, J [Julian], Machts, N., Möller, J., & Harms, U. (2021). Der Simulierte Klassenraum Biologie – Erfassung deklarativen und prozeduralen Wissens bei Lehramtsstudierenden der Biologie. *Zeitschrift Für Didaktik Der Naturwissenschaften*, 27(1), 215–229. <https://doi.org/10.1007/s40573-021-00136-z>

- Flores, M. A., Veiga Simão, A. M., Barros, A., & Pereira, D. (2015). Perceptions of effectiveness, fairness and feedback of assessment methods: a study in higher education. *Studies in Higher Education*, 40(9), 1523–1534. <https://doi.org/10.1080/03075079.2014.881348>
- Flower, L., & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*, 32(4), 365–387. <https://doi.org/10.2307/356600>
- Forzani, F. M. (2014). Understanding “Core Practices” and “Practice-Based” Teacher Education. *Journal of Teacher Education*, 65(4), 357–368. <https://doi.org/10.1177/0022487114533800>
- Fraefel, U. (2022). Mittels Kernpraktiken zu professionellem Unterrichten. *Journal Für LehrerInnenbildung*, 22(3), 16–29. <https://doi.org/10.35468/jlb-03-2022-01>
- Fraefel, U., & Grossman, P. (2024). Future Directions for the Field. In P. Grossman & U. Fraefel (Eds.), *Core practices in education. Core practices in teacher education: A global perspective* (pp. 139–148). Harvard Education Press.
- Freeman, D. (2017). The Case for Teachers’ Classroom English Proficiency. *RELC Journal*, 48(1), 31–52. <https://doi.org/10.1177/0033688217691073>
- Frey, A., & Jung, C. (Eds.). (2011). Kompetenzmodelle, Standardmodelle und Professionsstandards in der Lehrerbildung: Stand und Perspektiven [Special issue]. *Lehrerbildung auf dem Prüfstand*. Verlag Empirische Pädagogik.
- Fürstenau, B. (2019). *Lehr-Lern-Theorien: Behaviorismus, Kognitivismus, Konstruktivismus: Lernen und Expertise verstehen und fördern* (2nd ed.). *Studentexte Basiscurriculum Berufs- und Wirtschaftspädagogik: Vol. 6*. Schneider Verlag Hohengehren.
- Gallagher, G. (2017). Aligning for learning: Including feedback in the Constructive Alignment model. *All Ireland Journal of Higher Education*, 9(1), 3011–30112.
- Gartmeier, M., Bauer, J., Fischer, M. R., Hoppe-Seyler, T., Karsten, G., Kiessling, C., Möller, G. E., Wiesbeck, A., & Prenzel, M. (2015). Fostering professional communication skills of future physicians and teachers: effects of e-learning with video cases and role-play. *Instructional Science*, 43(4), 443–462. <https://doi.org/10.1007/s11251-014-9341-6>

- Gartmeier, M., Bauer, J., Fischer, M. R., Karsten, G., & Prenzel, M. (2011). Modellierung und Assessment professioneller Gesprächsführungskompetenz von Lehrpersonen im Lehrer-Elterngespräch. In O. Zlatkin-Troitschanskaia (Ed.), *Stationen Empirischer Bildungsforschung* (pp. 412–424). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94025-0_29
- Gass, S. M., & Mackey, A. (2016). *Stimulated Recall Methodology in Applied Linguistics and L2 Research*. Routledge. <https://doi.org/10.4324/9781315813349>
- Gerich, M., & Schmitz, B. (2016). Using Simulated Parent-Teacher Talks to Assess and Improve Prospective Teachers' Counseling Competence. *Journal of Education and Learning*, 5(2), 285–301. <https://doi.org/10.5539/jel.v5n2p285>
- Gerlach, D. (2022). Das Wissen der Fremdsprachenlehrpersonen - und Grundsätzliches zu seiner Beforschung. *Zeitschrift Für Fremdsprachenforschung*, 33(1), 29–50.
- Gerlach, D., & Götz, S. (2021a). *Manual to the Creative Writing in Language Teaching (CWILT) Corpus*. <https://doi.org/10.17192/fdr/57>
- Gerlach, D., & Götz, S. (2021b). Narratives Schreiben im Englischunterricht: Eine korpuslinguistische und genreanalytische Betrachtung von Schreibprodukten der Sekundarstufe I. *Zeitschrift Für Fremdsprachenforschung*, 32(2), 203–228.
- Gerlach, S., Rau-Patschke, S., Skorsetz, N., & Weber, N. (2023). Studierende unterrichten als „Vertretungslehrkräfte“: Ein „Professionalisierungsportal“ für die Hochschullehre einführen und evaluieren. In L. Mientus, C. Klemppin, & A. Nowak (Eds.), *Potsdamer Beiträge zur Lehrerbildung und Bildungsforschung: Vol. 4. Reflexion in der Lehrkräftebildung: Empirisch - phasenübergreifend - interdisziplinär* (pp. 423–429). Universitätsverlag Potsdam.
- Girod, M., & Girod, G. (2006). Exploring the Efficacy of the Cook School District Simulation. *Journal of Teacher Education*, 57(5), 481–497. <https://doi.org/10.1177/0022487106293742>
- Gläser, J., & Laudel, G. (2010). *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen* (4th ed.). VS Verlag.
- Glisan, E., & Donato, R. (2017). *Enacting the Work of Language Instruction, Vol. 1. ACTFL*.
- Glisan, E., & Donato, R. (2021). *Enacting the Work of Language Instruction, Vol. 2. ACTFL*.
- Goh, C. C. M., & Burns, A. (2012). *Teaching speaking: A holistic approach. Cambridge language education*. Cambridge University Press.

- Goh, C. C. M., & Vandergrift, L. (2021). *Teaching and Learning Second Language Listening*. Routledge. <https://doi.org/10.4324/9780429287749>
- Gormley, G., Sterling, M., Menary, A., & McKeown, G. (2012). Keeping it real! Enhancing realism in standardised patient OSCE stations. *The Clinical Teacher*, 9(6), 382–386. <https://doi.org/10.1111/j.1743-498X.2012.00626.x>
- Graham, S., & Harris, K. R. (2018). Evidence-Based Writing Practices: A Meta-Analysis of Existing Meta-Analyses. In R. Fidalgo Redondo, K. R. Harris, & M. Braaksma (Eds.), *Design Principles for Teaching Effective Writing* (pp. 13–37). BRILL. https://doi.org/10.1163/9789004270480_003
- Graham, S., Hebert, M., & Harris, K. R. (2015). Formative Assessment and Writing. *The Elementary School Journal*, 115(4), 523–547. <https://doi.org/10.1086/681947>
- Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99(3), 445–476. <https://doi.org/10.1037/0022-0663.99.3.445>
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337–350. <https://doi.org/10.1007/s10654-016-0149-3>
- Gronostay, D., Manzel, S., & Zischke, F. E. (2023). Konstruktion eines Testinstruments zur Erhebung des Professionswissens von Lehramtsstudierenden im Fach Sozialwissenschaften (SoWis-L). *Diagnostica*, 69(4), 169–181. <https://doi.org/10.1026/0012-1924/a000313>
- Grossman, P. (Ed.). (2018). *Teaching core practices in teacher education*. Harvard Education Press.
- Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching Practice: A Cross-Professional Perspective. *Teachers College Record*, 111(9), 2055–2100.
- Grossman, P., & Fraefel, U. (Eds.). (2024). *Core practices in education. Core practices in teacher education: A global perspective*. Harvard Education Press.
- Grossman, P., Hammerness, K., & McDonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teachers and Teaching*, 15(2), 273–289. <https://doi.org/10.1080/13540600902875340>
- Grossman, P., Schneider Kavanagh, S., & Pupik Dean, C. G. (2018). The Turn Towards Practice in Teacher Education: An Introduction to the Work of the Core Practice Consortium. In P. Grossman (Ed.), *Teaching core practices in teacher education* (pp. 1–14). Harvard Education Press.
- Grotegut, L. (2023). *GEProS - Entwicklung und Evaluation einer digitalen Simulation zur Förderung diagnostischer Kompetenz von Lehramtsstudierenden* [Dissertation]. <https://doi.org/10.17619/UNIPB/1-1753>

- Grotegut, L., Janzen, T., Wotschel, P., & Vogelsang, C. (in preparation). Nur eine Note? – Erfahrungen von Lehramtsstudierenden mit Prüfungen und Feedback im Studium.
- Grotegut, L., & Klingsieck, K. B. (2023). Kompetenzentwicklung in der Lehrkräftebildung: der Mehrwert einer digitalen Simulation. *ZeHf – Zeitschrift Für Empirische Hochschulforschung*, 7(1-2023), 56–73. <https://doi.org/10.3224/zehf.v7i1.05>
- Grotjahn, R. (Ed.). (2002a). *Fremdsprachen in Lehre und Forschung (FLF): Vol. 32. Der C-Test: Theoretische Grundlagen und praktische Anwendungen*. AKS-Verl.
- Grotjahn, R. (2002b). Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis. In R. Grotjahn (Ed.), *Fremdsprachen in Lehre und Forschung (FLF): Vol. 32. Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (pp. 211–225). AKS-Verl.
- Grotjahn, R., & Kleppin, K. (2017). Feedback zu schriftlichen Lernerproduktionen. In B. Akukwe, R. Grotjahn, & S. Schipolowski (Eds.), *Narr Studienbücher. Schreibkompetenzen in der Fremdsprache: Aufgabengestaltung, kriterienorientierte Bewertung und Feedback* (pp. 255–291). Narr Francke Attempo.
- Grünewald, A., Ketzer-Nöltge, A., Mehlhorn, G., Rossa, H., Settinieri, J., & Wilden, E. (Eds.). (2022). Fachdidaktisches Wissen in der Fremdsprachenlehrkräftebildung [Special issue]. *Zeitschrift Für Fremdsprachenforschung*, 33(1).
- Grünewald, A., & Levin, A. (2022). *DiaSim: Simulation didaktischer und pädagogischer Diagnostik in der Lehrer*innenbildung*. <https://www.stifterverband.org/file/11270/download?token=pskgU3A3> [Last retrieved: January 02, 2025]
- Gulikers, J. T., Bastiaens, T. J., & Kirschner, P. A. (2004). A Five-dimensional Framework for Authentic Assessment. *Educational Technology Research and Development*, 52(3), 67–86.
- Gulikers, J. T., Bastiaens, T. J., Kirschner, P. A., & Kester, L. (2008). Authenticity is in the eye of the beholder: student and teacher perceptions of assessment authenticity. *Journal of Vocational Education & Training*, 60(4), 401–412. <https://doi.org/10.1080/13636820802591830>
- Gulikers, J. T., Kester, L., Kirschner, P. A., & Bastiaens, T. J. (2008). The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes. *Learning and Instruction*, 18(2), 172–186. <https://doi.org/10.1016/j.learninstruc.2007.02.012>
- Hallet, W. (2016). *Genres im fremdsprachlichen und bilingualen Unterricht: Formen und Muster der sprachlichen Interaktion* (1st ed.). Klett | Kallmeyer.

- Halliday, M. A. K. (2014). *Halliday's introduction to functional grammar* (Fourth Edition). Routledge Tylor & Francis Group.
<https://doi.org/10.4324/9780203431269>
- Hamstra, S. J., Brydges, R., Hatala, R., Zendejas, B., & Cook, D. A. (2014). Reconsidering fidelity in simulation-based training. *Academic Medicine: Journal of the Association of American Medical Colleges*, 89(3), 387–392.
<https://doi.org/10.1097/ACM.0000000000000130>
- Hanser, C., Nussbaumer, M., & Sieber, P. (1994). Was sich in geschriebenen Texten zeigt. In P. Sieber (Ed.), *Sprachlandschaft: Vol. 12. Sprachfähigkeiten - besser als ihr Ruf und nötiger denn je! Ergebnisse und Folgerungen aus einem Forschungsprojekt* (pp. 187–301). Sauerländer.
- Harder, P. (2014). *Werthaltungen und Ethos von Lehrern: Empirische Studie zu Annahmen über den guten Lehrer. Schriften aus der Fakultät Sozial- und Wirtschaftswissenschaften der Otto-Friedrich-Universität Bamberg: Vol. 12.* Univ. of Bamberg Press.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hattie, J. (2023). *Visible learning: The sequel; a synthesis of over 2,100 meta-analyses relating to achievement* (1st ed.). Routledge.
<https://doi.org/10.4324/9781003380542>
- Hattie, J., & Clarke, S. (2019). *Visible learning: Feedback* (1st ed.). Routledge.
- Hattie, J., Gan, M., & Brooks, C. (2017). Instruction Based on Feedback. In R. E. Mayer & P. A. Alexander (Eds.), *Educational Psychology Handbook. Handbook of research on learning and instruction* (2nd ed., pp. 290–324). Routledge.
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hattie, J., & Wollenschläger, M. (2014). A conceptualization of feedback. In H. Ditton & A. Müller (Eds.), *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (pp. 135–149). Waxmann.
- Hawe, E., Dixon, H., & Watson, E. (2008). Oral Feedback in the Context of Written Language. *The Australian Journal of Language and Literacy*, 31(1), 43–58.
- Hayes, J. R. (2012). Modeling and Remodeling Writing. *Written Communication*, 29(3), 369–388. <https://doi.org/10.1177/0741088312451260>
- Hayes, J. R., & Flowers, L. (1980). Identifying the Organization of Writing Processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Lawrence Erlbaum Ass.

- Heine, L., & Schramm, K. (2022). Introspektion. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik* (pp. 173–181). Gunter Narr Verlag.
- Helmke, A. (2022). *Unterrichtsqualität und Professionalisierung: Diagnostik von Lehr-Lern-Prozessen und evidenzbasierte Unterrichtsentwicklung* (1st ed.). Klett | Kallmeyer.
- Henley, A. J., & DiGennaro Reed, F. D. (2015). Should You Order the Feedback Sandwich? Efficacy of Feedback Sequence and Timing. *Journal of Organizational Behavior Management*, 35(3-4), 321–335.
<https://doi.org/10.1080/01608061.2015.1093057>
- Hennes, A.-K. (2018). *Schreibprodukte bewerten – Die Rolle der Expertise bei der Bewertung der Textproduktionskompetenz* [Dissertation]. Universität zu Köln, Köln.
- Hennes, A.-K., Schmidt, B. M., Yanagida, T., Osipov, Igor, Rietz, C., & Schabmann, A. (2022). Meeting the Challenge of Assessing (Students') Text Quality: Are There any Experts Teachers Can Learn from or Do We Face a More Fundamental Problem? *Psychological Test and Assessment Modeling*, 64(3), 272–303.
- Herrington, J., Reeves, T. C., & Oliver, R. (2014). Authentic Learning Environments. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 401–412). Springer New York. https://doi.org/10.1007/978-1-4614-3185-5_32
- Hess, M., Werker, K., & Lipowsky, F. (2017). Was Wissen Lehramtsstudierende über gutes Feedback? Zur Erfassung konzeptuellen Wissens und zu dessen Zusammenhang mit der Selbsteinschätzung der Studierenden. In K. Zierer & K.-H. Arnold (Eds.), *Jahrbuch für Allgemeine Didaktik. Themen- teil: Allgemeine Didaktik und Lehrer/innenbildung* (pp. 11–29). Schneider-Verlag Hohengehren.
- Hilger, A., Lübbert, T., Pretzer, I., Reinartz, J., Theißen, J., & Schneider, M. (2015). Seminar. In M. Schneider & M. Mustafić (Eds.), *Gute Hochschullehre: Eine evidenzbasierte Orientierungshilfe* (pp. 39–62). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-45062-8_3
- Holstein, A., Weber, K. E., Prilop, C. N., & Kleinknecht, M. (2022). Analyzing pre- and in-service teachers' feedback practice with microteaching videos. *Teaching and Teacher Education*, 117, 103817.
<https://doi.org/10.1016/j.tate.2022.103817>
- Holzberger, D., Philipp, A., & Kunter, M. (2013). How teachers' self-efficacy is related to instructional quality: A longitudinal analysis. *Journal of Educational Psychology*, 105(3), 774–786. <https://doi.org/10.1037/a0032198>

- Homer, M. (2022). Pass/fail decisions and standards: The impact of differential examiner stringency on OSCE outcomes. *Advances in Health Sciences Education: Theory and Practice*, 27(2), 457–473.
<https://doi.org/10.1007/s10459-022-10096-9>
- Hopp, H., Bail, J., & Jackson, C. N. (2020). Frequency at the syntax–discourse interface: A bidirectional study on fronting options in L1/L2 German and L1/L2 English. *Second Language Research*, 36(1), 65–96.
<https://doi.org/10.1177/0267658318802985>
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). *Cambridge language teaching library*. Cambridge University Press.
- Hümbert-Schnurr, S., Thomas, M., Wackermann, R., & Degeling, M. (2019). Implementierung lernwirksamen Feedbacks in Physikdidaktik-Seminaren. In C. Maurer (Ed.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe: Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Kiel 2018*. (pp. 850–853). Universität Regensburg.
- Hutz, M. (2017). Storing Words in the Mind: The Mental Lexicon and Vocabulary Learning. In M. Eisenmann & T. Summer (Eds.), *Anglistische Forschungen: Vol. 420. Basic issues in EFL teaching and learning* (3rd ed., pp. 105–117). Winter.
- Hutz, M. (2018). Focus on Form - The Lexico-Grammar Approach. In C. Surkamp & B. Viebrock (Eds.), *Teaching English as a foreign language: An introduction* (pp. 133–158). J.B. Metzler Verlag.
- Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing*, 16(3), 148–164.
<https://doi.org/10.1016/j.jslw.2007.07.005>
- Hyland, K. (2019). *Second language writing* (2nd ed.). Cambridge University Press.
- Ingenkamp, K. (Ed.). (1977). *Beltz-Studienbuch. Die Fragwürdigkeit der Zensurengebung: Texte und Untersuchungsberichte* (7th ed.). Beltz.
- Jackson, C. N., & Hopp, H. (2020). Prediction error and implicit learning in L1 and L2 syntactic priming. *International Journal of Bilingualism*, 24(5-6), 895–911. <https://doi.org/10.1177/1367006920902851>
- Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V. F., & Hughey, J. (1981). *Testing ESL composition: A practical approach* (1st ed.). *English composition program*. Newbury House.

- Jansing, B., Haudeck, H., Keßler, J.-U., Nold, G., & Stancel-Piątak, A. (2013). Professionelles Wissen im Studienverlauf: Lehramt Englisch. In S. Blömeke, A. Bremerich-Vos, G. Kaiser, G. Nold, H. Haudeck, J.-U. Keßler, & K. Schwippert (Eds.), *Professionelle Kompetenzen im Studienverlauf: Weitere Ergebnisse zur Deutsch-, Englisch- und Mathematiklehrausbildung aus TEDS-LT* (pp. 77–105). Waxmann.
- Janzen, T., Vogelsang, C., & Rumlich, D. (2024). Feedbackkompetenz handlungsnah prüfen: Die Entwicklung einer rollenspielbasierten Simulation als Prüfungsformat für angehende Englischlehrkräfte. In D. Gerlach (Ed.), *Wissen, Können und Handeln von Fremdsprachenlehrpersonen: Professions- und Wissensforschung in den Fremdsprachendidaktiken* (pp. 221–234). Peter Lang.
- Jewitt, C., & Kress, G. (2010). Multimodality, literacy and school English. In D. Wyse, R. Andrews, & J. Hoffman (Eds.), *The Routledge International Handbook of English, Language and Literacy Teaching* (pp. 342–352). Routledge.
- Jossberger, H., Breckwoldt, J., & Gruber, H. (2022). Promoting Expertise Through Simulation (PETS): A conceptual framework. *Learning and Instruction*, 82, Article 101686. <https://doi.org/10.1016/j.learninstruc.2022.101686>
- Jung, E. (2010). *Kompetenzerwerb: Grundlagen, Didaktik, Überprüfbarkeit*. Oldenbourg Verlag.
- Jürgens, E., & Lissmann, U. (2015). *Pädagogische Diagnostik: Grundlagen und Methoden der Leistungsbeurteilung in der Schule. Pädagogik: Vol. 27*. Beltz. http://content-select.com/index.php?id=bib_view&ean=9783407293992
- Kadel, J., Buschmann, C., Haas, S., Meßner, M. T., & Adl-Amini, K. (2023). Plan-spiele und simulative Methoden in der Lehrkräftebildung – ein Literatur-überblick. *Zeitschrift Für Hochschulentwicklung*, 18(Sonderheft Plan-spiele), 19–39. <https://doi.org/10.21240/zfhe/SH-PS/02>
- Kalantzis, M., Cope, B., Chan, E., & Dalley-Trim, L. (2016). *Literacies* (2nd ed.). Cambridge University Press.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535. <https://doi.org/10.1037/0033-2909.112.3.527>
- Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, 38(4), 319–342. <https://doi.org/10.1111/j.1745-3984.2001.tb01130.x>
- Kane, M. T. (2006). Validation. In *ACE / Praeger series on higher education. Educational measurement* (4th ed., pp. 17–64). Praeger Publ.

- Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3–17. <https://doi.org/10.1177/0265532211417210>
- Kane, M. T. (2013a). The Argument-Based Approach to Validation. *School Psychology Review*, 42(4), 448–457. <https://doi.org/10.1080/02796015.2013.12087465>
- Kane, M. T. (2013b). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kaufman, D., & Ireland, A. (2019). Simulation as a Strategy in Teacher Education. In E. Terhart (Ed.), *Oxford Research Encyclopedia of Education* (pp. 1–19). Oxford University Press. <https://doi.org/10.1093/acrefore/9780190264093.013.478>
- Keijzer, M. (2007). *Last in first out? An investigation of the regression hypothesis in Dutch emigrants in Anglophone Canada*. LOT: Vol. 163. LOT.
- Keller, S. (2013). *Integrative Schreibdidaktik Englisch für die Sekundarstufe: Theorie, Prozessgestaltung, Empirie*. Giessener Beiträge zur Fremdsprachendidaktik. Narr Francke Attempto.
- Kellogg, R. T. (1996). A Model of Working Memory in Writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and application* (pp. 57–71). Lawrence Erlbaum Associates Publ.
- Khan, K. Z., Ramachandran, S., Gaunt, K., & Pushkar, P. (2013). The Objective Structured Clinical Examination (OSCE): Amee Guide No. 81. Part I: An historical and theoretical perspective. *Medical Teacher*, 35(9), 1437–46. <https://doi.org/10.3109/0142159X.2013.818634>
- Khatti, N., & Sweet, D. (2013). Assessment Reform: Promises and Challenges. In M. B. Kane & R. Mitchell (Eds.), *Implementing Performance Assessment: Promises, Problems, and Challenges* (pp. 1–21). Taylor and Francis.
- Kim, J., Park, J.-H., & Shin, S. (2016). Effectiveness of simulation-based nursing education depending on fidelity: A meta-analysis. *BMC Medical Education*, 16(152), 1–8. <https://doi.org/10.1186/s12909-016-0672-7>
- Kirchhoff, P. (2016). Was sollte eine gute Englischlehrkraft wissen? Über die Auswahl von Items im FALKO-E Test zum fachspezifischen Professionswissen. In M. Legutke & M. Schart (Eds.), *Giessener Beiträge zur Fremdsprachendidaktik. Fremdsprachendidaktische Professionsforschung: Brennpunkt Lehrerbildung* (pp. 75–98). Narr Francke Attempto.

- Kirchhoff, P. (2017). FALKO-E: Fachspezifisches professionelles Wissen von Englischlehrkräften: Entwicklung und Validierung eines domänenspezifischen Testinstruments. In S. Krauss, A. Lindl, A. Schilcher, M. Fricke, A. Göhring, B. Hofmann, P. Kirchhoff, R. H. Mulder, & J. Baumert (Eds.), *FALKO: Fachspezifische Lehrerkompetenzen: Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik: mit neuen Daten aus der COACTIV-Studie* (pp. 113–152). Waxmann.
- Kitchener, R. F. (2002). Folk epistemology: An introduction. *New Ideas in Psychology*, 20(2-3), 89–105. [https://doi.org/10.1016/S0732-118X\(02\)00003-X](https://doi.org/10.1016/S0732-118X(02)00003-X)
- Klempin, C. (2019). *Reflexionskompetenz von Englischlehramtsstudierenden im Lehr-Lern-Labor-Seminar: Eine Interventionsstudie zur Förderung und Messung. Literatur-, Kultur- und Sprachvermittlung*. J.B. Metzler.
- Klieme, E. (Ed.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie*. Beltz.
- Klieme, E., & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin, & H.-H. Krüger (Eds.), *Kompetenzdiagnostik* (pp. 11–29). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90865-6_2
- Klieme, E., Hartig, J., & Rauch, D. (2008). The Concept of Competence in Educational Contexts. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Hogrefe eLibrary. Assessment of competencies in educational contexts* (pp. 3–22). Hogrefe.
- Kline, P. (2015). *A handbook of test construction: Introduction to psychometric design. Psychology Revivals Ser.* Routledge.
- Klippel, F. (2018). Englischlehrerbildung - historische und aktuelle Perspektiven der Vorbereitung auf den Berufseinstieg. In B. Diehr (Ed.), *Kolloquium Fremdsprachenunterricht: Vol. 62. Universitäre Englischlehrerbildung: Wege zu mehr Kohärenz im Studium und Korrespondenz mit der Praxis* (pp. 17–33). Peter Lang.
- Klippel, F. (2019). The History of English Instruction in the German-Speaking World. In R. Hickey (Ed.), *English in the German-Speaking World* (pp. 77–95). Cambridge University Press. <https://doi.org/10.1017/9781108768924.005>
- Klippel, F. (2022). Was sollen Sprachenlehrer wissen und können? Grundmuster der Fremdsprachenlehrkräftebildung im 19. Jahrhundert. *Zeitschrift Für Fremdsprachenforschung*(2), 73–95.

- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Klusmann, U. (2013). Occupational Self-Regulation. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers* (pp. 291–308). Springer US. https://doi.org/10.1007/978-1-4614-5149-5_14
- Knorr, P., & Schramm, K. (2012). Datenerhebung durch Lautes Denken und Lautes Erinnern in der fremdsprachendidaktischen Empirie. In S. Doff (Ed.), *Narr-Studienbücher. Fremdsprachenunterricht empirisch erforschen: Grundlagen - Methoden - Anwendung* (pp. 184–217). Narr.
- Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. *Zeitschrift Für Psychologie / Journal of Psychology*, 216(2), 61–73. <https://doi.org/10.1027/0044-3409.216.2.61>
- Koller, H.-C. (2014). *Grundbegriffe, Theorien und Methoden der Erziehungswissenschaft: Eine Einführung* (7th ed.). Kohlhammer.
- König, J. (2014). Kompetenzen in der Lehrerbildung aus fächerübergreifender Perspektive der Bildungswissenschaften. In A. Bresges, B. Dilger, T. Henemann, J. König, H. Lindner, & A. Rohde (Eds.), *Kompetenzen diskursiv: Terminologische, exemplarische und strukturelle Klärungen in der LehrerInnenbildung* (1st ed., pp. 17–46). Waxmann.
- König, J., Blömeke, S., & Doll, J. (2011). Pädagogisches Wissen von Deutsch-, Englisch- und Mathematiklehramtsstudierenden. In S. Blömeke, K. Schwippert, A. Bremerich-Vos, H. Haudeck, G. Kaiser, G. Nold, & H. Wiltenberg (Eds.), *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen: Erste Ergebnisse aus TEDS-LT* (pp. 135–157). Waxmann.
- König, J., Blömeke, S., & Schwippert, K. (2013). Pädagogisches Professionswissen im Studienverlauf. In S. Blömeke, A. Bremerich-Vos, G. Kaiser, G. Nold, H. Haudeck, J.-U. Keßler, & K. Schwippert (Eds.), *Professionelle Kompetenzen im Studienverlauf: Weitere Ergebnisse zur Deutsch-, Englisch- und Mathematiklehrausbildung aus TEDS-LT* (pp. 145–166). Waxmann.
- König, J., Doll, J., Buchholtz, N., Förster, S., Kaspar, K., Rühl, A.-M., Strauß, S., Bremerich-Vos, A., Fladung, I., & Kaiser, G. (2017). Pädagogisches Wissen versus fachdidaktisches Wissen? *Zeitschrift Für Erziehungswissenschaft*, 21(3), 1–38. <https://doi.org/10.1007/s11618-017-0765-z>

- König, J., Lammerding, S., Nold, G., Rohde, A., Strauß, S., & Tachtsoglou, S. (2016). Teachers' Professional Knowledge for Teaching English as a Foreign Language. *Journal of Teacher Education*, 67(4), 320–337. <https://doi.org/10.1177/0022487116644956>
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kopp, B., & Mandl, H. (2014a). Aspekte der Feedbacknachricht. In H. Ditton & A. Müller (Eds.), *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (pp. 151–162). Waxmann.
- Kopp, B., & Mandl, H. (2014b). Lerntheoretische Grundlagen von Rückmeldungen. In H. Ditton & A. Müller (Eds.), *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (pp. 29–41). Waxmann.
- Korthagen, F., & Nuijten, E. (2022). *The Power of Reflection in Teacher Education and Professional Development*. Routledge. <https://doi.org/10.4324/9781003221470>
- Korthagen, F. A. (2010). How teacher education can make a difference. *Journal of Education for Teaching*, 36(4), 407–423. <https://doi.org/10.1080/02607476.2010.513854>
- Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L., & van der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. *Medical Education*, 37(2), 132–139. <https://doi.org/10.1046/j.1365-2923.2003.01429.x>
- Krauss, S. (2020). Expertise-Paradigma in der Lehrerinnen- und Lehrerbildung. In C. Cramer, J. König, M. Rothland, & S. Blömeke (Eds.), *Handbuch Lehrerinnen- und Lehrerbildung* (pp. 154–162). utb.
- Krauss, S., Bruckmaier, G., Lindl, A., Hilbert, S., Binder, K., Steib, N., & Blum, W. (2020). Competence as a continuum in the COACTIV study: the "cascade model". *ZDM*, 52(2), 311–327. <https://doi.org/10.1007/s11858-020-01151-z>
- Krauss, S., Lindl, A., Schilcher, A., Fricke, M., Göhring, A., Hofmann, B., Kirchhoff, P., Mulder, R. H., & Baumert, J. (Eds.). (2017). *FALKO: Fachspezifische Lehrkompetenzen: Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik: mit neuen Daten aus der COACTIV-Studie*. Waxmann.

- Krauss, S., Lindl, A., Schilcher, A., & Tepner, O. (2017). Das Forschungsprojekt FALKO - ein einleitender Überblick. In S. Krauss, A. Lindl, A. Schilcher, M. Fricke, A. Göhring, B. Hofmann, P. Kirchhoff, R. H. Mulder, & J. Baumert (Eds.), *FALKO: Fachspezifische Lehrerkompetenzen: Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik: mit neuen Daten aus der COACTIV-Studie* (pp. 9–65). Waxmann.
- Krings, H. P. (1989). Schreiben in der Fremdsprache - Prozeßanalysen zum 'vierten skill'. In K. Baumgärtner (Ed.), *Konzepte der Sprach- und Literaturwissenschaft: Vol. 48. Textproduktion: Ein interdisziplinärer Forschungsüberblick* (pp. 377–436). De Gruyter.
- Kron, S., Sommerhoff, D., Achtner, M., Stürmer, K., Wecker, C., Siebeck, M., & Ufer, S. (2022). Cognitive and Motivational Person Characteristics as Predictors of Diagnostic Performance: Combined Effects on Pre-Service Teachers' Diagnostic Task Selection and Accuracy. *Journal Für Mathematik-Didaktik*, 43(1), 135–172. <https://doi.org/10.1007/s13138-022-00200-2>
- Kron, S., Sommerhoff, D., Achtner, M., & Ufer, S. (2021). Selecting Mathematical Tasks for Assessing Student's Understanding: Pre-Service Teachers' Sensitivity to and Adaptive Use of Diagnostic Task Potential in Simulated Diagnostic One-To-One Interviews. *Frontiers in Education*, 6, Article 604568. <https://doi.org/10.3389/educ.2021.604568>
- Kuckartz, U., & Rädiker, S. (2022). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung: Grundlagentexte Methoden* (5th ed.). *Grundlagentexte Methoden*. Beltz Juventa.
- Kuckartz, U., & Rädiker, S. (2023). *Qualitative content analysis: Methods, practice and software* (2nd ed.). Sage.
- Kulgemeyer, C., Kempin, M., Weißbach, A., Borowski, A., Buschhüter, D., Enkrott, P., Reinhold, P., Riese, J., Schecker, H., Schröder, J., & Vogelsang, C. (2021). Exploring the impact of pre-service science teachers' reflection skills on the development of professional knowledge during a field experience. *International Journal of Science Education*, 43(18), 3035–3057. <https://doi.org/10.1080/09500693.2021.2006820>
- Kulgemeyer, C., & Riese, J. (2018). From professional knowledge to professional performance: The impact of CK and PCK on teaching quality in explaining situations. *Journal of Research in Science Teaching*, 55(10), 1393–1418. <https://doi.org/10.1002/tea.21457>

- Kulgemeyer, C., Riese, J., Vogelsang, C., Buschhüter, D., Borowski, A., Weißbach, A., Jordans, M., Reinhold, P., & Schecker, H. (2023). How authenticity impacts validity: Developing a model of teacher education assessment and exploring the effects of the digitisation of assessment methods. *Zeitschrift Für Erziehungswissenschaft*, 26, 601–625. <https://doi.org/10.1007/s11618-023-01154-y>
- Kulgemeyer, C., & Tomczyszyn, E. (2015). Physik erklären – Messung der Erklärungsfähigkeit angehender Physiklehrkräfte in einer simulierten Unterrichtssituation. *Zeitschrift Für Didaktik Der Naturwissenschaften*, 21(1), 111–126. <https://doi.org/10.1007/s40573-015-0029-5>
- Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, 1(4), 279–308. <https://doi.org/10.1007/BF01320096>
- Kultusministerkonferenz. (2004). *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Mittleren Schulabschluss*.
- Kultusministerkonferenz. (2019a). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung*.
- Kultusministerkonferenz. (2019b). *Standards für die Lehrerbildung: Bildungswissenschaften*.
- Kultusministerkonferenz. (2023a). *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Ersten Schulabschluss und den Mittleren Schulabschluss*.
- Kultusministerkonferenz. (2023b). *Information in English*. <https://www.KMK.org/KMK/information-in-english.html> [Last retrieved: January 02, 2025]
- Kunter, M. (2013). Motivation as an Aspect of Professional Competence: Research Findings on Teacher Enthusiasm. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers* (pp. 273–289). Springer US. https://doi.org/10.1007/978-1-4614-5149-5_13
- Kunter, M., Kleickmann, T., Klusmann, U., & Richter, D. (2013). The Development of Teachers' Professional Competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers* (pp. 63–77). Springer US. https://doi.org/10.1007/978-1-4614-5149-5_4

- Kürzinger, A., & Pohlmann-Rother, S. (2019). Niedrig und mittel inferente Kodierung: Feedback im Schreibunterricht. In F. Lipowsky, M. Hess, & A.-K. Denn (Eds.), *Dokumentation der Erhebungsinstrumente des Projekts "Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE). Technischer Bericht zu den PERLE-Videostudien. Band 2: Beobachtungssysteme zur Beschreibung und Qualität von Grundschulunterricht* (pp. 387–403).
- Kuster, W., Egli Cuenat, M., Klee, P., & Roderer, T. (2014). *Profession-related language competence profile for foreign language teachers at lower secondary level*. https://www.phsg.ch/sites/default/files/cms/Forschung/Projekte/Berufsspezifische%20Sprachkompetenzprofil%20f%C3%BCr%20Lehrpersonen%20f%C3%BCr%20Fremdsprachen/KP_Sek.I_EN_21.4.15.pdf [Last retrieved: January 02, 2025]
- Landesprüfungsamt für Lehrämter an Schulen NRW. (2022). *Vergleichende Statistik: Gymnasien und Gesamtschulen*. https://fragdenstaat.de/anfrage/anfrage-zu-durchfallquoten-beim-zweiten-staatsexamen/694428/anhang/20220425_Anlage1_Ergb._ZweiteStaatsprfungen2021_LA_GY-GE.pdf [Last retrieved: January 02, 2025]
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Laudel, H., & Narciss, S. (2023). The effects of internal feedback and self-compassion on the perception of negative feedback and post-feedback learning behavior. *Studies in Educational Evaluation*, 77, 101237. <https://doi.org/10.1016/j.stueduc.2023.101237>
- Lazarides, R., Buchholz, J., & Rubach, C. (2018). Teacher enthusiasm and self-efficacy, student-perceived mastery goal orientation, and student motivation in mathematics classrooms. *Teaching and Teacher Education*, 69, 1–10. <https://doi.org/10.1016/j.tate.2017.08.017>
- Lederman, J. (2023). Validity and Racial Justice in Educational Assessment. *Applied Measurement in Education*, 36(3), 242–254. <https://doi.org/10.1080/08957347.2023.2214654>
- Lee, I. (2017). *Classroom Writing Assessment and Feedback in L2 School Contexts*. Springer Singapore. <https://doi.org/10.1007/978-981-10-3924-9>
- Lee, I. (2019). Teacher written corrective feedback: Less is more. *Language Teaching*, 52(4), 524–536. <https://doi.org/10.1017/S0261444819000247>
- Legutke, M. (2022). Theoretische Forschung. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik* (pp. 40–50). Gunter Narr Verlag.

- Legutke, M., Saunders, C., & Schart, M. (2022). Zwischen den Disziplinen: anmerkungen zur Fachspezifik des Professionswissens von Fremdsprachenlehrkräften. *Zeitschrift Für Fremdsprachenforschung*, 33(1), 3–28.
- Lehmann-Grube, S. K., & Nickolaus, R. (2009). Professionalität als kognitive Disposition. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus, & R. H. Mulder (Eds.), *Beltz-Bibliothek. Lehrprofessionalität: Bedingungen, Genese, Wirkungen und ihre Messung* (pp. 59–70). Beltz.
- Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, 44(2), 325–343. <https://doi.org/10.3758/s13428-011-0146-0>
- Leuders, T. (2014). Modellierungen mathematischer Kompetenzen – Kriterien für eine Validitätsprüfung aus fachdidaktischer Sicht. *Journal Für Mathematik-Didaktik*, 35(1), 7–48. <https://doi.org/10.1007/s13138-013-0060-3>
- Levelt, W. J. M. (1995). *Speaking: From intention to articulation* (4th ed.). A Bradford book. MIT Press.
- Levin, O., Frei-Landau, R., Flavian, H., & Miller, E. C. (2023). Creating authenticity in simulation-based learning scenarios in teacher education. *European Journal of Teacher Education*, 1–22. <https://doi.org/10.1080/02619768.2023.2175664>
- Levin, O., Frei-Landau, R., & Goldberg, C. (2023). Development and validation of a scale to measure the simulation-based learning outcomes in teacher education. *Frontiers in Education*, 8, Article 1116626. <https://doi.org/10.3389/feduc.2023.1116626>
- Levine, T. H., Mitoma, G. T., Anagnostopoulos, D. M., & Roselle, R. (2023). Exploring the Nature, Facilitators, and Challenges of Program Coherence in a Case of Teacher Education Program Redesign Using Core Practices. *Journal of Teacher Education*, 74(1), 69–84. <https://doi.org/10.1177/00224871221108645>
- Lewis, M. (2009). *The lexical approach: The state of ELT and a way forward* (6th ed.). Heinle.
- Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6th ed.). Beltz.
- Lightbown, P. M., & Spada, N. (2013). *How Languages are Learned*. Oxford Handbook for Language Teachers (4th ed.). Oxford University Press.
- Lipnevich, A. A., & Panadero, E. (2021). A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions. *Frontiers in Education*, 6, Article 720195. <https://doi.org/10.3389/feduc.2021.720195>
- Lipowsky, F. (Ed.). (2013). *Persönlichkeits- und Lernentwicklung an staatlichen und privaten Grundschulen: Ergebnisse der PERLE-Studie zu den ersten beiden Schuljahren*. Waxmann. <https://elibrary.utb.de/doi/book/10.31244/9783830978268>

- Lipowsky, F., Faust, G., & Kastens, Claudia, Post, Swantje. (2013). Die PERLE-Studie: Überblick und Hintergründe. In F. Lipowsky (Ed.), *Persönlichkeits- und Lernentwicklung an staatlichen und privaten Grundschulen: Ergebnisse der PERLE-Studie zu den ersten beiden Schuljahren* (pp. 9–28). Waxmann.
- Lipowsky, F., Hess, M., & Denn, A.-K. (Eds.). (2019). *Dokumentation der Erhebungsinstrumente des Projekts "Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE). Technischer Bericht zu den PERLE-Videostudien. Band 2: Beobachtungssysteme zur Beschreibung und Qualität von Grundschulunterricht*. https://www.pedocs.de/volltexte/2019/17081/pdf/MatBild_Bd23_4.pdf [Last retrieved: January 02, 2025]
- Lock, S., & and agencies (2021, October 18). Prince William reveals Earthshot Prize winners in global bid to tackle climate crisis: Five global thinkers were each awarded a grant worth £1m to develop their ideas and technologies. *The Guardian*. <https://www.theguardian.com/uk-news/2021/oct/18/prince-william-reveals-earthshot-prize-winners-in-global-bid-to-tackle-climate-crisis> [Last retrieved: January 02, 2025]
- Lotz, M. (2016). *Kognitive Aktivierung Im Leseunterricht der Grundschule: Eine Videostudie Zur Gestaltung und Qualität Von Leseübungen Im Ersten Schuljahr*. Springer Fachmedien Wiesbaden GmbH.
- Lotz, M., Gabriel, K., & Lipowsky, F. (2013). Niedrig und hoch inferente Verfahren der Unterrichtsbeobachtung. Analysen zu deren gegenseitiger Validierung. *Zeitschrift für Pädagogik*, 59, 357–380. <https://doi.org/10.25656/01:11942>
- Machts, N., Chernikova, O., Jansen, T., Weidenbusch, M., Fischer, F., & Möller, J. (2023). Categorization of simulated diagnostic situations and the salience of diagnostic information. *Zeitschrift Für Pädagogische Psychologie*, Article 1010-0652/a000364. Advance online publication. <https://doi.org/10.1024/1010-0652/a000364>
- Manchón, R. M. (2016). Language and L2 Writing: Learning to write and writing to learn in academic contexts. In K. Hyland & P. Shaw (Eds.), *Routledge handbooks in applied linguistics. The Routledge handbook of English for academic purposes* (1st ed., pp. 139–151). Routledge Taylor & Francis Group.
- Mandouit, L. (2020). *Investigating how students receive, interpret, and respond to teacher feedback* [Dissertation]. University of Melbourne, Melbourne.
- Mandouit, L., & Hattie, J. (2023). Revisiting "The Power of Feedback" from the perspective of the learner. *Learning and Instruction*, 84, 101718. <https://doi.org/10.1016/j.learninstruc.2022.101718>

- Mao, Z., & Lee, I. (2020). Feedback scope in written corrective feedback: Analysis of empirical research in L2 contexts. *Assessing Writing*, 45, 100469. <https://doi.org/10.1016/j.asw.2020.100469>
- Marczynski, B., Kaltefleiter, L. J., Siebeck, M., Wecker, C., Stürmer, K., & Ufer, S. (2022). Diagnosing 6th Graders' Understanding of Decimal Fractions: Fostering Mathematics Pre-Service Teachers' Diagnostic Competences with Simulated One-on-One Interviews. In F. Fischer & A. Opitz (Eds.), *Springer eBook Collection. Learning to Diagnose with Simulations: Examples from Teacher Education and Medical Education* (1st ed., pp. 49–62). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-89147-3_5
- Marks, J. (2023). *Standards und Kompetenzen in der Lehrer*innenbildung: Eine fremdsprachendidaktische Perspektive*. Dissertation. *Fremdsprachendidaktik in globaler Perspektive: Vol. 10*. LIT.
- Martel, J. (2021). Implementing “enhanced rehearsals” in a practice-based TESOL methods course. *TESOL Journal*, 12(2), Article e540. <https://doi.org/10.1002/tesj.540>
- Martin, J. R., & Rose, D. (2008). *Genre Relations: Mapping culture. Equinox textbooks and surveys in linguistics*. Equinox Pub.
- Marty, S. *Aufbau formativer Beurteilungskompetenz bei angehenden Lehrpersonen: Wirkungen eines Lehr-Lernangebots zu lernförderlichem Feedback* [Dissertation]. Universität Zürich. <https://doi.org/10.5167/uzh-238062>
- Matsumoto-Royo, K., & Ramírez-Montoya, M. S. (2021). Core practices in practice-based teacher education: A systematic literature review of its teaching and assessment process. *Studies in Educational Evaluation*, 70, 101047. <https://doi.org/10.1016/j.stueduc.2021.101047>
- Mattheoudakis, M. (2007). Tracking changes in pre-service EFL teacher beliefs in Greece: A longitudinal study. *Teaching and Teacher Education*, 23(8), 1272–1288. <https://doi.org/10.1016/j.tate.2006.06.001>
- Matz, F., Ritter, M., & Zeineddine, N. (n.d.). *Analyseschwerpunkt Sprechförderlicher Englischunterricht. Fach Englisch*. https://www.uni-muenster.de/imperia/md/content/ProVision/hinweisezurnutzung/provision_analyseschwerpunkt_sprechfoerderlicher_englischunterricht.pdf [Last retrieved: January 02, 2025]
- Mayr, J. (2014). Der Persönlichkeitsansatz in der Forschung zum Lehrerberuf Konzepte, Befunde und Folgerungen. In E. Terhart, H. Bennewitz, & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (2nd ed., pp. 189–215). Waxmann.
- Mayring, P. (2022). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (13th ed.). Julius Beltz GmbH & Co. KG.

- McDonald, M., Kazemi, E., & Schneider Kavanagh, S. (2013). Core Practices and Pedagogies of Teacher Education. *Journal of Teacher Education*, 64(5), 378–386. <https://doi.org/10.1177/0022487113493807>
- McKenney, S., & Reeves, T. C. (2019). *Conducting educational design research* (2nd ed.). Routledge.
- Mehrens, W. A. (1997). The Consequences of Consequential Validity. *Educational Measurement: Issues and Practice*, 16(2), 16–18.
- Meier, J. (2022). „LauRa - In der Lehramtsausbildung Reflexionskompetenz analysieren“: Untersuchungen zum Verhältnis von Theorie, Praxis und Reflexion im Rahmen der Professionalisierung angehender Lehrpersonen [Dissertation]. Universität Paderborn. <https://doi.org/10.17619/UNIPB/1-1274>
- Meier, J., Janzen, T., Wotschel, P., & Vogelsang, C. (2023). Rollenspielbasierte Simulationen als Übungs- und Prüfungsformate im Lehramtsstudium. Eine explorative Studie zu Erfahrungen und Einschätzungen aus Studierendensicht. *Die Hochschullehre*, 9, 85–100. <https://doi.org/10.3278/HSL2307W>
- Meier, J., Vogelsang, C., & Rotärmel, T. (2024). "Das Berufsfeldpraktikum empfinde ich als nicht notwendig im Studium". Evaluationsergebnisse zum Berufsfeldpraktikum aus Sicht von Lehramtsstudierenden. In A. Schöning, M. Heer, M. Pahl, F. Diehr, E. Parusel, A. Tinnefeld, & J. Walke (Eds.), *Das Berufsfeldpraktikum als Professionalisierungselement. Grundlagen, Konzepte, Beispiele für das Lehramtsstudium* (pp. 310–324). Verlag Julius Klinkhardt. <https://doi.org/10.35468/6068-31>
- Meier, S. (2015). *Kompetenzen von Lehrkräften: Eine empirische Studie zur Entwicklung fachübergreifender Kompetenzeinschätzungen*. Dissertation. *Empirische Erziehungswissenschaft: Vol. 55*. Waxmann.
- Meinhardt, C. (2018). *Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern*. Dissertation. *Studien zum Physik- und Chemielernen: Vol. 256*. Logos Verlag.
- Messick, S. (1988). The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 33–45). Lawrence Erlbaum Associates, Inc.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *The American Council on Education Macmillan series in higher education. Educational measurement* (3rd ed., pp. 13–103). American Council on Education.
- Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist*, 50(9), 741–749.

- Messick, S. (1996). Validity of Performance Assessments. In G. W. Phillips (Ed.), *Technical Issues in Large-scale Performance Assessment* (pp. 1–18).
- Meuser, M., & Nagel, U. (2009). Experts and Changes in Knowledge Production. In A. Bogner, B. Littig, & W. Menz (Eds.), *Research methods series. Interviewing experts* (pp. 17–42). Palgrave Macmillan.
- Meyer, J., Köller, O., & Bernholt, A. (2023). Die Rolle von Persönlichkeitseigenschaften und Studienwahlmotiven für die Vorhersage von Leistungen im Lehramtsstudium und Vorbereitungsdienst. In T. Kauper, A. Bernholt, J. Möller, & O. Köller (Eds.), *PaLea: professionelle Kompetenzen und Studienstrukturen im Lehramtsstudium* (pp. 179–203). Waxmann.
- Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine: Journal of the Association of American Medical Colleges*, 65(9 Suppl), 63–67. <https://doi.org/10.1097/00001888-199009000-00045>
- Misoch, S. (2019). *Qualitative Interviews*. De Gruyter. <https://doi.org/10.1515/9783110545982>
- Molitor-Lübbert, S. (1996). Schreiben als mentaler und sprachlicher Prozeß. In H. Günther, O. Ludwig, J. Baurmann, F. Coulmas, K. Ehlich, P. Eisenberg, H. W. Giese, H. Glück, K. B. Günther, U. Knoop, B. Pompino-Marschall, E. Scheerer, & R. Weingarten (Eds.), *Schrift und Schriftlichkeit* (pp. 1005–1027). Walter de Gruyter. <https://doi.org/10.1515/9783110147445.2.7.1005>
- Moosbrugger, H., & Kelava, A. (Eds.). (2020). *Lehrbuch. Testtheorie und Fragebogenkonstruktion* (3rd ed.). Springer.
- Mory, E. H. (2004). Feedback Research Revisited. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (2nd ed., pp. 745–783). Lawrence Erlbaum Associates Publishers.
- Müller-Fohrbrodt, G., Cloetta, B., & Dann, H.-D. (1978). *Der Praxisschock bei jungen Lehrern: Formen, Ursachen, Folgerungen ; eine zusammenfassende Bewertung theoretischer und empirischer Erkenntnisse* (1st ed.). Lehrer-ausbildung und -weiterbildung. Klett.
- Müller-Hartmann, A., & Schocker-von Ditfurth, M. (2014). *Introduction to English language teaching* (4th ed.). *Uni-Wissen Anglistik/Amerikanistik*. Klett Lerntraining.
- Müller-Ruckwitt, A. (2008). "Kompetenz" - *Bildungstheoretische Untersuchungen zu einem aktuellen Begriff*. *Bibliotheca Academica Reihe Pädagogik: Vol. 6*. Ergon-Verl.

- Myers, P., Starr, A., & Mullins, K. (2018). Flight Simulator Fidelity, Training Transfer, and the Role of Instructors in Optimizing Learning. *International Journal of Aviation, Aeronautics, and Aerospace*. Advance online publication. <https://doi.org/10.15394/ijaaa.2018.1203>
- Narciss, S. (2006). *Informatives tutorielles Feedback: Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse. Pädagogische Psychologie und Entwicklungspsychologie: Vol. 56*. Waxmann Verlag GmbH.
- Narciss, S. (2008). Feedback Strategies for Interactive Learning Tasks. In J. M. Spector, M. D. Merrill, J. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology* (3rd ed., pp. 125–144). Routledge.
- Narciss, S. (2014). Modelle zu den Bedingungen und Wirkungen von Feedback in Lehr-Lernsituationen. In H. Ditton & A. Müller (Eds.), *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (43–82). Waxmann.
- Narciss, S. (2017). Conditions and Effects of Feedback Viewed Through the Lens of the Interactive Tutoring Feedback Model. In D. Carless, S. M. Bridges, C. Ka Yuk Chan, & R. Glofcheski (Eds.), *The Enabling Power of Assessment. Scaling up Assessment for Learning in Higher Education* (Vol. 5, pp. 173–189). Springer Singapore. https://doi.org/10.1007/978-981-10-3045-1_12
- Nelles, C., Langenbeck, S., Lubbe, D., Doering, B., Flach, E., Götz, C., Hinding, B., & Lindenmeyer, J. (2024). Shows How! Aufwand, Schwierigkeitsgrad und Testgütekriterien der anwendungsorientierten Parcoursprüfung nach der neuen Approbationsordnung für Psychotherapeut_innen. Ergebnisse einer Pilotstudie. *Psychologische Rundschau*, 75(1), 3–16. <https://doi.org/10.1026/0033-3042/a000632>
- Neuweg, G. H. (2014). Das Wissen der Wissensvermittler. In E. Terhart, H. Bennewitz, & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (2nd ed., pp. 583–614). Waxmann.
- Neuweg, G. H. (2022a). *Lehrerbildung: Zwölf Denkfiguren im Spannungsfeld von Wissen und Können*. Waxmann.
- Neuweg, G. H. (2022b). Lehrpersonenkompetenz im Spannungsfeld von Wissen und Können. In T. Hascher, T.-S. Idel, & W. Helsper (Eds.), *Handbuch Schulforschung* (pp. 1287–1307). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-24729-4_62
- Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. Sage.

- Nickl, M. (2023). *Developing Effective Scaffolding for Future Mathematics Teachers' Assessment Skills in Simulations: Towards Adaptivity* [Dissertation]. Technische Universität München.
- Nickl, M., Sommerhoff, D., Böheim, R., Ufer, S., & Seidel, T. (2024). Fostering pre-service teachers' assessment skills in a video simulation. *Zeitschrift Für Pädagogische Psychologie*, 38(1-2), 27–34.
<https://doi.org/10.1024/1010-0652/a000362>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.
<https://doi.org/10.1080/03075070600572090>
- Nückles, M., & Fraefel, U. (2024). *Core Practices. Projekte: Forschung, Entwicklung, Implementierung*. <https://www.corepractices.education/projekte/> [Last retrieved January 02, 2025]
- Nunberg, G. (1993). Indexicality and Deixis. *Linguistics and Philosophy*, 16(1), 1–43.
- Nussbaumer, M. (1991). *Was Texte sind und wie sie sein sollen*. De Gruyter.
<https://doi.org/10.1515/9783111632308>
- Nussbaumer, M., & Sieber, P. (1994). Texte analysieren mit dem Zürcher Textanalyseraster. In P. Sieber (Ed.), *Sprachlandschaft: Vol. 12. Sprachfähigkeiten - besser als ihr Ruf und nötiger denn je! Ergebnisse und Folgerungen aus einem Forschungsprojekt* (pp. 141–186). Sauerländer.
- Obliers, R., Koerfer, A., & Albus, C. (2018). Die OSCE-Methode mit Simulationspatienten. In A. Koerfer & C. Albus (Eds.), *Kommunikative Kompetenz in der Medizin: Ein Lehrbuch zur Theorie, Didaktik, Praxis und Evaluation der ärztlichen Gesprächsführung* (pp. 1867–1888). Verlag für Gesprächsforschung.
- OECD. (2023). *Pisa 2022 Results (Volume I): The State of Learning and Equity in Education. PISA*. OECD Publishing. https://www.oecd-ilibrary.org/education/pisa-2022-results-volume-i_53f23881-en [Last retrieved: January 02, 2025] <https://doi.org/10.1787/53f23881-en>
- Oser, F. (1996). Wann lernen Lehrer ihr Berufsethos? *Die Institutionalisierung von Lehren und Lernen. Beiträge zu einer Theorie der Schule, Zeitschrift für Pädagogik, Beiheft*. <https://doi.org/10.25656/01:9800>
- Oser, F., & Oelkers, J. (Eds.). (2001). *Nationales Forschungsprogramm 33 - Wirksamkeit unserer Bildungssysteme. Die Wirksamkeit der Lehrerbildungssysteme: Von der Allrounderbildung zur Ausbildung professioneller Standards*. Rüegger.

- Pajares, M. F. (1992). Teachers' Beliefs and Educational Research: Cleaning Up a Messy Construct. *Review of Educational Research*, 62(3), 307–332.
<https://doi.org/10.3102/00346543062003307>
- Paltridge, B. (2001). *Genre and the language learning classroom. Michigan teacher training*. University of Michigan Press.
<http://www.loc.gov/catdir/description/umich051/2001002843.html>
- Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22, 74–98. <https://doi.org/10.1016/j.edurev.2017.08.004>
- Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, 35, 100416.
<https://doi.org/10.1016/j.edurev.2021.100416>
- Pang, M. (2019). Developing Core Practices for EFL/ESL Teaching: A Framework for Methodology Course Design. *TESOL Quarterly*, 53(1), 258–273.
<https://doi.org/10.1002/tesq.487>
- Pant, H. A., Zlatkin-Troitschanskaia, O., Lautenbach, C., Toepper, M., & Molerov, D. (Eds.). (2016). *Modeling and Measuring Competencies in Higher Education: - Validation and Methodological Innovations (KoKoHs) Overview of the Research Projects*.
- Parkes, J., Abercrombie, S., & McCarty, T. (2013). Feedback sandwiches affect perceptions but not performance. *Advances in Health Sciences Education: Theory and Practice*, 18(3), 397–407. <https://doi.org/10.1007/s10459-012-9377-9>
- Parkinson, J. (2017). The student laboratory report genre: A genre analysis. *English for Specific Purposes*, 45, 1–13.
<https://doi.org/10.1016/j.esp.2016.08.001>
- Pasternack, P., Baumgarth, B., Burkhardt, A., Paschke, S., & Thielemann, N. (2017). *Drei Phasen: Die Debatte zur Qualitätsentwicklung in der Lehrer_innenbildung. GEW-Materialien aus Hochschule und Forschung: Vol. 124*. wbv.
- Pereira, D., Assunção Flores, M., & Barros, A. (2017). Perceptions of Portuguese undergraduate students about assessment: a study in five public universities. *Educational Studies*, 43(4), 442–463.
<https://doi.org/10.1080/03055698.2017.1293505>
- Pereira, D., Cadime, I., Brown, G., & Flores, M. A. (2022). How do undergraduates perceive the use of assessment? A study in higher education. *European Journal of Higher Education*, 12(1), 1–17.
<https://doi.org/10.1080/21568235.2020.1871393>

- Pereira, D., Cadime, I., Flores, M. A., Pinheiro, C., & Santos, P. (2022). Investigating the Effect of the Programme of Study on University Students' Perceptions about Assessment. *Educational Assessment*, 27(1), 71–92. <https://doi.org/10.1080/10627197.2022.2027753>
- Pereira, D., Niklasson, L., & Flores, M. A. (2017). Students' perceptions of assessment: a comparative analysis between Portugal and Sweden. *Higher Education*, 73(1), 153–173. <https://doi.org/10.1007/s10734-016-0005-0>
- Perleth, C., & Neumann, M. (2011). Vergleich von Präsenzlehrveranstaltungen mit inhaltsgleichen E-Learning-Angeboten im Lehramtsstudium. In *Gesellschaft für Informatik 2011* (p. 454).
- Philip, T. M., Souto-Manning, M., Anderson, L., Horn, I., J. Carter Andrews, D., Stillman, J., & Varghese, M. (2019). Making Justice Peripheral by Constructing Practice as "Core": How the Increasing Prominence of Core Practices Challenges Teacher Education. *Journal of Teacher Education*, 70(3), 251–264. <https://doi.org/10.1177/0022487118798324>
- Pilypaitytė, L. (2013). *Fremdsprachenlehrerausbildung zwischen Qualitätsansprüchen und Realität: Auseinandersetzung mit exemplarisch ausgewählten Forderungen an die deutsche Fremdsprachenlehrerausbildung. Mehrsprachigkeit und multiples Sprachenlernen / Multilingualism and Multiple Language Acquisition and Learning: Vol. 10*. Schneider Verl. Hohengehren.
- Polivka, L., Delcour, C., Dufresne, H., Bartoli, S., Bataille, P., Bekel, L., Bonigen, J., Deladrière, E., Dimarcq, S., Felix, A., Havas, C., Le Goff, H., Levy, M., Riback, E., Welfringer-Morin, A., Houdouin, V., Hadj-Rabia, S., Bodemer, C., Faye, A., & Melki, I. (2024). Breaking bad news: An active learning method for medical students. *BMC Medical Education*, 24(1), Article 994. <https://doi.org/10.1186/s12909-024-05821-4>
- Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2), 157–170. <https://doi.org/10.1007/s10798-011-9189-x>
- Popham, W. J. (1997). Consequential validity: Right Concern-Wrong Concept. *Educational Measurement: Issues and Practice*, 16(2), 9–13. <https://doi.org/10.1111/j.1745-3992.1997.tb00586.x>
- Porsch, R. (2010). Schreibkompetenzvermittlung im Englischunterricht - Eine bundesweite Umfrage unter Lehrkräften und Schülern der Sekundarstufe I. *Zeitschrift Für Fremdsprachenforschung*, 21(1), 53–77.
- Porsch, R. (2020). Fremdsprachliches Schreiben in der Schule lehren. *Fremdsprachen Lehren Und Lernen*, 49(1), 67–82.

- Porsch, R., & Gollub, P. (2018). Veränderungen von Angst zu unterrichten bei Lehramtsstudierenden nach einem schulpraktischen Aufenthalt Ergebnisse der InPraxis-Studie zum Praxissemester in Nordrhein-Westfalen. In N. Hericks (Ed.), *Hochschulen im Spannungsfeld der Bologna-Reform* (pp. 239–256). Springer Fachmedien Wiesbaden.
https://doi.org/10.1007/978-3-658-21290-2_14
- Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: all that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, 35(3), 277–289. <https://doi.org/10.1080/02602930903541007>
- Prilop, C. N., Schwedler-Diesener, A., Holstein, A., Bussmann, A., & Kleinknecht, M. (2024). Videobasierte Förderung der Core Practice Feedback-Geben-in-Schüler:innen-Arbeitsphasen – Evaluation eines Flipped-Classroom-Designs. In M. Besser, C. Herzog, P. Kuhl, & M. Ahlers (Eds.), *Digitales Lehren und Lernen im Fachunterricht: Aktuelle Entwicklungen, Gegenstände und Prozesse* (1st ed., pp. 184–204). Juventa Verlag.
- Prochazka, J., Ovcari, M., & Durinik, M. (2020). Sandwich feedback: The empirical evidence of its effectiveness. *Learning and Motivation*, 71, 101649. <https://doi.org/10.1016/j.lmot.2020.101649>
- Puig-Mayenco, E., Chaouch-Orozco, A., Liu, H [Hong], & Martín-Villena, F. (2023). The LexTALE as a measure of L2 global proficiency: A cautionary tale based on a partial replication of Lemhöfer and Broersma (2012). *Linguistic Approaches to Bilingualism*, 13(3), 299–314.
- Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, 28(1), 4–13. <https://doi.org/10.1002/bs.3830280103>
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21–33.
- Riegger, M., & Rotthoff, T. (2022). Simulieren statt Studieren in der Lehrerbildung? *Religionspädagogische Beiträge*, 45(1), 17–25.
<https://doi.org/10.20377/rpb-156>
- Rodríguez Moreno, J., Agreda Montoro, M., & Ortiz Colón, A. (2019). Changes in Teacher Training within the TPACK Model Framework: A Systematic Review. *Sustainability*, 11(7), 1870. <https://doi.org/10.3390/su11071870>
- Roeger, C. (2016). *Philosophieunterricht zwischen Kompetenzorientierung und philosophischer Bildung*. Verlag Barbara Budrich.
<https://doi.org/10.2307/j.ctvddzfrz>
- Römhild, R. (2023). *Global citizenship, ecomedia and English language education*. Palgrave studies in education and the environment. Palgrave Macmillan.

- Rosenberg, M. B., & Holler, I. (2016). *Gewaltfreie Kommunikation: Eine Sprache des Lebens*. Junfermannsche Verlagsbuchhandlung.
- Rossa, H., & Wilden, E. (2024). Was bedeutet es, eine Fremdsprache zu können? Bildungsstandards für die erste Fremdsprache als unterrichtsbezogener Referenzrahmen und Instrument politischer Steuerung. *Zeitschrift Für Fremdsprachenforschung*, 35(1), 5–22.
- Roters, B., König, J., Tachtsoglou, S., & Nold, G. (2013). Fachdidaktisches Wissen angehender Englischlehrkräfte. Theoretischer Rahmen und empirische Ergebnisse zur Struktur eines Testinstruments. *Lehrerbildung Auf Dem Prüfstand*, 6(2), 155–177.
- Roters, B., Lammerding, S., Tachtsoglou, S., Nold, G., & König, J. (2015). Professionelle Kompetenz von angehenden Englischlehrkräften (PKE). In J. König & B. Roters (Eds.), *Posterausstellung zur Eröffnungsfeier* (pp. 14–15). Universität zu Köln.
- Roters, B., Nold, G., Haudeck, H., Keßler, J.-U., & Stancel-Piątak, A. (2011). Professionelles Wissen von Studierenden des Lehramts Englisch. In S. Blömeke, K. Schwippert, A. Bremerich-Vos, H. Haudeck, G. Kaiser, G. Nold, & H. Willenberg (Eds.), *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen: Erste Ergebnisse aus TEDS-LT* (pp. 77–99). Waxmann.
- Rothland, M. (2020). Theorie-Praxis-Verhältnis in der Lehrerinnen- und Lehrerbildung. In C. Cramer, J. König, M. Rothland, & S. Blömeke (Eds.), *Handbuch Lehrerinnen- und Lehrerbildung* (pp. 133–140). utb.
<https://doi.org/10.35468/hblb2020-015>
- Ruelmann, M., Charalambous, C. Y., & Praetorius, A.-K. (2023). The representation of feedback literature in classroom observation frameworks: an exploratory study. *Educational Assessment, Evaluation and Accountability*, 35(1), 67–104. <https://doi.org/10.1007/s11092-022-09403-0>
- Rumlich, D. (2015). *Evaluating bilingual education in Germany*. Dissertation. *Mehrsprachigkeit in Schule und Unterricht: Vol. 15*.
<https://doi.org/10.3726/978-3-653-06460-5>
- Rütti-Joy, O. *Fostering and Assessing Pre-Service English Teachers' Oral Teacher Language Competence through an Assessment Rubric and Peer Feedback: An LSP Approach [Dissertation]*. Universität Freiburg. <https://sonar.ch/documents/319592/files/R%C3%BCtti-JoyO.pdf> [Last retrieved: January 02, 2025]
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, 25(1), 54–67. <https://doi.org/10.1006/ceps.1999.1020>

- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.
<https://doi.org/10.1007/BF00117714>
- Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, 30(2), 175–194. <https://doi.org/10.1080/0260293042000264262>
- Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, 23(4), 349–371.
[https://doi.org/10.1016/S0191-491X\(97\)86215-3](https://doi.org/10.1016/S0191-491X(97)86215-3)
- Sauvé, L., Renaud, L., Kaufman, D., & Marquis, J.-S. (2007). Distinguishing between games and simulations: A systematic review. *Educational Technology & Society*, 10(3), 247–256.
- Scherer, S [Sina], Graham, S., & Busse, V. (2024). How effective is feedback for L1, L2, and FL learners' writing? A meta-analysis. *Learning and Instruction*, 93, 101961. <https://doi.org/10.1016/j.learninstruc.2024.101961>
- Schluer, J. (2022). *Digital Feedback Methods*. Gunter Narr Verlag.
<https://doi.org/10.24053/9783823395324>
- Schluer, J., Rütli-Joy, O., & Unger, V. (2023). Feedback Literacy. In T. Philipp & T. Schmohl (Eds.), *Hochschulbildung: Lehre und Forschung. Handbook Transdisciplinary Learning* (Vol. 6, pp. 155–164). Transcript Verlag.
<https://doi.org/10.14361/9783839463475-017>
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 3–32). Cambridge University Press.
- Schmidt, R. C. (1997). Managing Delphi Surveys Using Nonparametric Statistical Techniques *Decision Sciences*, 28(3), 763–774.
<https://doi.org/10.1111/j.1540-5915.1997.tb01330.x>
- Schnabel, K. (2018). Simulation aus Fleisch und Blut: Schauspielpatienten. In M. St. Pierre & G. Breuer (Eds.), *Simulation in der Medizin: Grundlegende Konzepte - Klinische Anwendung* (2nd ed., pp. 125–130). Springer.
- Schneider, K. (2019). What Does Competence Mean? *Psychology*, 10(14), 1938–1958. <https://doi.org/10.4236/psych.2019.1014125>
- Schnell, A. K. (2018). *Schreibprozesse und Schreibentwicklung in der Fremdsprache*. Dissertation. *Kolloquium Fremdsprachenunterricht: Vol. 66*. Peter Lang.
- Schreiber, N., & Gut, C. (2022). Die Validierungspraxis bei hands-on Experimentiertests in der naturwissenschaftsdidaktischen Forschung. *Zeitschrift Für Didaktik Der Naturwissenschaften*, 28(1), 127–148.
<https://doi.org/10.1007/s40573-022-00145-6>

- Schröder, J., Riese, J., Vogelsang, C., Borowski, A., Buschhüter, D., Enkrott, P., Kempin, M., Kulgemeyer, C., Reinhold, P., & Schecker, H. (2020). Die Messung der Fähigkeit zur Unterrichtsplanung im Fach Physik mit Hilfe eines standardisierten Performanztests. *Zeitschrift Für Didaktik Der Naturwissenschaften*, 26(1), 103–122. <https://doi.org/10.1007/s40573-020-00115-w>
- Schulentwicklung NRW. (n.d.). *Classroom Management*. <https://www.schulentwicklung.nrw.de/materialdatenbank/material/download/3746#:~:text=Funktionale%20Einsprachigkeit%20meint%2C%20dass%20die,Probleme%20auf%20die%20Muttersprache%20zur%C3%BCckgreift> [Last retrieved: January 02, 2025]
- Schumacher, S. (2017). *Lehrerprofessionswissen im Kontext beschreibender Statistik: Entwicklung und Aufbau des Testinstruments BeSt Teacher mit ausgewählten Analysen. Bielefelder Schriften Zur Didaktik der Mathematik Ser: v.4*. Springer Fachmedien Wiesbaden. <https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4839077>
- Schustereder, S. (2020). Von ABB bis ZfsL - Ein Erfahrungsbericht zur schulpraktischen Lehrerausbildung im Rahmen des Vorbereitungsdienstes am Zentrum für schulpraktische Lehrerausbildung Bon. In R. A. Ißler & U. Küchler (Eds.), *Wissenschaft und Lehrerbildung: Vol. 5. Impulse zur Fremdsprachendidaktik: - Issues in foreign language education* (pp. 349–366). V&R unipress; Bonn University Press.
- Schutz, K. M., Grossman, P., & Shaughnessy, M. (2018). Approximations of Practice in Teacher Education. In P. Grossman (Ed.), *Teaching core practices in teacher education* (pp. 57–83). Harvard Education Press.
- Schütze, B., Rakoczy, K., Klieme, E., Besser, M., & Leiss, D. (2017). Training effects on teachers' feedback practice. The mediating function of feedback knowledge and the moderating role of self-efficacy. *ZDM - Mathematics Education*, 49(3), 475–489.
- Setyonugroho, W., Kennedy, K. M., & Kropmans, T. J. B. (2015). Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Education and Counseling*, 98(12), 1482–1491. <https://doi.org/10.1016/j.pec.2015.06.004>
- Shepard, L. A. (1993). Evaluating Test Validity. *Review of Research in Education*, 19(1), 405–450. <https://doi.org/10.3102/0091732x019001405>
- Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2), 4–14. <https://doi.org/10.3102/0013189X015002004>

- Shulman, L. S. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review*, 57(1), 1–21.
- Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Siekmann, L., & Busse, V. (2022). Process-oriented writing and formative feedback in EFL classes: A comparison of teachers' and learners' perceptions. In E. Wilden, L. Alfes, K. F. Cantone, S. Cikrikci, & D. Reimann (Eds.), *Standortbestimmungen in der Fremdsprachenforschung* (pp. 274–290). Schneider bei wbv.
- Siekmann, L., Müller, N., & Busse, V. (2022). Wie können Lehrkräfte Feedback zur Schreibförderung in größeren Lerngruppen einsetzen? In V. Busse, N. Müller, & L. Siekmann (Eds.), *Schreiben fachübergreifend fördern: Grundlagen und Anregungen für Schule, Unterricht und Lehrkräftebildung* (1st ed., pp. 268–287). Klett | Kallmeyer.
- Siekmann, L., Parr, J., van Ophuysen, S., & Busse, V. (2023). Text quality and changing perceptions of teacher feedback and affective-motivational variables: a study with secondary EFL students. *Frontiers in Education*, 8, Article 1171914. <https://doi.org/10.3389/feduc.2023.1171914>
- Siker, J. (2023). Simulation in Preparing Teachers to Work with Learners with Diverse Abilities. In M. L. Angelini & R. Muñiz (Eds.), *Springer Texts in Education. Simulation for Participatory Education* (pp. 337–345). Springer International Publishing. https://doi.org/10.1007/978-3-031-21011-2_14
- Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation*, 70, 101030. <https://doi.org/10.1016/j.stueduc.2021.101030>
- Sopka, S., Simon, M., & Beckers, S. K. (2018). „Assessment drives learning“: Konzepte zur Erfolgs- und Qualitätskontrolle. In M. St. Pierre & G. Breuer (Eds.), *Simulation in der Medizin: Grundlegende Konzepte - Klinische Anwendung* (2nd ed., pp. 97–109). Springer. https://doi.org/10.1007/978-3-662-54566-9_8
- St. Pierre, M. (2018). Blick zurück: Die Geschichte der Patientensimulation. In M. St. Pierre & G. Breuer (Eds.), *Simulation in der Medizin: Grundlegende Konzepte - Klinische Anwendung* (2nd ed., pp. 1–18). Springer.
- Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology*, 60, 100832. <https://doi.org/10.1016/j.newideapsych.2020.100832>

- Stadler, W. (2018). Assessment for, assessment of und assessment as learning. In B. Hinger & W. Stadler (Eds.), *Testen und Bewerten fremdsprachlicher Kompetenzen* (pp. 169–179). Narr Francke Attempto.
- Stadler, W., & Kremmel, B. (2018). Testprinzipien. In B. Hinger & W. Stadler (Eds.), *Testen und Bewerten fremdsprachlicher Kompetenzen* (pp. 39–55). Narr Francke Attempto.
- Stanat, P., Schipolowski, S., Schneider, R., Weirich, S., Henschel, S., & Sachse, K. A. (2023). *IQB-Bildungstrend 2022. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich*. Waxmann Verlag GmbH. <https://doi.org/10.31244/9783830997771>
- Ständige Wissenschaftliche Kommission. (2023). *Lehrkräftegewinnung und Lehrkräftebildung für einen hochwertigen Unterricht*. <https://www.swk-bildung.org/content/uploads/2024/02/SWK-2023-Gutachten-Lehrkraeftebildung.pdf> [Last retrieved: January 02, 2025]
- Stevens, J. A., & Kincaid, J. P. (2015). The Relationship between Presence and Performance in Virtual Simulation Training. *Open Journal of Modelling and Simulation*, 03(02), 41–48. <https://doi.org/10.4236/ojmsi.2015.32005>
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18
- Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: a review. *Assessment & Evaluation in Higher Education*, 30(4), 325–341. <https://doi.org/10.1080/02602930500099102>
- Stürmer, K., Fütterer, T., Kron, S., Sommerhoff, D., & Ufer, S. (2024). What makes a simulation-based learning environment for preservice teachers authentic? The role of individual learning characteristics and context-related features. *European Journal of Psychology of Education*, 3277–3299. <https://doi.org/10.1007/s10212-024-00837-2>
- Sturpe, D. A., Layson-Wolf, C., Lynn McPherson, M., & Haines, S. T. (2013). An objective structured teaching examination (OSTE) for faculty development. *Currents in Pharmacy Teaching and Learning*, 5(6), 616–619. <https://doi.org/10.1016/j.cptl.2013.07.015>
- Summer, T., & Plötz, E. (2024). Core Practices in der Fachdidaktik Englisch: Relevanz für Lehrkräfteprofessionalisierung und Praktika. In E. Plötz, T. Summer, D. Grötzbach, & B. Drechsel (Eds.), *Schulpraktische Studien an der Universität Bamberg: Impulse für die Lehrkräftebildung* (Vol. 14, pp. 157–185). University of Bamberg Press. <https://doi.org/10.20378/irb-97307>
- Surkamp, C. (Ed.). (2017). *Metzler Lexikon Fremdsprachendidaktik: Ansätze – Methoden – Grundbegriffe* (2nd ed.). J.B. Metzler Verlag.

- Sutton, P. (2012). Conceptualizing feedback literacy: knowing, being, and acting. *Innovations in Education and Teaching International*, 49(1), 31–40. <https://doi.org/10.1080/14703297.2012.647781>
- Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.
- Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Sweller, J. (2011). Cognitive Load Theory. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation: Vol. 55. Cognition in education* (pp. 38–78). Elsevier.
- Tachtsoglou, S., & König, J. (2018). Der Einfluss von Lerngelegenheiten in der Lehrerbildung auf das pädagogische Wissen angehender Englischlehrkräfte. *Journal of Educational Research Online*, 10(2), 3–33. <https://doi.org/10.25656/01:16131>
- Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. *Higher Education*, 76(3), 467–481. <https://doi.org/10.1007/s10734-017-0220-3>
- Tai, J., Bearman, M., Gravett, K., & Molloy, E. (2023). Exploring the notion of teacher feedback literacies through the theory of practice architectures. *Assessment & Evaluation in Higher Education*, 48(2), 201–213. <https://doi.org/10.1080/02602938.2021.1948967>
- TeachingWorks. (n.d.). *High-leverage practices*. <https://www.teachingworks.org/high-leverage-practices/> [Last retrieved: January 02, 2025]
- Telgmann, L., & Müller, K. (2023). Training & prompting pre-service teachers' noticing in a standardized classroom simulation – a mobile eye-tracking study. *Frontiers in Education*, 8, Article 1266800. <https://doi.org/10.3389/feduc.2023.1266800>
- Terhart, E. (2002). *Standards für die Lehrerbildung: Eine Expertise für die Kultusministerkonferenz*. <https://d-nb.info/1141683334/34> [Last retrieved: January 02, 2025]
- Terhart, E. (2007). Teacher Induction in Germany: Traditions and Perspectives. In M. Valenčič Zuljan (Ed.), *Professional inductions of teachers in Europe and elsewhere* (pp. 154–166). Faculty of Education University of Ljubljana.
- Terhart, E. (2012a). "Bildungswissenschaften". Verlegenheitslösung, Sammeldisziplin, Kampfbegriff? *Zeitschrift für Pädagogik*, 58(1), 22–39. <https://doi.org/10.25656/01:10493>

- Terhart, E. (2012b). Wie wirkt Lehrerbildung? Forschungsprobleme und Gestaltungsfragen. *Zeitschrift Für Bildungsforschung*, 2(1), 3–21.
<https://doi.org/10.1007/s35834-012-0027-3>
- Terhart, E. (2014). Standards für die Lehrerbildung: Bildungswissenschaften - nach zehn Jahren. *Die deutsche Schule*, 106, 300–323.
<https://doi.org/10.25656/01:25897>
- Terhart, E. (2021). Die Entwicklung der Schulforschung nach PISA. In T. Hascher, T.-S. Idel, & W. Helsper (Eds.), *Handbuch Schulforschung* (pp. 57–82). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-24734-8_3-1
- Theelen, H., van den Beemt, A., & den Brok, P. (2019). Classroom simulations in teacher education to support preservice teachers' interpersonal competence: A systematic literature review. *Computers & Education*, 129, 14–26.
<https://doi.org/10.1016/j.compedu.2018.10.015>
- Toulmin, S. E. (2003). *The uses of argument: Updated Edition* (Updated ed.). EBSCOhost eBook Collection. Cambridge University Press.
- Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse. *Zeitschrift Für Pädagogische Psychologie*, 21(1), 11–27.
- Truscott, J. (1996). The Case Against Grammar Correction in L2 Writing Classes. *Language Learning*, 46(2), 327–369. <https://doi.org/10.1111/j.1467-1770.1996.tb01238.x>
- Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing*, 16(4), 255–272.
<https://doi.org/10.1016/j.jslw.2007.06.003>
- Ulrich, I., Klingebiel, F., Bartels, A., Staab, R., Scherer, S [Sonja], & Gröschner, A. (2020). Wie wirkt das Praxissemester im Lehramtsstudium auf Studierende? Ein systematischer Review. In I. Ulrich & A. Gröschner (Eds.), *Edition ZfE. Praxissemester im Lehramtsstudium in Deutschland: Wirkungen auf Studierende* (Vol. 9, pp. 1–66). Springer Fachmedien Wiesbaden.
https://doi.org/10.1007/978-3-658-24209-1_1
- Universität Paderborn. (2023). *Zahlenspiegel: Studierende und Abschlüsse 2023*. https://www.uni-paderborn.de/fileadmin/zv/1-3/Statistiken/Studierendenspiegel_2023/3mm_Studierendenspiegel_2023.pdf [Last retrieved: January 02, 2025]
- van Beuningen, C. G., de Jong, N. H., & Kuiken, F. (2012). Evidence on the Effectiveness of Comprehensive Error Correction in Second Language Writing. *Language Learning*, 62(1), 1–41. <https://doi.org/10.1111/j.1467-9922.2011.00674.x>

- van den Bergh, L., Ros, A., & Beijaard, D. (2013). Teacher feedback during active learning: Current practices in primary schools. *The British Journal of Educational Psychology*, 83(Pt 2), 341–362. <https://doi.org/10.1111/j.2044-8279.2012.02073.x>
- Voerman, L., Meijer, P. C., Korthagen, F. A., & Simons, R. J. (2012). Types and frequencies of feedback interventions in classroom interaction in secondary education. *Teaching and Teacher Education*, 28(8), 1107–1115. <https://doi.org/10.1016/j.tate.2012.06.006>
- Vogelsang, C. (2014). *Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften: Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*. Dissertation. *Studien zum Physik- und Chemielernen: Vol. 174*. Logos Verlag.
- Voss, T. (2019). Lehrkraftwissen und dessen Erwerb: Was muss eine Lehrkraft wissen und wo lernt sie es? In N. McElvany, F. Schwabe, W. Bos, & H. G. Holtappels (Eds.), *IFS-Bildungsdialoge: Vol. 3. Lehrerbildung – Potentiale und Herausforderungen in den drei Phasen* (pp. 9–28). Waxmann.
- Vygotskij, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.
- Vygotskij, L. S. (1978). *Mind in Society: The Developments of Higher Psychological Processes*. Harvard University Press.
- Wagener, U., Reimer, M., Lüschen, I., Schlesier, J., & Moschner, B. (2019). „Krass lehramtsbezogen“ – Lehramtsstudierende wünschen sich mehr Kohärenz in ihrem Studium. *Herausforderung Lehrer_innenbildung*, 2(1), 210–226. <https://doi.org/10.4119/hlz-2488>
- Walker, J. M. T., & Dotger, B. H. (2012). Because Wisdom Can't Be Told. *Journal of Teacher Education*, 63(1), 62–75. <https://doi.org/10.1177/0022487111419300>
- Walz, K. (2023). *Testing the Usability of Role-Plays to Assess Performance-Based Competences in Higher Education under Consideration of FAIR Data Management* [Dissertation]. Justus-Liebig-Universität Gießen. <https://doi.org/10.22029/jlupub-17878>
- Walz, K., Schwabe, U., & Braun, E. (2024). The role of performance skills in German teacher education: empirical relations between general pedagogical knowledge, communication skills, and grades. *Frontiers in Education*, 9, Article 1327827. <https://doi.org/10.3389/feduc.2024.1327827>
- Wass, V., van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *Lancet (London, England)*, 357(9260), 945–949. [https://doi.org/10.1016/S0140-6736\(00\)04221-5](https://doi.org/10.1016/S0140-6736(00)04221-5)

- Watanabe, H. (2017). An Examination of Written Genres in English Language Textbooks in Japan. *The Journal of AsiaTEFL*, 14(1), 64–80.
<https://doi.org/10.18823/asiatefl.2017.14.1.5.64>
- Webb, N. M., & Shaughnessy, M. (2005). Generalizability Theory: Overview. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (pp. 717–719). Wiley.
- Weger, D. (2019). Professional Vision - State of the art. *Fremdsprachen Lehren Und Lernen*, 48(1), 14–31.
- Weigle, S. C. (2002). *Assessing writing* (1st ed.). *Cambridge language assessment series*. Cambridge Univ. Press.
- Weinert, F. E. (2001). Concept of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Hogrefe & Huber.
- Weinert, F. E. (2014). *Leistungsmessungen in Schulen* (3rd ed.). Beltz.
- Weir, C., & Khalifa, H. (2008). A cognitive processing approach towards defining reading comprehension. *Cambridge ESOL*(31), 2–10.
- White, P. R. R. (1998). *Telling Media Tales: the news story as rhetoric*. University of Sydney, Sydney.
- Whitehurst, J. (n.d.). *Screencast Feedback for Clear and Effective Revisions of High-stakes Process Assignments*. <https://cccc.ncte.org/cccc/owi-open-resource/screencast-feedback> [Last retrieved: January 02, 2025]
- Wideen, M., Mayer-Smith, J., & Moon, B. (1998). A Critical Analysis of the Research on Learning to Teach: Making the Case for an Ecological Perspective on Inquiry. *Review of Educational Research*, 68(2), 130–178.
<https://doi.org/10.3102/00346543068002130>
- Widlak, H. (1983). Praxisschock: Das Scheitern der Anwendung von Wissen. *Bildung Und Erziehung*, 36(jg), 213–226. <https://doi.org/10.7788/bue-1983-jg18>
- Wiesbeck, A. (2015). *An Evaluation of Simulated Conversations as an Assessment of Pre-Serie Teachers' Communication Competence in Parent-Teacher Conversations*.
- Wieser, C. (2016). Teaching and personal educational knowledge – conceptual considerations for research on knowledge transformation. *European Journal of Teacher Education*, 39(5), 588–601.
<https://doi.org/10.1080/02619768.2016.1253673>
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, 25(1), 68–81.
<https://doi.org/10.1006/ceps.1999.1015>

- Wigfield, A., & Eccles, J. S. (2020). 35 years of research on students' subjective task values and motivation: A look back and a look forward. In *Advances in Motivation Science* (Vol. 7, pp. 161–198). Elsevier.
<https://doi.org/10.1016/bs.adms.2019.05.002>
- Wilkinson, T. J., Frampton, C. M., Thompson-Fawcett, M., & Egan, T. (2003). Objectivity in Objective Structured Clinical Examinations: Checklists Are No Substitute for Examiner Commitment. *Academic Medicine: Journal of the Association of American Medical Colleges*, 78(2), 219–223.
- Wirtz, M. A., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Hogrefe Verlag für Psychologie.
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology*, 10, Article 3087. <https://doi.org/10.3389/fpsyg.2019.03087>
- Woolfolk, A. (2010). *Educational psychology* (11th ed.). Pearson Education.
- Wotschel, P., & Janzen, T. (in preparation). Rollenspielbasierte Simulationen als Prüfungsformat im Lehramtsstudium. In C. Vogelsang, L. Grotegut, J. Bruns, J. Riese, & S. Fechner (Eds.), *Handlungsorientierung in der Ausbildung von Lehrkräften und pädagogischen Fachkräften - Konzeptionen, Herausforderungen & Forschungsperspektiven*.
- Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, 105(2), 249–265.
<https://doi.org/10.1037/a0031311>
- Yan, X., & Fan, J. (2021). Reliability and dependability. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 477–494). Routledge. <https://doi.org/10.4324/9781003220756-37>
- Zamel, V. (1985). Responding to Student Writing. *TESOL Quarterly*, 19(1), 79–101. <https://doi.org/10.2307/3586773>
- Zeichner, K. (2010). Rethinking the Connections Between Campus Courses and Field Experiences in College- and University-Based Teacher Education. *Journal of Teacher Education*, 61(1-2), 89–99.
<https://doi.org/10.1177/0022487109347671>
- Zhai, L. (2019). Illuminating the Enactment of High-Leverage Teaching Practices in an Exemplary World Language Teaching Video Library. *American Educational Research Journal*, 56(5), 1681–1717.
<https://doi.org/10.3102/0002831218824289>
- Zierer, K. (2019). Erfolgreiches Feedback ist nicht einfach, aber es ist wirkmächtig! Klaus Zierer im Interview mit John Hattie. *Friedrich Jahresheft*, 6–9.

- Zierer, K., & Arnold, K.-H. (Eds.). (2017). *Jahrbuch für Allgemeine Didaktik. Thementeil: Allgemeine Didaktik und Lehrer/innenbildung*. Schneider-Verlag Hohengehren.
- Zimmermann, R. (2000). L2 writing: subprocesses, a model of formulating and empirical findings. *Learning and Instruction*, 10(1), 73–99.
[https://doi.org/10.1016/S0959-4752\(99\)00019-5](https://doi.org/10.1016/S0959-4752(99)00019-5)
- Zlatkin-Troitschanskaia, O., & Seidel, J. (2011). Kompetenz und ihre Erfassung – das neue „Theorie-Empirie- Problem“ der empirischen Bildungsforschung? In O. Zlatkin-Troitschanskaia (Ed.), *Stationen Empirischer Bildungsforschung* (pp. 218–233). VS Verlag für Sozialwissenschaften.
https://doi.org/10.1007/978-3-531-94025-0_16

List of Abbreviations

Abbreviation	Explanation
A1-C2	Levels of language competence in the CEFR (Council of Europe, 2001, 2020)
AERA	American Educational Research Association
AI	Artificial intelligence
α	Cronbach's α , indicator for consistency of items within a scale [Alpha] (Döring, 2023, p. 274)
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
AOI	Aptitude and Orientation Internship (<i>Eignungs- und Orientierungspraktikum</i>)
APA	American Psychological Association
AR	Augmented reality
β	Coefficient for regression [Beta]
BISTA(s)	<i>Bildungsstandard(s) (für die erste Fremdsprache Englisch/Französisch)</i> (Educational standards (for the first foreign language English/French)) (KMK, 2004, 2023a)
CA	Constructive Alignment
CEFR	Common European Framework of References (Council of Europe, 2001, 2020)
CI	Confidence interval
CK	Content knowledge
COACTIV	Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers (Baumert & Kunter, 2013)
COSIMA	Project " <i>Förderung von Diagnosekompetenzen in simulationsbasierten Lernumgebungen in der Hochschule</i> " (Fostering diagnostic competences in simulation-based learning environments in higher education)
d	Cohen's coefficient for standardized effect size (Döring, 2023, p. 799)
DESI	<i>Deutsch Englisch Schülerleistungen International</i> (International study on German and English learner achievement)
DFG	<i>Deutsche Forschungsgemeinschaft</i> (German Research Foundation)
DiaSim	Project " <i>Simulation didaktischer und pädagogischer Diagnostik in der Lehrer*innenbildung</i> " (Simulation of didactic and pedagogic diagnosis in teacher education) (Grünewald & Levin, 2022)
η^2	In an AN(C)OVA this value describes the variance of a factor [Eta-squared]
EF	Elaborated feedback
EFL	English as a foreign language
ES	Effect size
ESL	English as a second language
FALKO	Project " <i>Fachspezifische Lehrerkompetenzen</i> " (Subject-specific teacher competences) (Krauss, Lindl, Schilcher, Fricke, et al., 2017)
FALKO-E	Project " <i>Fachspezifische Lehrerkompetenzen Englisch</i> " (Subject-specific teacher competences English) (Kirchhoff, 2017)
FALKO-E PCK	The PCK part of the FALKO-E test
FI	Feedback intervention

List of Abbreviations

FL	Foreign language
GDPR	General Data Protection Regulation
GEProS	<i>Game- und e-learningbasierte, problemorientierte und selbst-gesteuerte Lernumgebung</i> (Game- and e-learningbased problemoriented and self-regulated learning environment) (Grotegut, 2023)
H1a-H4c	Hypotheses tested in this chapter
ICC	Intra-class correlation for interval-scaled rater agreement (Döring, 2023, p. 559)
ITF	Interactive tutoring feedback (Narciss, 2006)
IUA	Interpretation/Use Argument
IQB	<i>Institut zur Qualitätssicherung im Bildungswesen</i> (Institute for Educational Quality Improvement)
K	Coefficient for nominal rater agreement. Cohen's κ for two raters; Fleiss' κ for two or more raters [Kappa] (Döring, 2023, p. 559)
KCR	Knowledge of correct result
KMK	<i>Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [Kultusministerkonferenz]</i> (The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany)
KMK04/19	<i>Standards für die Lehrerbildung: Bildungswissenschaften</i> (Standards for teacher education: educational sciences) (KMK, 2019b)
KMK08/19	<i>Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung</i> (KMK subject-specific standards for teacher education) (KMK, 2019a)
KoKoHs	Project " <i>Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen</i> " (Modeling and Measuring Competencies in Higher Education – Validation and Methodological Innovations)
KomPrü	Project " <i>Kompetenzorientierte Prüfung kommunikativer Fähigkeiten von Studierenden</i> " (Competence-oriented assessment of students' communicative skills)
KR	Knowledge of result
L1	First language(s)
L2	Second language
LP	Language proficiency
M	Arithmetic mean
N	Number of subjects in a full sample
N	Number of subjects in a subsample
NRW	North Rhine-Westphalia
NCME	National Council on Measurement in Education
OI	Occupational Internship (<i>Berufsfeldpraktikum</i>)
OSCE	Objective Structured Clinical Examination
OSTE	Objective Structured Teaching Examination
P	Level of significance
PCK	Pedagogical content knowledge

PERLE	Project " <i>Persönlichkeits- und Lernentwicklung von Grundschulkindern</i> " (Personality- and learning development of primary school children) (Lipowsky, 2013)
PISA	Programme for International Student Assessment (OECD, 2023)
<i>PISA 2000</i>	Results of the first PISA study (Baumert, 2001)
PK	Pedagogical knowledge
PKE	Project " <i>Professionelle Kompetenz von Englischlehrkräften: Fachdidaktisches Wissen angehender Englischlehrkräfte – Konzeption, Messung, Validierung</i> " (EFL teachers' professional competence: Future EFL teachers' PCK – Conception, measurement, validation) (König et al., 2016).
ProfKom	Project " <i>Professionelle Kommunikation</i> " (Professionalization of future physicians and teachers in the area of communication competence) (Gartmeier et al., 2015)
PSET(s)	Pre-service EFL teacher(s)
PST(s)	Pre-service teacher(s)
QPT	Quick Placement Test
r	Pearson's correlation coefficient for parametric data (Bortz & Schuster, 2010, p. 153; Döring, 2023, p. 559)
ρ	Spearman's rank correlation coefficient for non-parametric data [Rho] (Döring, 2023, p. 667)
RobS	Role-play-based simulation(s)
RQT	Overarching research question for this thesis
RQ1-14	Research questions of this thesis
SBL	Simulation-based learning
SCR ^{Bio}	Project " <i>Simulated Classroom Biology</i> " (Julian Fischer et al., 2022)
SD	Standard deviation
SE	Standard error
SIS	School Internship Semester (<i>Praxissemester</i>)
SL	Standardized learner
SP	Standardized patient
SPK	Specific pedagogical knowledge
SRI	Stimulated recall interview
STEM	Science, Technology, Engineering and Mathematics (comparable to MINT)
τ -b	Kendall's rank correlation coefficient [Tau-b] (Döring, 2023, p. 667)
TEDS-LT	Teacher Education and Development Study – Learning to Teach (Blömeke et al., 2011)
TEFL	Teaching English as a Foreign Language
TLA	Teaching and Learning Activities (as part of constructive alignment)
TP	Test person
U	Test statistic for the Mann-Whitney U test
VA	Validity Argument
VR	Virtual reality
VT(s)	Verbal trigger(s)

List of Abbreviations

W	Kendall's coefficient of concordance for rank agreement among multiple raters (Wirtz & Caspar, 2002, p. 133)
W	Value of the test statistic in the Shapiro-Wilk test for normal distribution
WCF	Written corrective feedback
χ^2	Used in this thesis to determine the significance of agreement in Kendall's W [Chi-squared]
Z	Score of distribution for the discussion of normal distribution (A. Field, 2009, p. 139)
ZFF	<i>Zeitschrift für Fremdsprachenforschung</i> (Journal for foreign language education)
ZPD	Zone of proximal development (Vygotskij, 1978)
ZTA	<i>Zürcher Textanalyseraster</i> (Zurich text analysis instrument) (Nussbaumer, 1991)

List of Figures

Figure 1: Causal flow chart depicting the influence of teacher education on teacher actions and the consequential influence on student learning (Terhart, 2012b, p. 7; translated by TJ).....	8
Figure 2: Flow diagram of impact on student learning (Diez, 2010, p. 442).	8
Figure 3: The Cascade Model (Krauss et al., 2020, p. 316).	10
Figure 4: Model of professional competence used in the COACTIV study (Baumert & Kunter, 2013, p. 29).....	18
Figure 5: Facets of PK (Baumert & Kunter, 2013, p. 35).	20
Figure 6: Concepts of "teacher knowledge" (Neuweg, 2014, p. 585; translated by TJ in concurrence with Wieser, 2016, p. 595).....	24
Figure 7: Model of competence viewed as a continuum (Blömeke et al., 2015, p. 7).....	26
Figure 8: Constructive alignment in higher education (Biggs, 1996; figure adapted from Deibl et al., 2018, 295; Gallagher, 2017, p. 3013).....	41
Figure 9: Core Practices in the Foreign Language Classroom (Glisan & Donato, 2017, 2021 as cited in Summer & Plötz, 2024, p. 164.	46
Figure 10: Subject practices for the EFL classroom according to Pang (2019, p. 265).....	46
Figure 11: Framework for clinical assessment (Miller, 1990, p. 63; adapted and emphasis added by TJ).	58
Figure 12: Relation of assessment formats in teacher education, assessment levels (Miller, 1990) and competence facets (Blömeke et al., 2015).	67
Figure 13: Model of feedback, according to Hattie and Timperley (2007); figure adapted from Zierer (2019).....	73
Figure 14: The Hayes-Flower model (1980, as cited in Hayes, 2012, p. 370).....	80
Figure 15: Teaching-learning cycle for teaching genre writing (Hyland, 2019, p. 20).	87
Figure 16: Schematic procedure of the RobS (Adapted and translated from Wotschel & Janzen, in preparation).....	120
Figure 17: Still image of a simulation (AAS13).	126
Figure 18: Toulmin's model of inference (M. T. Kane, 2006, p. 28).	157
Figure 19: Framework for an argument-based approach to validation according to Schreiber and Gut (2022).	162
Figure 20: Overview of the IUA.	164
Figure 21: Research design of the project.	174
Figure 22: Constellation during the simulation.	191
Figure 23: Perspective of SRI video stimulus in which only the actor is visible.	196
Figure 24: Code overlap of relevant event-based items.	277
	451

Figure 25: Overview of the rater training.	282
Figure 26: Overview of the inter-rating procedure.	285
Figure 27: Screenshot of the performance by GJS22 (on the left) [0:05:25]. ..	293
Figure 28: Boxplot of the RobS test scores.	311
Figure 29: Distribution of RobS test scores.	312
Figure 30: Histogram of the test scores in the full sample.	313
Figure 31: Histogram of the inter-rating subsample rater 1 - rating I.	315
Figure 32: Histogram of the inter-rating subsample rater 2 - rating I.	316
Figure 33: Histogram of the inter-rating subsample rater 3 - rating I.	316
Figure 34: Scatter plot of rating I of rater 1 and raters 2 / 3.	317
Figure 35: Scatter plot of rating I of rater 2 and 3.	318
Figure 36: Histogram of the inter-rating subsample of rater 2 - rating II.	321
Figure 37: Histogram of the inter-rating subsample of rater 3- rating II.	321
Figure 38: Scatter plot of rating II of rater 1 and raters 2 / 3.	322
Figure 39: Scatter plot of rating II of rater 2 and 3.	322
Figure 40: RobS score difference with regard to phase of study.	349
Figure 41: Boxplot of the PCK scores.	352
Figure 42: Visual representation of the correlation results for H3.	362

List of Tables

Table 1: Framework for a performance-oriented assessment of feedback competence on writing	98
Table 2: Text features to diagnose and address in the feedback conversation.	122
Table 3: Verbal triggers used in the RobS.....	129
Table 4: Scoring instrument for the RobS.....	132
Table 5: Evidence organizer.....	176
Table 6: Overview of text samples and correct identification.....	181
Table 7: Data collection process in the pilot survey.	188
Table 8: Sample characteristics of the pilot survey.	194
Table 9: Overview VT "Good morning."	197
Table 10: Overview VT "Können Sie das auf Deutsch sagen?" / "Can you say that in German?"	198
Table 11: Overview VT "Ist vom Inhalt her alles drin?" ("Is everything included with regard to content?")	198
Table 12: Overview VT "Ist das vom Aufbau her in Ordnung?" ("Is the structure okay?").....	199
Table 13: Overview VT "Was habe ich hier falsch gemacht?" ("What did I do wrong here?")	199
Table 14: Overview of the VT "Ist das hier richtig?" ("Is this correct here?")...	200
Table 15: Overview of the VT "Das habe ich nicht verstanden." ("I did not understand that.")	200
Table 16: Overview of the VT "Wie kann ich das üben?" ("How can I practice that?")	201
Table 17: Overview of the VT "Muss ich den Text nochmal neu schreiben?" ("Do I have to write the text again?")	202
Table 18: Data collection process in the main survey.....	212
Table 19: Sample characteristics of the main survey.....	214
Table 20: Example category system for the authenticity of actors' behavior.	218
Table 21: Percentage agreement for the subcategories within the main categories of the guideline interview.....	219
Table 22: Code frequencies of participants' perception of authenticity of the RobS.....	220
Table 23: First level of subcategories of competences (general) required in the RobS mentioned by the participants.....	223
Table 24: Second level of subcategories of competences (general) required in the RobS mentioned by the participants.....	224

Table 25: First level of subcategories of EFL-specific competences required in the RobS mentioned by the participants.....	225
Table 26: Second level of subcategories of EFL-specific competences required in the RobS mentioned by the participants.....	225
Table 27: Overview of participants' perception of fairness in the RobS.....	226
Table 28: Overview of participants' perspective of the implementation of RobS for summative assessment.....	227
Table 29: Overview of experts.....	238
Table 30: Preliminary grades of the videos in the expert interview.....	239
Table 31: Questions for the evaluation of the scoring instrument.....	240
Table 32: Criteria for the assessment of feedback performance mentioned by the experts.....	246
Table 33: Item-specific means, standard deviations and distribution of points for each item.....	276
Table 34: Cross tabulation Deep: Process - Process.....	277
Table 35: Cross tabulation Self-regulation - Process.....	278
Table 36: Cross tabulation Surface: Language correctness - Surface: Language appropriateness.....	278
Table 37: Cross tabulation Deep: Structure - Deep: Content.....	278
Table 38: Cross tabulation Deep: Process - Deep: Content.....	279
Table 39: Cross tabulation Deep: Process - Deep: Structure.....	279
Table 40: Cross tabulation Dialogue orientation - Checking understanding.....	279
Table 41: Cross tabulation Dialogue orientation - Adaptive error correction.....	280
Table 42: Cross tabulation Checking understanding - Adaptive error correction.....	280
Table 43: Statistics for internal consistency of the 27-item scale.....	281
Table 44: Description of the inter-rating subsample.....	284
Table 45: Calculation of overall Fleiss' Kappa.....	287
Table 46: Fleiss' κ of rating I.....	287
Table 47: Pairwise comparison of inter-rater reliability scores of rating I.....	288
Table 48: Fleiss' κ of rating II.....	297
Table 49: Pairwise comparison of inter-rater reliability scores of rating II.....	298
Table 50: Selected discrepancies between rater 1 and rater 2 / 3 in rating II.....	300
Table 51: Descriptive statistics of RobS test scores.....	311
Table 52: Percentiles of the RobS test scores.....	311
Table 53: Descriptive statistics of participants in five-point intervals of the RobS test score.....	312
Table 54: Achieved RobS test score in %.....	312
Table 55: Statistics on normal distribution of RobS test score.....	313

Table 56: Descriptive statistics of the RobS test scores of the three raters in rating I.....	315
Table 57: Statistics on the normal distribution of rating I of all three raters..	317
Table 58: Pearson's <i>r</i> of the correlation of the RobS test scores between the three raters in rating I.....	318
Table 59: Spearman's ρ of the correlation of the RobS test scores between the three raters in rating I.....	318
Table 60: Corrected item-total correlations in rating I of raters 1, 2 and 3....	319
Table 61: Descriptive statistics of the test scores of all three raters in rating II.....	320
Table 62: Statistics on the normal distribution of rating II of all three rater in rating II.....	320
Table 63: Pearson's <i>r</i> of the correlation of the RobS test scores between the three raters in rating II.....	323
Table 64: Spearman's ρ of the correlation of the RobS test scores between the three raters in rating II.....	323
Table 65: Corrected item-total correlations in rating II for raters 1, 2 and 3..	323
Table 66: Descriptive statistics of the RobS test score of all three actors based on rater 1.....	328
Table 67: Spearman's ρ of the correlation of FALKO-E PCK scores between four raters.....	342
Table 68: Spearman's ρ of individual items in FALKO-E PCK between rater 1 and rater 3.	343
Table 69: Grading system for experts' global rating.....	345
Table 70: Participants self-reported data on curricular internships.	347
Table 71: Descriptive statistics of the external tests.....	351
Table 72: Pearson's <i>r</i> of the correlation between RobS, SPK, PCK and CK / LP.	354
Table 73: Spearman's ρ of the correlation between RobS, SPK, PCK and CK / LP.	355
Table 74: Pearson's <i>r</i> controlled for phase of study of the correlation between RobS, SPK, PCK and CK / LP.....	355
Table 75: Spearman's ρ controlled for phase of study of the correlation between RobS, SPK, PCK and CK / LP.....	355
Table 76: Pearson's <i>r</i> controlled for graduation grade of the correlation between RobS, SPK, PCK and CK / LP.....	355
Table 77: Spearman's ρ controlled for graduation grade of the correlation between RobS, SPK, PCK and CK / LP.	356
Table 78: Ratings of the performances based on grades.	357

Table 79: Spearman's rank-oriented correlations (ρ) between experts and the RobS score.....	358
Table 80: Kendall's τ -b rank-oriented correlations between experts and the RobS score.....	358
Table 81: Intercoding sub-codes for the PSETs guideline interview.....	685
Table 82: Discrepancies between rater 1 and rater 2 in rating I.....	686
Table 83: Discrepancies between rater 1 and rater 3 in rating I.....	687
Table 84: Inter-Item Correlation matrix of the 27 items.....	689
Table 85: RobS test score frequencies.....	692
Table 86: Statistics on the normal distribution of RobS test scores in the actor-specific subsamples.....	693
Table 87: Statistics on the normal distribution of the external tests of participants in their Bachelor program.....	693
Table 88: Statistics on the normal distribution of the external tests of participants in their Master of Education program.....	693
Table 89: Statistics on the normal distribution of the external tests.....	693
Table 90: Descriptive statistics of the three external tests of participants in the Bachelor program.....	694
Table 91: Descriptive statistics of the three external tests of participants in the Master of Education program.....	694
Table 92: Statistics on the normal distribution of the external tests of participants in their Bachelor (BA) and Master of Education (MEd) program.....	694
Table 93: Bivariate correlations for the external tests and graduation grade.....	694

Appendix

A. Test Material

A.1. M0: Vignette

Vignette

About the situation:

You have been working as an English teacher at the Heinz-Nixdorf-Gesamtschule for 4 months now. In your 9th grade you made an offer to provide your students with feedback in the writing process. One of your students has thus handed in her first draft of a newspaper report after the 6th lesson in your plan. She stated the following: "Können Sie bitte schauen, ob das so in Ordnung ist? Ich bin mir nicht sicher, ob das alles so passt."

Your task:

You have 15 Minutes to prepare for a **feedback conversation**, based on the learner text as well as the additional material. The conversation is to take place during the 10-minute-break after the lesson.

You can also add some written remarks for Mia to the text.

About the learner:

Mia is 15 years old and rather average in English, her grades are between 3 and 4, though the last class test was a 5+.

Material

- M1 Learner text
- M2 The task for the text
- M3 Excerpt from the series of lessons on the topic
- M4 Sample text from the series of lessons
- M5 Empty sheet for notes

We will collect all your notes and the material after the simulation.

Tuesday the Heinz-Nixdorf-Gesamtschule won the NRW Umweltpreis. It is for the action Plant trees for bees. In the gymnasium we all meet and environmental minister Schulze too was there. He says „Sie haben mit ihren vielen Taten gezeigt, dass die junge Generation ihre Zukunft selber in die Hand nimmt“ We become even 1000€ prize money for plant more trees! It gives different projects. We raise money with selling waffles, selling art and picking up garbage that the bank buys. Also the bank paid 5€ per collected kilo garbage: 207kg of garbage was collected and 1035€ donated. We too become money from parents and grandparents. Headmaster Meier speaks that it was great. It was an amazing event.

[illegible]

A.3. M2: The Task for the Text¹⁶⁴

M2 The task for the text

Writing a news report

Task: Yesterday, your school received the "NRW Umweltpreis" for its project "Plant trees for bees". As a junior journalist, it is your task to **write a short news report** on the ceremony and project for the local newspaper.

Keep in mind the **structure** and **style** of a news report.

Use the **notes** and **pictures** below.

- **Project:** Plant trees for bees
- **Aim:** Plant more trees to create living space for insects
 - 500 trees planted
 - Donations made by parents, grandparents, children
 - Special event "Friday for our bees" different projects:



Selling waffles



Selling art



Picking up garbage

- The bank paid 5€ per collected kilogram of garbage: 207kg of garbage was collected and 1,035€ donated

- **Ceremony:** Last Tuesday in the gymnasium
 - NRW Minister for the Environment Schulze gave a speech:
"Sie haben mit ihren vielen Taten gezeigt, dass die junge Generation ihre Zukunft selber in die Hand nimmt."
 - 1000€ prize money
 - Headmaster Meier: "Diese Leistung ist einfach spitze!"

¹⁶⁴ References for the images, from left to right: <https://pixabay.com/de/photos/waffel-waffeleisen-waffel-backen-203024/> (Pixabay_Efraimstochter); <https://pixabay.com/de/photos/junge-malen-kreativ-baby-6680403/> (Pixabay_Vika_Glitter); <https://pixabay.com/de/photos/%C3%B6kologie-schmutz-reinigung-m%C3%BCII-4202701/> (Pixabay_nastya_gepp) (December 14th, 2024).

A.4. M3: Excerpt from the Series of Lessons

M3 Excerpt from the series of lessons on the topic

Sitzung	Thema	Lerninhalt	Lernziel
1+2	Newspapers & The Environment	Die Lernenden lernen englischsprachige Zeitungen kennen, durchsuchen sie nach Artikeln zum Thema Umwelt und stellen diese kurz vor. Hausaufgabe: News report lesen	Die Lernenden können... ... Zeitungen als Gesamtstruktur verstehen. ... Artikel kurz präsentieren.
3	The news report	Die Lernenden beantworten Verständnisfragen zu dem Artikel und erstellen eine mind map mit Vokabeln zum Thema „environment“ aus dem gelesenen Artikel. Hausaufgabe: 2. news report lesen (M4), mind-map ergänzen	Die Lernenden können... ... inhaltliche Fragen zu einem news report beantworten. ... Wörter aus dem Themenfeld „environment“ vernetzen und benutzen.
4+5	The news report genre	Die Lernenden lernen Merkmale und sprachliche Mittel des Genres <i>news report</i> (Überschrift –Einführung– weitere Informationen) durch den Vergleich der beiden Artikel kennen.	Die Lernenden können... ... das genre <i>news report</i> beschreiben. ... sprachliche Mittel aus <i>news reports</i> erkennen.
6 (Aktuelle Stunde)	Writing a news report	Die Lernenden schreiben einen <i>news report</i> , basierend auf dem Arbeitsblatt	Die Lernenden können... ... einen ersten Entwurf eines <i>news report</i> schreiben.
7+8	Writing a news report II	Die Lernenden holen sich Feedback von ihren Mitschüler*innen in einer <i>writing conference</i> und überarbeiten ihre <i>news reports</i> .	Die Lernenden können... ... die <i>news reports</i> anderer Schüler*innen rückmelden ... Rückmeldungen nutzen um die eigenen <i>news reports</i> überarbeiten.
9	Project: Creating a class news paper	Die Lernenden beginnen mit einem Projekt in dem sie gemeinsam eine Klassenzeitung erstellen.	Die Lernenden können... ... gemeinsam eine Klassenzeitung erstellen.
10...	..	...	...

A.5. M4: Sample Text from the Series of Lessons

Due to reasons of copyright, the sample text cannot be reproduced here. The sample text is an abridged and adapted version, including vocabulary aid, of the following article:

Lock, S., & and agencies (2021, October 18). Prince William reveals Earthshot Prize winners in global bid to tackle climate crisis: Five global thinkers were each awarded a grant worth £1m to develop their ideas and technologies. *The Guardian*. <https://www.theguardian.com/uk-news/2021/oct/18/prince-william-reveals-earthshot-prize-winners-in-global-bid-to-tackle-climate-crisis> [Last retrieved: July 21st, 2025]

Please contact Thomas Janzen for further information.

A.6. M5: Empty Sheet for Notes

M5 Empty sheet for notes

Space for your notes

A.7. Role description of Mia (German)

Rollenbeschreibung Mia

Situationsbeschreibung:

Du bist Mia, 15 Jahre und gehst in die 9. Klasse einer Gesamtschule. Du bist in Englisch meistens durchschnittlich (zwischen 3 und 4 in Schulnoten) aber hast dir in der Vergangenheit auch nicht wirklich Mühe gegeben. Da die letzte Klassenarbeit mit einer 5+ eher schlecht für dich verlief, möchtest du dich bei der nächsten mehr anstrengen und dafür gut üben, daher suchst du das Feedback deiner Englischlehrkraft zum 1. Entwurf eines Zeitungsberichts. In der Stunde, in der ihr Merkmale eines Zeitungsberichts besprochen habt, hast du aber nicht alles verstanden, weil ihr es auf Englisch gemacht habt. Geschrieben hast du den Text wie immer, einfach drauf los und dann geschaut was so in der Aufgabe steht.

Da du dich auf Englisch unsicher fühlst, möchtest du das Gespräch gerne auf Deutsch führen. Du willst deiner Lehrkraft aber nicht direkt das Gefühl geben, dass du sie nicht verstehst, daher versuchst du im Gespräch deutlich zu machen, dass du das Gesagte nicht ganz verstehst, erst wenn du das Gefühl hast dein Lehrer reagiert nicht darauf, bittest du ihn direkt um den Sprachwechsel.

Du selbst schätzt den Text als okay ein. Du denkst, dass sprachlich ein paar Fehler drin sind aber der Text ist ein guter Bericht, die meisten Informationen aus der Aufgabe sind mit im Text.

Dein Sprachniveau entspricht ungefähr B1. Dein Hauptproblem sind Verben da du die unregelmäßigen Verben nie wirklich gelernt hast, bisher hat es auch so immer geklappt, wenn du das -ed anfügst. Generell schaffst du es bisher ganz gut dich mit einem Mix aus Deutsch und Englisch durchzuschlagen. Wenn man dir erklären würde, welche feststehenden Begriffe und unregelmäßige Verbformen du im Englischen gebrauchen müsstest, könntest du es aber verstehen.

Dein Ziel:

Als Mia möchtest du aus dem Feedbackgespräch folgendes unbedingt mitnehmen:

- Wie soll ein Zeitungsbericht aufgebaut sein?
- Was für sprachliche Fehler sind in deinem Text?
 - o Dir ist wichtig, dass die „Grammatik“ richtig ist.
 - o Du bist dir unsicher, ob „we“ angebracht ist
- Welche Punkte müssen noch in deinen Text mit rein?
 - o Du willst wissen, ob du was vergessen hast
- Was musst du tun um dich zu verbessern?

Mögliche Fragen & Aussagen:

- Can you say that again? Können Sie das nochmal sagen?
- I don't understand. Ich verstehe nicht.
- Können Sie nochmal erklären, was genau wir machen sollen?
- Können Sie das an einem Beispiel machen?
- Können Sie das aufschreiben?
- Und warum ist das falsch?
- Ich kann einfach kein Englisch.
- Was muss ich tun damit der Text gut ist?
- Ich kann das nicht. Warum bin ich schlechter als die anderen?

- Muss ich den Text nochmal schreiben?
- Habe ich so viel falsch gemacht in dem Text? Ich fand ihn jetzt nicht so schlecht wie Sie das sagen.
- Wie kann ich üben Texte zu schreiben?
- Ist das hier denn so okay?

Was zu beachten ist:

- Es soll ein Dialog entstehen – sollte die Versuchsperson zu Beginn in einen Monolog verfallen solltest du durch Signale dein Unverständnis zum Ausdruck bringen
- Versuch nicht deine Stimme, Sprache o.Ä. zu verstellen (kein absichtlich schlechtes Englisch, starke Akzente etc., lieber einfache Sätze)
- Versuche das Gespräch – sofern es auf Englisch beginnt – ins Deutsche zu überführen, dadurch dass du die Versuchsperson wissen lässt, dass du es auf Englisch nicht verstehst
- Wenn die Versuchsperson deine Frage gut beantwortet solltest du mindestens einmal versuchen, dass Sie dir noch 2-3 alternative Antworten gibt, also z.B. weitere Fragen nach Regeln, Erklärungen, Beispielen, dass Sie etwas aufschreibt, oder einen Lerntipp gibt
- Es sollten möglichst die unter den Zielen genannten Dinge angesprochen werden. So können wir Vergleichbarkeit unter den Versuchspersonen schaffen
- Nach 8-9 Minuten sollte ein Signal kommen „Ich hab nicht mehr so viel Zeit“
- Nach 11-12 Minuten sollte das Signal kommen „Können wir jetzt aufhören? Ich muss weiter“
- Wenn das Gespräch ins stocken kommt, stelle Nachfragen wie „ist das so okay?“ und zeig auf Stellen im Text die ihr noch nicht besprochen habt

A.8. Text Sample A from the Pilot Survey

A1 Text A

Plant trees for bees

Lastly Tuesday we got an prise for Envirement. The Prise is
for Plant trees for bees. For the prise we got 1000€. By Plant
trees for bees maked we trees for bees. We maked 500
plants. On Friday for our bees its gived a festival where we
selled waffeln and art and collected trash and than the bank
give there money for it. Sie haben mit ihren vielen Taten
gezeigt, dass die junge Generation ihre Zukunft selber in die
Hand nimmt. say minister for envirement Schulze and to
that say our headmaster that it was amazing. It was an
amazing event.

B. Rating manual

Rating Manual for a Role-play-based Simulation Assessment

Feedback on Writing in the EFL Context

PERFORM-LA

Version of July 24th, 2024

Thomas Janzen

Instructions for rating and documentation

1st viewing: Focus on events & dichotomous criteria

- Check all dichotomous criteria (p. 4-21)
- Check all tiered event-based scaled criteria (p. 22-27)
- Check conditions for conditional highly inferential criteria (p. 28-32)

For qualitative documentation:

When an event has occurred, pause the video and document the time stamp or copy the transcript's excerpt into the document.

2nd viewing: Focus on frequency

- Check highly inferential criteria (p. 28-47)
-

For qualitative documentation:

Add brief comments for highly-inferential criteria if two points have not been coded.

3rd viewing: Optional

- Review uncertain decisions.

When grading, the behavior of the learner/actor must not be considered.

On relative frequency-based criteria:

Relative frequency-based criteria operationalize the occurrences of the specified indicator relative to the duration of the dialogue:

0: "never" relates to 0 % - no occasion.

1: "sometimes" relates to up to about 50 %. One occasion (= not never!) is sufficient.

2: "often" relates to about 50 % and above. It does not have to be "always" (= 100 %), some lack of performance is acceptable.

Event-based categories

Self-evaluation

In accordance with the ITF model of feedback by Narciss (2006), there is an external feedback loop (teacher) and an internal feedback loop (learner). In order for the learner to make the best use of the feedback that is provided, it is important that they self-evaluate their performance. They can then compare the feedback they are receiving to their own evaluation and adjust accordingly. Furthermore, this allows the teacher to get information on the learner as well, which they can use in the conversation and by asking this question they engage the learner in the process more actively.

Category	Explanation	Specifications	Anchor	Points
Self-evaluation	TP asks the learner for their own evaluation of the performance (task, process or self-regulation), e.g., by asking the learner for specific challenges or specific aspects the learner succeeded in.	Self-evaluation must occur before specific feedback is given. Opening remarks that do not contain specific information are acceptable. <u>Key phrases:</u> "What do you think about your text?" "What were challenges during the (writing) process?" "What was easy/difficult for you?" "How was writing the text for you?" <u>Unsuccessful attempts:</u> "How was it?" "Do you like your text?"	Successful [0:00:26] TP: [...] what do you think about your TEXT (.) it-self (MMM07_Sim) – [0:00:35] TP: mh wie ist das DENN hattest du probleme beim schreibEN oder [0:00:40] S: mh (BJP10_Sim) – [0:01:09] TP: mhmh (4 Sek.) die struktur (.) okay gibt es denn etwas was du an deinem ähm bericht sehr gut fandest MIA [0:01:21] S: mh (..) weiß nicht ich habe ja hier auch irgendwo ein zitat drin (GBP25_Sim) Unsuccessful [0:00:49] TP: ähm und du hast auch einige wörter falsch geschrieben [0:00:54] S: okay [0:00:55] TP: ähm aber ansonsten war es an sich nicht falsch also du hast nichts falsches reingeschrieben [0:01:03] S: mhmh [0:01:04] TP: äh in den newspaper article (.) ähm (..) ja (.) ansonsten wie ist dir denn die aufgabe gefallen also fandest es war LEICHT oder nicht so LEICHT (CCF27_Sim)	2

Positive feedback

In accordance with the idea of self-efficacy, it is important to also include aspects in the feedback conversation that the learner has successfully accomplished (Bandura, 1997, pp. 80–81; Hattie & Timperley, 2007, pp. 96–97). It also is part of the diagnosis of the actual state, as this includes aspects that have been realized successfully. Positive feedback should, however, refer to specific aspects and not be in a general matter. Also, it should not deflect from the general message and purpose of feedback, which is to use the provided information to improve in the future. If positive feedback is used to mitigate the feedback content, it is deflecting.

Category	Explanation	Specifications	Anchor	Points
Positive feedback	TP mentions specific aspects that the learner has achieved successfully concerning the task, process or self-regulation (not self) at least once.	The aspect that is praised must be correct. Highlighting positive aspects that are actually incorrect are not considered, but when other positive aspects are mentioned two points can be coded. The positive feedback does not deflect from the overall feedback message (i.e., need for revision through using the feedback). It is sufficient to refer to broader aspects of the text here (e.g. 'content', 'structure'). It does not have to be a specific aspect as in the latter categories (deep/surface). Depending on the context, descriptive statements can be seen as positive feedback. <u>Key Phrases:</u> "You successfully..." "You did.... – that was good" "You included the relevant information." <u>Unsuccessful attempts:</u> "You did a good job!"	Successful [0:00:24] TP: [...] i would like to start off by saying that i really liked your structure (.) how you structured your text (SNA28_Sim) [0:01:10] TP: lassen wir (.) ähm okay also zum aufbau du hast ähm in der einleitung so die we fragen geklärt [0:01:16] S: [mhmh] [0:01:16] TP: [also] das fand ich sehr gut (.) dann weiß man am anfang erstmal worum es geht und das ist ja auch wichtig in so einem zeitungsartikel (MMK21_Sim) Unsuccessful Deflecting + general [00:00:24] TP: i have (.) looked at (.) your text [00:00:26] S: mhm [00:00:26] TP: and ähm i find it very good [00:00:31] S: mhm [00:00:32] TP: there are some aspects now where you can change things or write something differently (unv.) in all together it was good and now i will look at the points with you (unv.) (CHS10_Sim) Incorrect praise [0:07:33] TP: und it was an	2

Appendix B - Rating manual

Category	Explanation	Specifications	Anchor	Points
		"That was a good text"	amazing event alles richtig geschrieben kein problem (NDA16_Sim)	

Learning objectives (Feed Up)

This criterion is based on the model of feedback proposed by Hattie & Timperley (2007). The test person should include a context-relevant goal of the learning process. Learning objectives could also be set in dialogue with the learner. As it is self-evident that the ability to write a first draft of a news report following style and structure is the overall aim, aspects that fall within that main objective suffice as being specific learning objectives.

However, criteria that may be logical in the context of the task and its assessment/evaluation are not necessary context-relevant learning objectives, e.g., correct spelling or using the past simple because the event was in the past (because that logic also extends to other texts as well, like a diary entry or a blog post. Arguing that the news report is about a past event and therefore needs the past simple, however, fits the notion of learning objective here.

Category	Explanation	Specifications	Anchor	Points
Learning objectives (Feed Up)	TP clearly states at least one context-relevant and specific learning objective.	The mentioned aspect must be correct. It is sufficient to imply that criteria for the news report are learning objectives, including a reference to the sample text. It is not sufficient to say that because the event took place in the past, past tense should be used – it must be in reference to reporting/writing a news report. <u>Key phrases:</u> „The goal was to write a first draft...” “One aspect of a news report is to use past simple” “In a news report you have to...” <u>Unsuccessful attempts:</u> “You had to write a news report”	Successful [0:06:03] TP: wie gesagt das lernziel war es ja ähm dass ihr erstmal einen ersten entwurf äh ähm erfassen ne entwerfen konntet [0:06:11] S: mhm [0:06:12] TP: ähm und das lernziel hast du auf jeden fall erfüllt [0:06:15] S: okay (MMM07_Sim) Unsuccessful [0:01:22] TP: also du musst die gesamt äh die vergangenheitsform hier wählen [0:01:26] S: mhm [0:01:26] TP: das geht darum dass äh du äh hier jetzt hier geschrieben hast tuesday the heinz nixdorf (unv.) win würde dann ja bedeuten dass wäre heute passiert [aber ähm] [0:01:34] S: [ach so] [0:01:35] TP: da es auf tuesday bezogen ist und dass das gestern war [0:01:37] S: mhm [0:01:38] TP: musst du die vergangenheitsform benutzen [sprich] (NRS18_Sim)	2

Process

This criterion is based on the model of feedback proposed by Hattie & Timperley (2007). Feedback should provide information on the process of solving the task at hand. In order to distinguish this level from the level of self-regulation, it focuses on the process of writing a news report. When self-regulation and process overlap in a single statement, the question is on the focus (see specifications).

Category	Explanation	Specifications	Anchor	Points
Process	TP addresses aspects of the process of solving the task (feed up/back/forward) at least once.	Focus: Process of solving the task at hand (writing/news report) 1) Focus on writing 2) Focus on the news report The mentioned aspects must be correct. <u>Key phrases:</u> "You tried out new linguistic formulations " "When you write news report, you should try to include..." "When doing this task, you should include the aspects from the task sheet" "When structuring a text you should..." "We are in the process of writing and this includes feedback." (Feed-Up) <u>Unsuccessful attempts:</u> A reference to the scaffolding in the sample news report does not support solving the task at hand. <u>Clarification:</u> Difference to "Self-Regulation" The distinction of the two categories requires a view of the context and the whole conversation. Process looks at how TP can provide feedback on the process of solving the task at hand. The focus should thus be on the process of writing or on writing the news report. If a generalization for self-regulation is made based on the writing of the news report, only the process level should be considered.	<u>Successful</u> [0:08:06] TP: dann bist du direkt weitergegangen und hast so ein bisschen (.) ein bisschen so die handlung erzählt das da dann der ähm environmental minister schulz da gewesen ist und das zitat das kommt ja erst so wirklich weiter unten [0:08:16] S: ja [0:08:17] TP: da hast du so ein bisschen die informationen schon ein bisschen vermischt oder ein bisschen herumgespielt was ja auch ähm (.) zum beispiel wenn man an das kreative schreiben denkt ja gar nicht mal so äh also eine sehr gute strategie ist mit den informationen ein bisschen herumzuspielen (GBP25_Sim) — <u>Successful-Overlap to self-regulation, focus still in task at hand</u> [0:05:37] TP: also ja das du vielleicht nochmal (.) wenn du am ende als kleinen tipp wenn du am ende deinen	2

Category	Explanation	Specifications	Anchor	Points
		Self-regulation focuses on strategies that can apply to other tasks at well. When the task at hand is only used as an example for self-regulation strategies, only self-regulation is considered. <i>Difference to excluded category 'Deep-level: Process'</i> <i>This category looks at information the TP provides with regard to feed up/back/ forward on the process of solving the task (writing a news report), whereas "Deep-level: process" only considers <u>feedback</u> on the process the learner went through in order to write text.</i>	report durchliest [oder] so [0:05:42] S:[mhmh] [0:05:43] TP: das du da (.) ja in je- dem satz ein biss- chen auf die struk- tur deines satzes achtest (ACG22_Sim)	

Self-regulation

This criterion is based on the model of feedback proposed by Hattie & Timperley. Feedback should provide information on general aspects that are relevant when solving tasks like using the material for information, or seeking feedback as a means to improve learning.

Category	Explanation	Specifications	Anchor	Points
Self-regulation	TP addresses the use of specific strategies (feed up/back/forward) on a metacognitive level at least once.	Focus: More abstract than just the report. Specific tips, advice and strategies on writing, solving tasks (in general), learning, memorizing etc. Specificity refers to the fact that the strategy should directly address the question or topic. The mentioned aspects must be correct. <u>Key phrases:</u> „I think it is good that you came to me to get feedback“ “Next time, you can use the dictionary when you are unsure“ “You can also ask your peers for feedback“ “Check a text again before handing it in the next time.“ “You can use vocabulary from worksheets“ <u>Unsuccessful attempts:</u> General advice on language learning, e.g., through engagement with texts in the English language is not coded. <u>Difference to “Process”:</u> The distinction of the two categories requires a view of the context and the whole conversation. Process looks at how TP can provide feedback on the process of solving the task at hand. The focus should thus	Successful [0:06:54] TP: ähm das entwickelt sich auf jeden fall in die richtige richtung ich merke ja bei dir auch dass du jetzt feedback möchtest zu dem [text] [0:06:59] S: [ja] (räusp-ern) [0:07:00] TP: um dich selber zu verbessern ähm deswegen will ich dir da auf jeden fall schon dankbar und ich glaube dass das in eine gute richtung gehen kann [wenn du die sachen] (MMM07_Sim) Successful – Overlap with process but focus is on self-regulation [0:05:45] TP: [und] dann noch hier tress kleiner rechtschreibfehler wäre einmal mit doppel e [0:05:48] S: okay also so wie es hier auch drüber STEHT [0:05:51] TP: genau ja genau und hier in der aufgabe sieht man das auch immer [und das ist auch immer so ein kleiner] [0:05:54] S: [oh ja (lachen)] [0:05:55] TP: trick wenn wir so englische wörter in der aufgabe haben kannst du da dich immer dran orientieren [0:05:59] S: okay [0:05:59] TP: ähm genau wie hier unten hattest du bist du auch auf die parents und grandparents eingegangen	2

Category	Explanation	Specifications	Anchor	Points
		be on the news report. If a generalization for self-regulation is made based on the writing of the news report, only the process level should be considered. Self-regulation focuses on strategies that can apply to other tasks at well. When the task at hand is only used as an example for self-regulation strategies, only self-regulation is considered.	[0:06:04] S: mhm [0:06:05] TP: ähm parents hier hast du beides mit a geschrieben und in der aufgabe sieht man dass es mit e geschrieben wird [0:06:10] S: mh [0:06:10] TP: NE [da kannst du] [0:06:11] S: [okay] [0:06:11] TP: dich immer drauf zurück verweisen (BJP10_Sim) Unsuccessful - contradictory [0:08:01] auf jeden fall nochmal machen du kannst sowieso immer zu mir kommen und mir irgendwas was du (unv.) ja wenn du nochmal feedback oder so was brauchst weil besser du fragst nochmal nach anstatt das du dir unsicher bist oder so ich meine du kannst natürlich auch irgendwen aus der klasse fragen aber die können dir dann vielleicht auch nicht so weiterhelfen dann [0:08:21] S: ja okay gut dann muss ich jetzt auch schon weiter zu der nächsten stunde (AHW03_Sim) Unsuccessful – incorrect [0:03:01] TP: [wenn also also hier hier] habe ich jetzt zum beispiel so hilfestellung gegeben [0:03:07] S: ah okay [ja] [0:03:08] TP: [und] genau (.) so hilfestellung auf die kannst du vielleicht auch mehr achten [0:03:13] S: mhmh [0:03:13] TP: das würde auf jeden fall allgemein für alle arten von texten helfen (MMK21_Sim)	

Self

This criterion is based on the model of feedback proposed by Hattie & Timperley (2007). Feedback should not be directed at the person but rather on specific aspects the can improve on. This includes generic praise or criticism that is unfounded and unreasoned.

Two points are coded if TP does not address the learner's self.

Category	Explanation	Specifications	Anchor	Points
Self	TP does not address the learner's personal level, e.g., no generic praise or criticism of the learner is given.	Two points are coded if TP does NOT make statements according to the key phrases or the anchor. <u>Key phrases:</u> "You're a good/bad student" "You're not good at English" Advice for tutoring is not coded as relating to the learner's self.	<u>Unsuccessful – addresses the self-level</u> [0:04:05] TP: eigentlich bist du ja eine stärker eine starke schülerin (MMK21_Sim) – [0:06:14] TP: das wurde wie gesagt so der anfang wurde gut umgesetzt nur noch ein bisschen vertiefen und ich meine jeder hat ja seine fächer äh die man gut kann ich zum beispiel kann mathe immer noch nicht deswegen also das man zum beispiel zeitformen falsch verwendet oder auch become habe ich auch ganz lange falsch gesagt das kann man aber auch ändern (ELH30_Sim)	2

Prioritization of deep-level features

As one aspect of the writing process it is essential to focus on deep-level features in first drafts such as is the case in this situation (Ferris, 2013). Deep-level features focus on the structure and content of the text. In order to operationalize this notion, the TP must prioritize deep-level features when they address aspects in the text that need to be improved during a potential revision. Prioritization is taken literally here, in that it must occur before feedback on surface-level features. Considering that positive feedback addresses aspects that are not essential to revise in a second draft, aspects that are/could be coded as positive feedback and deep-level feature are not considered for this category. However, it is acceptable to give any kind of positive feedback before providing feedback on aspects that need to be improved.

Feedback in this context applies to feed back, e.g., information on the actual state or discrepancy between actual state and target state.

Possible orders of events:

- A) 1. Positive feedback on deep/surface-level 2. Feedback on deep-level features that need improvement → Criterion is coded.
- B) 1. Feedback on deep-level features that need improvement 2. Anything else → Criterion is coded.
- C) 1. Feedback on surface-level features that need improvement 2. Feedback on deep-level features that need improvement or anything else → Criterion is not coded.
- D) 1. Positive feedback on deep-level 2. Feedback on surface-level features that need improvement → Criterion is not coded.

Category	Explanation	Specifications	Anchor	Points
Prioritization of deep-level features	TP first provides feedback on deep-level features that need improvement, e.g., content or structure.	TP can provide positive feedback (deep/surface-level) before. TP cannot provide feedback on surface-level features that must be improved before providing feedback on deep-level features. If TP provides	Successful [0:01:08] TP: [ähm] und zwar ist mir erstmal aufgefallen ihr hattet ja hier dieses ähm arbeitsblatt dazu [mit den] [0:01:12] S: [mhm] [0:01:13] TP: informationen drauf und ich finde es gut dass du wirklich viele informationen dazu ähm genutzt hast um in den text zu integrieren [0:01:19] S: hmh [0:01:19] TP: das ist ja ein news report also ist es auch wichtig dass viele informationen drinne sind [0:01:24] S: mhm [0:01:25] TP: das hast du sehr gut gemacht	2

Category	Explanation	Specifications	Anchor	Points
		surface-level feedback before providing feedback on deep-level features this criterion is not fulfilled. Providing only positive feedback on deep-level features is not sufficient. <u>Key phrases:</u> "First, let's look at the structure/content" "Could you tell me first, how you wrote the text? What was your process like?" <u>Unsuccessful attempts:</u> - Aspect is addressed but not prioritized "Here you made a spelling mistake. Also, you forgot a title for the report." - Aspect is addressed prioritized but only positively "Your structure was good, but you should look at the tense."	[0:01:26] S: okay [0:01:26] TP: ähm (...) genau ähm (.) hier ist ja auch ähm steht ja auch dass ihr die die struktur benutzen solltet [ähm würdest du] [0:01:36] S: [JA] [0:01:36] TP: sagen dass das hier ähm wie es hier aufgebaut ist die struktur ist die in einem news report gebraucht werden sollte (...) [0:01:45] S: mh JA [0:01:46] TP: mhm [0:01:46] S: eigentlich schon (lachen) [0:01:47] TP: ja ähm kannst du dich noch an deinen text erinnern und würdest sagen dass du das ähm so auch genutzt hast (...) [0:01:54] S: mh ja ich habe es auf jeden fall versucht [0:01:55] TP: mhm (.) ich ich finde auch dass äh die struktur dir eigentlich ganz ähm ganz in ordnung gelungen ist [0:02:01] S: okay [0:02:01] TP: ähm was am ende oder was vielleicht ein bisschen zu kurz gekommen ist wenn ich dir das mal so als tipp mitgeben darf [0:02:07] S: hmh [0:02:07] TP: ist dass ähm dieser aim dieses ziel [ähm] [0:02:11] S: [mhm] [0:02:12] TP: hättest du vielleicht noch ein bisschen mehr informationen mit in den text schreiben können [0:02:16] S: okay also inhaltlich [hier] [0:02:17] TP: [genau] [0:02:18] S: okay mhm [0:02:18] TP: vielleicht noch ähm weil ähm hier wird ja so richtig klar ähm (.) warum eure schule diesen preis äh gewonnen hat NE ihr habt ja fünfhundert bäume gepflanzt [du hast] (MMM07_Sim) <u>Unsuccessful – other prioritization</u> [0:00:41] TP: so let's start with your first sentence you äh wrote down tuesday the heinz nix dorf gesamtschule win the nrw umweltpreis as you already know it was ähm the event from yesterday so you have to choose the past so it should be äh won instead of win (NRS18_Sim) <u>Unsuccessful – mentioned but not prioritized</u>	

Category	Explanation	Specifications	Anchor	Points
			[0:02:31] TP: da stand ja auch was das ziel also der aim des projektes ist das könntest du vielleicht nochmal ein bisschen genauer reinpacken das ähm das halt für mehr (.) lebensort für die insekten ist (SLG21_Sim) <u>Unsuccessful – only positive prioritization</u> [0:00:21] TP: good morning (8 Sek.) so genau ähm ich habe jetzt deinen text nochmal durchgelesen mia [0:00:32] S: [mhmh] [0:00:32] TP: [und] zwar sind mir da ein paar fehler aufgefallen (.) ähm (.) allgemein kann ich erstmal nur sagen inhaltlich ist das schon mal gut was du geschrieben hast [0:00:42] S: [okay] [0:00:42] TP: [also] das hast du schon mal gut rübergebracht (.) nur ähm wie du das ähm ja mit der rechtschreibung da müsstest du nochmal ein paar sachen äh (.) durchschauen und zwar wenn du hier einmal schaust (.) ähm in dem satz mit äh tuesday the heinz nixdorf gesamtschule ähm won the nrw da fällt dir hier direkt auf da habe ich das win durchgestrichen ähm weil du die (.) ähm zeitform nicht richtig angepasst hast also ich habe won statt win und (AAS13 Sim)	

Error patterns

According to Ferris (2013), it is more efficient to focus on error patterns that the learners have made when giving feedback instead of chronologically addressing each mistake. In order to obtain points, one error pattern has to be identified in the text and addressed in the conversation. It must be clear to the learner that the error occurred multiple times within the text and feedback should be provided on the pattern, either by addressing multiple examples or by giving overall advice on that issue.

Category	Explanation	Specifications	Anchor	Points
Error patterns	TP explicitly addresses error patterns (e.g. tense, irregular verbs, L1- inferential errors like false friends or cognates) and provides feedback on them (KR, KCR or further elaboration) at least once.	TP groups errors that share the same pattern and addresses them under the umbrella of the pattern. Explicitness can include physically pointing at aspects in the text. Repeated errors or error types are not considered as error patterns. Possible aspects: - Present simple instead of past tense forms - Forming of irregular verbs - L1-inferences, e.g., false friends <u>Key phrases:</u> "You still have an issue with using and forming the past simple"	Successful [0:01:28] TP: und zwar ähm ist mir einmal aufgefallen das du hauptsächlich im present tense geschrieben hast [0:01:33] S: mhmh [0:01:34] TP: und das man (.) aber in in news reports im past tense schreibt [weil] (SNA28_Sim) — [0:03:29] TP: überleiten und zwar ähm ist mir nämlich sehr oft aufgefallen dass du doch äh viele denglische formulierungen machst ähm wörter die man bei denen man denken könnte dass sie so richtig sind [ähm] [0:03:41] S: [okay] [0:03:42] TP: im englischen aber eher aus dem deutschen kommen ich möchte dir dafür das das wichtigste das größte den größten fehler denke ich mal geben und zwar it gived ähm (.) und zwar ist da besser there are ich weiß nicht ob du dich daran erinnerst darüber haben wir ähm im unterricht auch bestimmt schon gesprochen ähm aber da sind	2

Category	Explanation	Specifications	Anchor	Points
		"I can see you make some mistakes typical for German learners" <u>Unsuccessful attempts:</u> "You make a lot of spelling mistakes."	eben noch weitere also ähm zum beispiel to become ist auch ein ganz typischer fehler [ähm] (RNM25_Sim)	

Surface: Language correctness

This category targets TP's diagnostic abilities and is based on the framework by Becker-Mrotzek (2014).

Language correctness refers the accurate and precise identification of elements in the learner text that do not fulfill the linguistic norms in terms of orthography and grammar (syntax, morphology etc.). It includes the choice of words, if the chosen lexis is incorrect, e.g., a transfer error from the learner's first language. If the choice of words potentially fulfills the communicative purpose but is not appropriate for the genre, see language appropriateness.

Category	Explanation	Specifications	Anchor	Points
Surface: Language correctness	TP refers to aspects of language correctness, e.g., forming of past simple at least once.	Diagnosis must be correct and specific. Specificity can also be achieved through use of material. Possible aspects: - Forming of past simple - Syntax - Punctuation - Spelling - Incorrect use of words <u>Key phrases:</u> "This word is spelled incorrectly." "This is how you accurately form the past simple." <u>Unsuccessful attempts:</u> "There were many spelling mistakes."	<u>Successful</u> [0:02:55] TP: das ist nämlich so im laufe des textes öfter deutlich geworden (.) DANN ist es ganz wichtig ähm vokabeln sowas wie jetzt hier ähm (..) environMETAL ist erstmal environment NE minister wenn du jetzt einmal schaust (.) (CML16_Sim) [0:04:43] TP: äh (.) ja also es war du hast es nicht überall falsch ähm du das problem an einigen stellen ist zum beispiel du äh die irregular verbs nicht verwendet hast (GBP25_Sim) <u>Unsuccessful – unspecific</u> [0:00:49] TP: ähm und du hast auch einige wörter falsch geschrieben (CCF27_Sim) <u>Unsuccessful – incorrect</u> [00:02:09] TP: auf deutsch wäre es einfach okay (3.0) win the nrw umweltpreis (.) da ist ein grammatikfehler (.) wins the nrw umweltpreis (CHS10_Sim) [Context: Past tense forms should be used here.]	2

Surface: Language appropriateness

This category targets TP's diagnostic abilities and is based on the framework by Becker-Mrotzek (2014).

Language appropriateness refers to the accurate diagnosis of linguistic elements in the learner text that are not necessarily incorrect, but not appropriate for the text genre (e.g. news report). This refers to (lexico-)grammatical aspects when the element is potentially correct in other contexts, but not appropriate here. It also includes other linguistic and stylistic elements that are not suitable for the news report, e.g., lack of objectivity, the use of German without providing context/translations or variable use of language structures.

Category	Explanation	Specifications	Anchor	Points
Surface: Language appropriateness	TP refers to aspects of language appropriateness, e.g., inappropriate use of tenses at least once.	Diagnosis must be correct and specific. Specificity can also be achieved through use of material. Possible aspects: - Use of past simple instead of present simple because of the genre/report of a past event - Use of 'we' as unobjective for a news report - Using an indirect English quote instead of a direct German quote - Diverse sentence structure <u>Key phrases:</u> "This word/tense is not appropriate for this context" "This does not fit here" <u>Unsuccessful attempts</u> "You used the language inappropriately" "You forgot the past simple"	Successful [0:03:56] S: [okay] wo habe ich das denn falsch geMACHT [0:03:58] TP: ähm zum beispiel als äh beispiel einmal hier oben ähm du erzählst hier ja über die zeremonie [0:04:04] S: mhm [0:04:05] TP: und die war ja am diensttag das heißt das liegt in der vergangenheit [deswegen] [0:04:08] S: [ah okay] [0:04:08] TP: musst du hier nicht win nutzen sondern das unregelmäßige verb vielleicht weißt du es grad auch selber aus dem KOPF (MMM07_Sim) Unsuccessful - unspecific [0:00:42] TP: ähm und zwar hast du bei einigen sätzen das simple past vergessen (CCF27_Sim)	2

Deep: Structure

This category targets TP's diagnostic abilities and is based on the framework by Becker-Mrotzek (2014).

Text structure refers to TP's diagnosis of structural elements of the text. This includes overt structural elements like a headline for the news report or paragraphs, as well as the text-immanent elements like text structuring elements, cohesion and reader guidance like the provision of explanations/context for the reader. This also includes aspects of content if the positioning of that information is also relevant.

Category	Explanation	Specifications	Anchor	Points
Deep: Structure	TP refers to aspects of structure, e.g., cohesion, headline or audience-orientation at least once.	Diagnosis must be correct and specific. Specificity can also be achieved through use of material. Possible aspects: - Missing headline - Introduction, main part, conclusion - Additional explanations - Cohesion - Paragraphs - Missing information IF it is addressed in tandem with the structure/where it should be in the text	Successful [0:01:58] TP: [...] (.) du hast einen guten einleitungssatz geschrieben das ist ja auch wichtig bei einem zeitungsartikel (.) ähm was vielleicht noch wichtig wäre wäre ein title das die leute wissen worum es überhaupt geht (.) ähm (BNP09_Sim) – [0:06:51] TP: gerade am anfang nochmal mit reinnehmen äh warum das alles sattgefunden hat das ist bestimmt auch was was den leser interessiert (MJO28_Sim) Unsuccessful- too unspecific [0:07:27] S: okay war das denn auch so vom aufbau her wie das beim zeitungsartikel sein SOLL [0:07:31] TP: äh ja der aufbau hat auf jeden fall auch gestimmt (AAS13_Sim)	2

Deep: Content

This category targets TP's diagnostic abilities and is based on the framework by Becker-Mrotzek (2014).

Text content refers to TP's diagnosis whether the learner has fulfilled the task by including the information in the task sheet or other content-related aspects. This includes the facts as well as the quotes that are given in writing as well as the images in the task sheet, which were also included in the task description for the learner.

Category	Explanation	Specifications	Anchor	Points
Deep: Content	TP refers to aspects of content, e.g., missing information at least once.	Diagnosis must be correct and specific. Specificity can also be achieved through use of material. Possible aspects: - Images - Specific facts and figures - Information on the project's aim - Positive: Inclusion of quotes in general <u>Key phrases:</u> "You used a lot of information from the task sheet" "You forgot to include..." <u>Unsuccessful attempts:</u> "You included a lot of information."	Successful [0:02:31] TP: da stand ja auch was das ziel also der aim des projektes ist das könntest du vielleicht nochmal ein bisschen genauer reinputten das ähm das halt für mehr (.) lebensort für die insekten ist [0:02:42] S: ah [ja okay] [0:02:43] TP: das wir fünf hundert bäume gepflanzt haben [0:02:44] S: mhmh [0:02:45] TP: den rest hast du eigentlich eingebracht außer das du nicht friday for our bees (.) genannt hast das hättest du eigentlich auch noch reinbringen können aber du hast halt nur das ähm (..) (SLG21_Sim) Unsuccessful- too unspecific [0:07:24] TP: aber rein inhaltlich war das auf jeden fall schon gelungen (AAS13_Sim) Unsuccessful- incorrect [00:10:13] TP: äh (.) das müsste (unv.) also für einen zeitungsartikel ja also man darf ja (.) also es war auch nicht zu viel text oder inhalt (.) man soll ja auch keinen ewig langen text schreiben (.) der soll ja auch informieren (.) und da war soweit alles drin (CHS10 Sim)	2

Tiered event-based criteria

Adaptive error correction

This aim of this category is to evaluate TP's ability to give the learners an opportunity for self-correction by providing tiered, adaptive feedback. Self-correction must be related to text specific aspects which can be identified using the broader context as well. Simple comprehension questions are not considered as self-correction opportunities. Two points are coded if the TP presents an opportunity for self-correction and adapts with a second tier if the learner indicates inability to answer.

One point is coded if the TP presents an opportunity for self-correction but does not adapt with a second tier if the learner indicates inability to answer.

Zero points are coded if no opportunities for self-correction are given.

Category	Explanation	Specifications	Anchor	Points
Adaptive error correction	TP at least once provides opportunities for learner self-correction through graded feedback that begins with a question and includes at least one intermediate step before providing explicit feedback.	In order for the learner to self-correct an error has to be indicated previously. Questions/prompts that aim at the learner's knowledge but are not tied to a previously indicated error are not coded. <u>Key phrases:</u> "What do you think?" "Do you know..." "In which tense do you have to..." Question ↓ Implicit/indirect feedback ↓ Explicit feedback If the learner is unable to self-correct, appropriate scaffolding (implicit/indirect feedback) and, subsequently, explicit/direct feedback is provided.	Successful [0:03:11] TP: ähm kannst du dich vielleicht so dran erinnern was wir da vielleicht gelernt haben wenn wir nicht unbedingt wortwörtlich etwas wiedergeben möchten [0:03:18] S: mh (.) muss das dann auch auf englisch SEIN [0:03:21] TP: wäre schön wenn wir das auf englisch hinbekommen könnten [0:03:24] S: [ach so] [0:03:24] TP: [gibt es] da irgendwie eine strategie die wir kennengelernt haben im unterricht die wir da anwenden können [0:03:28] S: mh (.) ne weiß ich grade nicht [0:03:31] TP: äh total in ordnung also wenn wir doch äh von einer wörtlichen rede etwas wiedergeben [möchten] [0:03:36] S: [mhmh] [0:03:37] TP: ähm (.) dann müssen wir das ja indirekt irgendwie wiedergeben und [hier] [0:03:41] S: [mhmh] [0:03:41] TP: würde sich zum beispiel ganz gut anbieten mit indirect speech zum beispiel zu arbeiten (GBP25_Sim)	2

Category	Explanation	Specifications	Anchor	Points
			Example: "Look at this word, what do you think?" ↓ "Look at the tense, this is not quite right. Do you know why?" ↓ "Here you should use past simple, so 'won' would be correct. In a news report we talk about something that has happened, so we need the past simple."	
	TP at least once provides opportunities for learner self-correction but explicit feedback is provided subsequently without an intermediate step. Opportunity for self-correction can be a question or implicit/indirect feedback.	In order for the learner to self-correct an error has to be indicated previously. Questions/prompts that aim at the learner's knowledge but are not tied to a previously indicated error are not coded. <u>Key phrases:</u> "What do you think?" "Do you know..." 1) Question ↓ Explicit feedback 2) Implicit/indirect feedback ↓ Explicit feedback If the learner is unable to self-correct explicit/direct feedback is provided.	1) "Look at this word, what do you think?" ↓ "Here you should use past simple, so 'won' would be correct. In a news report we talk about something that has happened, so we need the past simple." [0:07:06] TP: [...] hast du vielleicht selber eine idee wie du die struktur besser machen könntEST oder eine idee wo du sagen würdest okay vielleicht das wir uns so ein bisschen davon ablösen können [0:07:32] S: mh ne weiß ich gerade nicht ↓ [0:07:34] TP: ne ist auch total ist auch vollkommen in ordnung lass uns doch da mal gemeinsam drüber gucken genau du hast jetzt ja angefangen das du erstmal uns diese schöne einleitung gegeben hast die würde ich auch immer beibehalten also das du immer irgendwie eine einleitung hast ähm lass mich mal gucken wo wir das hinpacken können können das gemeinsam ja hier schon mal also eine einleitung ist immer irgendwie eine pflicht	1

Category	Explanation	Specifications	Anchor	Points
			[0:07:52] S: mhmh (GBP25_Sim) 2) "Look at the tense, this is not quite right. Do you know why?" ↓ "Here you should use past simple, so 'won' would be correct. In a news report we talk about something that has happened, so we need the past simple." [0:04:08] TP: musst du hier nicht win nutzen sondern das unregelmäßige verb vielleicht weißt du es grad auch selber aus dem KOPF [0:04:13] S: nein weiß ich nicht ↓ [0:04:14] TP: okay ähm das ist dann won [0:04:16] S: mhm [0:04:16] TP: also w o n (.) wie man sieht gibt es da keine regeln für man kann nicht einfach ein ed hinten dranhängen wie das normalerweise ist (MMM07_Sim) <u>Unsuccessful (No error indicated – tests knowledge)</u> [0:03:38] TP: du schreibst ja also das ist ja auch in der echten welt so wenn du einen zeitungsartikel liest die zeitung da steht nur drin was passiert ist JA [0:03:44] S: mhm [0:03:44] TP: also die vergangenheit steht da drin (.) äh manchmal kündigen wichtige leute in zeitungen auch posts für die zukunft an da steht das natürlich in dem futur [0:03:55] S: mhm [0:03:55] TP: ja wir werden was verändern oder werden lösungen suchen JA	

Category	Explanation	Specifications	Anchor	Points
			[0:03:59] S: okay [0:03:59] TP: aber wir WERDEN bäume pflanzen das kann man machen aber grundsätzlich wenn events ereignisse in der vergangenheit waren da steht ja yesterday brauchen wir WAS [0:04:10] S: das simple past [0:04:11] TP: das simple past okAY (AOS13_Sim)	
	TP does not provide an opportunity for self-correction and only provides explicit and direct feedback.	<u>Key phrases:</u> "This is not correct" "The correct spelling is..." "You should have written..." Example: "Here you should use past simple, so 'won' would be correct. In a news report we talk about something that has happened, so we need the past simple. "	See TP23. [0:03:45] TP: ähm da hast du das deutsche geschrieben das passt danach hast du geschrieben we become even thousand dollar äh euro prize money for plant more trees also da hätte ich geschrieben become da wäre receive besser gewesen planting anstatt plant also ist eine ing form weil du sagst ja für das unternehmen äh der des einplanzens da wäre eine ing form angemessen gewesen und trees hast du falsch geschrieben das wird mit doppel e und nicht doppel s geschrieben (NDA16_Sim)	0

Revision

In process writing approaches (Hyland, 2019), writing needs a revision. Also, feedback should be used for it to be worthwhile. Thus, this category aims at TP's understanding of the writing process. As it is clear in the material, a revision phase is part of the broader series of lessons. Thus, a revision must not occur by the learner at home but potentially also in class. However, the TP must clearly indicate that a revision is necessary to achieve two points. If revision is addressed positively but presented as optional, only one point should be coded.

If instruction is unclear, contradictory, vague or a revision is negated, zero points should be coded.

Category	Explanation	Specifications	Anchor	Points
Revision	TP clearly instructs that the text needs to be revised using the feedback. This includes revision in class or at home.	It has to be sufficiently clear to the learner that it is required to rewrite the text e. g. that it is not optional. When it is formulated in a way, that it could be interpreted as optional, as in use of conjunctive, one point (see below) should be coded. If contradictory or vague messages are made (see anchor unsuccessful), the demand is not clear and zero points should be coded. <u>Key phrases:</u> "Yes, you should rewrite the text" "Please improve the text using the feedback"	Successful [0:07:34] S: okay (.) mh muss ich das denn jetzt nochmal komplett neu schreiben oDER [0:07:38] TP: genau ich würde dir empfehlen dass du das ganze einmal also du sollst das nochmal bis zur nächsten woche einmal komplett ab-schreiben und das du die verben die du falsch geschrieben hast nochmal durchgehst (AAS13_Sim) Unsuccessful – contradictory [0:08:30] S: mhmh okay (.) muss ich den text denn jetzt nochmal neu schreiben [0:08:34] TP: also komplett neu schreiben musst du den nicht also ich habe wie gesagt noch eine überschrift die fehlt [0:08:40] S: mhmh [0:08:41] TP: dann halt so ein bisschen einen absatz machen nach dieser (.) introduction (.) und das größte ist dann halt noch dieses zitat von dem ähm (.) das war von dem miniSTER ja von dem minister das du das irgendwie wenn du das nicht übersetzen willst das du es dann halt umschreibst weil zu deutsch da drin ist halt ein bisschen schlecht (.) aber ansonsten (.) so von der grundform ähm ja (.) also (KMP05_Sim)	2
	TP advises (but does not	TP indicates that a revision is useful but it is seen as optional.	Successful [0:05:48] S: heißt ich soll den text jetzt einmal neu	1

Category	Explanation	Specifications	Anchor	Points
	instruct) the learner to use the feedback, e.g., for revision or for future tasks. This includes revision in class or at home.	<u>Key phrases:</u> "If you want to, you can revise the text" "I would advise you to do it" "You should consider the feedback for next time" "That is a good idea." "Das wäre super." ["That would be great."] (Use of conjunctive) <u>Unsuccessful attempts:</u> "Try to do it better next time" "Just read through my comments again"	schreibEN und dann versuchen die sachen noch miteinzubauEN (.) [0:05:53] TP: wenn du die motivation hast ist das natürlich ein optimaler nächster schritt (MMM07_Sim) – [0:07:40] S: soll ich den text jetzt nochmal neu schreiben DANN [0:07:43] TP: also wenn du möchtest wenn du möchtest und das nochmal quasi umsetzen möchtest was ich dir gesagt habe [0:07:47] S: [ja] [0:07:47] TP: [genau] dann wäre das gut wenn du das nochmal schreibst [oder] [0:07:49] S: [okay] [0:07:49] TP: ähm ja genau verbessern ist eigentlich ja nochmal schreiben [wäre wahrscheinlich sinnvoller genau (lacht)] [0:07:51] S: [okay ja (lacht)] ja [0:07:55] TP: ähm und dann kannst du mir den auch gerne nochmal geben das ich dann da nochmal drüber gucke (SNA28_Sim) – [0:08:12] S: ja muss ich den text denn jetzt nochmal neu schreibEN [0:08:14] TP: ich würd sagen das ist eine gute übung wenn du das nochmal schreibst damit du das dann lernst und dann vielleicht auch für die klassenarbeit wenn das dann dran kommt [0:08:21] S: okay [0:08:22] TP: das ähm kannst du mir dann auch nochmal abgeben wenn du das möchtest (APD10_Sim)	
	TP does not state that the feedback needs to be used for	If contradictory or vague messages are made (see anchor unsuccessful), the demand is not clear and zero points should be coded.	[0:08:04] S: okay aber muss ich das dann jetzt nochmal alles komplett neu schreiben den artikEL [0:08:08] TP: nein [0:08:10] S: okay [0:08:10] TP: ich würde das so gelten lassen (.) wie gesagt	0

Category	Explanation	Specifications	Anchor	Points
	revision or for future tasks or TP is not clear/contradictive in their instruction.	<u>Key phrases:</u> "No, you don't have to write it again" "Do it better next time" "Just read through my comments again"	nur dass du versuchst durch deine fehler zu lernen nimm es mit und lies es dir nochmal durch und versuch einfach es das nächste mal besser zu machen (NDA16_Sim)	

Highly-inferential categories

Conditional highly-inferential categories

Dialogue orientation

Condition

As has been pointed out in the expert interviews, the dialogue should have some substance. Therefore, at least one event in the conversation that the TP triggers must be cognitively demanding for the learner so that they engage more deeply with the substance. Only if one such event has occurred, the condition is fulfilled and the general dialogic character, which includes questions about well-being or other prompts, is taken into consideration.

Criterion

In the current understanding of feedback, a dialogic approach is favored. This includes engaging the learner in the conversation in different ways. This category looks at the overall engagement of the learner while other criteria, later, look at specific aspects like self-correction and checking understanding. Codes of 0, 1, and 2 should still be coded although the condition is not fulfilled.

Category	Explanation	Specifications	Anchor	Points
Dialogue orientation	The triggered engagement is at least once <u>cognitively demanding</u> for the learner and relevant for the feedback conversation.	Cognitively demanding questions or prompts engage the learner more deeply with the context-relevant substance. This includes questions/prompts regarding self-evaluation and self-correction. This does not include simple comprehension checks and/or requests for translation as this is not part of the overall learning aim, e.g., not context-relevant. <u>Key phrases (cognitively demanding):</u> "Do you remember how the past simple is formed?" "What is the structure of a news report like?" "What were issues when writing the text?" <u>Unsuccessful attempts:</u> "I don't know what you mean here." "How are you doing?"	<u>Successful</u> [0:03:58] TP: ähm zum beispiel als äh beispiel einmal hier oben ähm du erzählst hier ja über die zeremonie [0:04:04] S: mhm [0:04:05] TP: und die war ja am diensttag das heißt das liegt in der vergangenheit [deswegen] [0:04:08] S: [ah okay] [0:04:08] TP: musst du hier nicht win nutzen sondern das unregelmäßige verb vielleicht weißt du es grad auch selber aus dem KOPF (MMM07_Sim) — <u>Unsuccessful for the condition</u> [0:04:25] TP: so you ähm it's something you can	

Category	Explanation	Specifications	Anchor	Points
		"Do you have questions?" "Do you understand?" "Can I go on in English?"	write in a diary for example you write dia a diary (APD10_Sim)	
	TP often involves the learner in the feedback conversation, e.g., through questions or prompts.	This includes all potentially learner-involving aspects, including those that are not cognitively demanding. Questions that are not part of the simulation or not directed at the learner are excluded (see unsuccessful). <u>Key phrases</u> "Do you remember how the past simple is formed?" "What is the structure of a news report like?" "What were issues when writing the text?" "Let me know if you have any questions" "Do you have questions?" "Do you understand?" "How are you doing?"	See above. <u>Unsuccessful</u> [0:03:00] TP: in dem zusammenhang (...) ne in dem zusammenhang nicht (...) was hatte ich jetzt gesaGT 0:03:10] S: äh there are glaube ich (ELH30_Sim)	2
	TP sometimes involves the learner in the feedback conversation, e.g., through questions or prompts.	See above & comments on frequency.	See above.	1
	TP does not involve the learner in the feedback conversation, e.g., through questions or prompts which are at least once cognitively demanding for the learner.	See above & comments on frequency.	See above.	0

Quality of Feed Back

Condition

Feed back should elaborate on the discrepancy between actual and a desired state which is described in objectives. Consequently, the TP should at least once explicitly state feed back against the learning objective.

Criterion

Considering the axioms of nonviolent communication by Rosenberg, feedback should be provided on a descriptive level. Furthermore, feedback should be clear (e.g. Ferris, 2013) for the learner and naturally be correct.

Cate- gory	Explanation	Specifications	Anchor	Points
Qual- ity of Feed Back	The feedback is at least once explicitly provided against specific learning objectives.	It must be clear what the objective is. It is sufficient to imply that criteria for the news report are learning objectives. The event must contain correct information with regards to both the objective as well as the diagnosis. Feedback provided can also include positive feedback. Example: "In order to write a great news report, you need to use the past tense. Looking at that objective, we can see that you used present tense when you wrote 'We win'."	Successful [0:00:58] TP: okay so if you take a look maybe at this article äh here (.) hm it has a title and you need to have one too and then it starts with celebrities have joined Prince William and so on so if ähm somebody reads your article ähm he oder she doesn't really know ähm maybe what the nrw umweltpreis is so (AHW03_Sim)	
	TP provides feedback that is often accurate, clear and on a descriptive level.	Accuracy refers to the necessity of the information of the feedback to be correct. Clarity refers to the necessity of being	Successful [0:02:18] TP: genau das hätte ich nämlich auch gesagt das was okay mir ist nämlich aufgefallen dass du ähm in manchen bereichen äh ganz gut bescheid wusstest vor allem über das projekt ähm was ihr gemacht habt und äh zum beispiel picking up	2

Category	Explanation	Specifications	Anchor	Points
		precise regarding what the issue is. Descriptiveness refers to the necessity that the feedback is provided on neutral grounds and no judgement is made. See also comments on frequency.	garbage und so weiter ähm das projekt hast du beschrieben und ähm man hat gemerkt dass du dich ähm mit dem themenfeld ähm ganz gut auseinander gesetzt hast ähm was mir ähm aufgefallen ist ist dass du ähm einige fehler gemacht hast die trotzdem aus diesem themenfeld kommen zum beispiel (unv.) kann ich dir direkt am text zeigen [ähm] [0:02:54] S: [ja] [0:02:54] TP: das wort action (.) ähm das ist sehr deutsch das ist ein typisches [denglisches] [0:02:59] S: [ja] [0:02:59] TP: wort ähm was besser wäre wäre zum beispiel project so wie es auch hier in deinem in deiner aufgabenstellung steht (RNM25_Sim)	
	TP provides feedback that is sometimes accurate, clear and on a descriptive level.	See above & comments on frequency.	See above.	1
	TP never provides feedback that is accurate, clear and on a descriptive level.	See above & comments on frequency.	<u>Unsuccessful - not accurate</u> [0:00:41] TP: und zwar ist mir aufgefallen das du vor allem immer den deutsche satzbau mehr benutzt hast und nicht den englischen [0:00:48] S: ah okay ja [0:00:50] TP: beispielsweise (.) hier (.) in the gymnasium we all meet [0:00:56] S: mhmh [0:00:57] TP: und das wäre dann halt eher (.) we all met in the gymnasium (SLG21_Sim) <u>Unsuccessful - not clear</u> [0:04:15] TP: [...] (.) ähm it give different projects also das habe ich	0

Category	Explanation	Specifications	Anchor	Points
			jetzt nicht ganz verstanden was du meintest da habe ich jetzt (.) also ich hätte dir vorgeschlagen dass du geschriebe hättest there WERE different projects [0:04:31] S: mhm also das waren so verschiedene sachen die wir gemacht haben da auf dem [0:04:34] TP: also es gab verschiedene projekte [0:04:35] S: genau [0:04:36] TP: genau dann hätte ich geschrieben there were different projects [0:04:39] S: okay [0:04:39] TP: oder many more projects hätte ich dann tatsächlich äh geschrieben [0:04:44] S: aber weil sich das dann besser anhört oder [0:04:45] TP: ja genau also äh das was du geschrieben hast ist so also ich weiß nicht falsch ist es nicht doch it give different also du sagst (.) es also nach dem motto irgendetwas gibt [0:04:59] S: also es gibt verschiedene projekte [0:05:00] TP: nein nein du also was du hier versuchst zu sagen in den worten es bedeutet irgendetwas gibt projekte raus als würde irgendetwas projekte geben [0:05:09] S: ach so [0:05:09] TP: aber du meinst ja es gibt also deswegen there were [0:05:11] S: ja [0:05:12] TP: different projects (NDA16_Sim) <u>Unsuccessful - not descriptive</u> [0:03:37] TP: okay so ähm another thing so the next (.) i mean it's not a big	

Category	Explanation	Specifications	Anchor	Points
			problem but the next thing that you should really work on is you have this direct quote by the ähm environmental minister and it's just in german so you just copied this quote and i think you should you know if you can't translate it directly you should have just summarized what he said you know because just having a quote in german (.) in a news-report that's (.) ja that's not so great (KMP05_Sim)	

Relative frequency-based categories

Body language

TP should indicate through their body language that they include the learner in the conversation. This is evident in having appropriate eye-contact with the learner, nodding when they are talking, facing the learner with their posture or projecting a comfortable atmosphere

Category	Explanation	Specifications	Anchor	Points
Body language	TP consistently conveys attentiveness through body language, appropriate eye-contact and facial expressions.	All these characteristics should be fulfilled to a certain degree - Upper body turned towards the learner - appropriate eye-contact in adequate situations - projecting comfort with facial expressions <u>Key events:</u> TP faces the learner. TP's upper body is turned towards the learner. TP looks at the learner. TP greets the learner with a hand shake. TP indicates listening through nodding. TP nods at the learner.	<u>Successful</u> [0:06:59] TP: richtet den Blick in die Richtung der S [0:06:59] S: nickt mit dem Kopf [0:07:00] TP: richtet den Blick auf das Material [0:07:08] S: nickt leicht mit dem Kopf [0:07:15] TP: richtet den Blick in die Richtung der S (MMM07_Sim) – [0:01:03] TP: richtet den Blick auf das Material und dann in die Richtung der S und senkt die Hand in kreisenden Bewegungen (NRS18_Sim) <u>Unsuccessful</u> [0:05:14] TP: nimmt den Zettel in die Hand [0:05:15] TP: bewegt den Kopf nach unten [0:05:17] TP: zeigt mit dem Stift auf den Zettel (M1) [0:05:32] TP: bewegt die Hand zur Seite [0:05:32] S: bewegt den Kopf hoch und runter [0:05:34] TP: bewegt die Hand zur Seite (NDA16_Sim)	2

Quality of Feed Forward

This category looks at the feed forward element in the feedback model by Hattie & Timperley (2007). Feed forward should contain precise actions that the learner should carry out following the information that is provided. It should potentially decrease the gap between actual and desired state. Following the notions of nonviolent communication, these calls to action should be formulated positively (Rosenberg & Holler, 2016). The TP should be clear in that their calls to action will not be misinterpreted by the learner. The call to action should be specific to a diagnosed element in the text and serve as a means to improve on that aspect and also contain precise instructions (Ferris, 2014).

Category	Explanation	Specifications	Anchor	Points
Quality of Feed Forward	TP provides calls to action that are often positive, unambiguous and specific.	A call to action contains a demand/advice for the learner which they should enact and follow up on to achieve a specific learning objective. Positive: Stating what the learner should do instead of what they should not do. Unambiguous: The statement can only be understood by the learner in one possible way and is not contradictory in itself. It clearly states what the learner should do next and how they should do it. Specific: The suggested or demanded action targets a specific diagnosed error and how the learner can improve on that aspect as well as	Successful [0:03:14] TP: vielleicht ähm kann man das nochmal kurz anreißen auf jeden fall ähm war das jetzt das auffälligste da würde ich dir vielleicht als tipp mitgeben nochmal die ähm unREGelmäßigen verben zu lernen [0:03:24] S: mhm [0:03:25] TP: ähm und dass du da weißt wie man von von der infinitiv die vergangenheit formt [0:03:30] S: aber wie kann ich das den lerNEN [0:03:31] TP: das äh steht zum beispiel hinten in den büchern immer [0:03:34] S: hmh [0:03:34] TP: ähm die vokabeln werden wahrscheinlich jetzt nicht mehr in den units der vokabeln sein aber ganz hinten im buch [ist] [0:03:40] S: [hmh] [0:03:40] TP: äh so eine seite da ist so eine ganz große tabelle [0:03:43] S: okay (räuspern) [0:03:44] TP: und da sind immer die unregelmäßigen verben aufgelistet ähm hinter dem infinitiv [0:03:48] S: mhm okay [0:03:48] TP: das heißt da kann man das nochmal ganz gut lernen weil bei denen ist es ja leider so dass man die auswendig lernen muss (MMM07_Sim)	2

Category	Explanation	Specifications	Anchor	Points
		providing specific instructions on what the learner needs to do. <u>Key phrases</u> (providing specific next steps) "One piece of advice/tip for the future..." "You need to..." "You can look at the text/page..." "I will give you a task for revision" <u>Unsuccessful attempts:</u> "We will talk about it in class." "You need to revise the tenses." "You need to learn them by heart." (without providing specifics) "There is a list."		
	TP provides calls to action that are sometimes positive, unambiguous and specific.	See above & comments on frequency.	See above.	1
	TP provides calls to action that are never positive, unambiguous and specific.	Instances of feed forward that do not address the specific case and/or do not provide more detailed information on how the objective can be achieved are not sufficient. <u>Key phrases:</u> Not positive:	<u>Unsuccessful – not positive</u> [0:05:28] S: aber das habe ich dann ja öfter falsch gemacht mit der zeitform oDER [0:05:31] TP: genau das simple past solltest du auf jeden fall nochmal durchgehen insgesamt also generell die zeitformen [0:05:36] S: machen wir das auch nochmal oder soll ich das auch nochmal alleine üben DANN [oder wie kann ich das üBEN] [0:05:40] TP: [ja wir werden bei der klassen] vor der	0

Category	Explanation	Specifications	Anchor	Points
		"You need to avoid..." Not unambiguous: "You could ask your peers, but they might not be helpful" Not specific: "You just need to learn them by heart" "You have to read and speak more in English" "Watch more English films"	klassenarbeit werden wir nochmal eine übung darauf ähm basierend also auf den fehlern der anderen schülerinnen und schülern nochmal (.) basierend die übung machen [0:05:50] S: okay [0:05:51] TP: und dann (.) um das in der klassenarbeit zu vermeiden (AAS13_Sim) <u>Unsuccessful - not unambiguous</u> [0:08:01] auf jeden fall nochmal machen du kannst sowieso immer zu mir kommen und mir irgendwas was du (unv.) ja wenn du nochmal feedback oder sowas brauchst weil besser du fragst nochmal nach anstatt das du dir unsicher bist oder so ich meine du kannst natürlich auch irgendwen aus der klasse fragen aber die können dir dann vielleicht auch nicht so weiterhelfen dann (AHW03_Sim) <u>Unsuccessful - not specific</u> [0:03:09] TP: also vor allem was mir halt auch aufgefallen ist das du die unregelmäßigen verben nochmal lernen musst [0:03:17] S: wie könnte ich das denn üBEN [0:03:19] TP: immer wiederholen die musst leider die musst du einfach auswendig lernen [0:03:23] S: [okay] [0:03:23] TP: [du musst] da wirklich immer wiederholen immer wieder weil (.) da gibt es leider keine richtige regel die du halt lernen kannst damit du es lernen kannst [0:03:33] S: [okay] [0:03:33] TP: [sondern] immer wiederholen und learning by heart [das ist (unv. Ca. 1 Sek.)] [0:03:36] S: [okay also einf]ach auswendig lernEN [0:03:38] TP: genau genau und äh schreibe sie dir immer wieder auf immer wieder	

Category	Explanation	Specifications	Anchor	Points
			und also (.) was was noch anderen schülerinnen und schülern (.) ähm geholfen hat oft ist halt immer wie- der dieses (.) wenn man was sagt vielleicht mal kurz nur in sich gehen und sagen wie hätte ich das jetzt auf eng- lisch geSAGT (CCF27 Sim)	

Target-oriented error selection

Feedback should enable the learner to achieve certain learning objectives. Consequently, the feedback conversation should predominantly revolve around errors the learner has made that need to be addressed in order to achieve this target. This includes single errors relevant for the learning objective as well as error patterns.

In order to code two points or one point, a selection of errors needs to be evident, thus a chronological discussion of the text without conscious selection of aspects will lead to t

e coding of zero points. A TP can start chronologically but still show that they select errors and receive points.

Non-target-related errors can include issues of orthography, punctuation, syntax.

Two points are coded if non-target-oriented errors are discussed marginally but a clear selection of errors for the target is still evident.

One point is coded if a selection of errors is evident but the TP also included non-target-error oriented errors equally.

Zero points are coded if no selection is evident, or if a selection is evident that only focuses on non-target specific errors.

Category	Explanation	Specifications	Anchor	Points
Target-oriented error selection	TP often provides feedback on selected errors, error patterns or mistakes that are relevant for achieving the learning objectives. Other errors, error patterns or mistakes that are not directly related to the learning objectives can be discussed marginally.	A conscious selection of errors must be evident. When target-specific errors are addressed in the context of a chronological discussion of the text, this is not sufficient.	Successful See TP 33. [0:01:08] TP: äh und zwar in der ersten stunde ging es vor allem um news reports verstehen wie der ähm aufbau war und wie wie eben ein typischer news report aufgebaut ist und ähm mir fällt auf in deinem text dass du das ähm sehr gut gemacht hast du hast am anfang das projekt vorgestellt du hast ähm gesagt wer alles da war worum es grob geht und ähm mit einem (.) äh mit einem ganz guten ende beendet soweit ist da alles in ordnung [0:01:41] S: okay [0:01:42] TP: ähm wie siehst du DA ähm wie siehst du DAS hättest du	2

Category	Explanation	Specifications	Anchor	Points
			hat dir das leicht gefal- LEN [0:01:47] S: mh ja also ich habe das einfach so runter geschrieben dann [0:01:52] TP: okay (.) ja das ist dann sehr gut ge- laufen ähm dann äh die zweite sache die ähm du hättest können sollen war ähm oder die du k äh kön- nen sollst ist ähm das themenfeld environment be- schreiben [ähm] (RNM25_Sim)	
	TP sometimes provides feed-back on se- lected errors, error patterns or mistakes that are rele- vant for achiev- ing the learn- ing objectives. Other errors, error patterns or mistakes that are not di- rectly related to the learning objectives are discussed equally.	See above.	See TP1.	1
	TP does not provide feed-back on se- lected errors, error patterns or mistakes that are rele- vant for achiev- ing the learn- ing objectives. Errors, error patterns or mistakes are discussed chronologically	TP discusses aspects in the text chrono- logically as they ap- pear. Starting with aspects that occur at the be- ginning of the text does not, however, necessarily indicate that the errors and/or error patterns are ad- dressed in a chrono- logical manner, as a subsequent con- scious selection is still possible.	Unsuccessful [0:03:45] TP: ähm da hast du das deutsche geschrie- ben das passt danach hast du geschrieben we become even thousand dollar äh euro prize money for plant more trees also da hätte ich geschrieben become da wäre receive besser gewe- sen planting anstatt plant also ist eine ing form weil du sagst ja für das unternehmen äh der des einplanzens da wäre eine ing form angemessen gewe- sen und trees hast du falsch geschrieben das	0

Category	Explanation	Specifications	Anchor	Points
	without selecting specific errors.	TP selects and discusses aspects in the text randomly that are not relevant for achieving the learning objectives.	wird mit doppel e und nicht doppel s geschrieben (NDA16_Sim)	

Detailedness

This category aims at the complexity and contextualization of the feedback. When information is provided to the learner, it should contain some form of elaboration as in explaining the rules of why a specific aspect is incorrect or give examples of how it could be done. It should not only rely on providing a correct solution (KCR) or providing information on whether the answer is correct or incorrect (KR). The extent is not decisive as this category seeks to account for the question of whether elaboration as occurred with the feedback.

For coding of two points, the feedback should more often than not be elaborated, in that explanations are included, reasons are given or examples provided. As in some instances feedback does not require elaboration, e.g., spelling mistakes that could be identified as slips in performance, some instances of KCR are appropriate.

For one point, elaboration should occur in some instances but the TP tends to give more simplistic feedback through KCR and KR.

Zero points are coded if the TP does not provide any (correct) elaboration on any aspect.

Category	Explanation	Specifications	Anchor	Points
Detailedness	TP often provides specific elaborated feedback, e.g., explains rules, gives reasons or provides examples. In some instances, only KCR is provided i.e., only the correct solution is given.	TP, upon requests or self-initiating, provides correct details on why certain aspects have been achieved successfully or need to be improved on. This includes: • explanations of rules, e.g., forming of regular past simple forms • giving reasons for the use of past simple, e.g., the genre and the fact that the event that is reported on has already happened	Successful See TP 6. [0:02:00] TP: und genau generell einfach alles ein bisschen ähm mehr schreiben können zu zum beispiel auch den projekten die ihr gemacht habt du hättest ähm genau die einfach vielleicht erklären können ähm zum beispiel hattet ihr ja hier ihr hattet ja waffeln verkauft oder kunst verkauft oder halt ähm müll müll ähm aufgesammelt das hättest du einfach vielleicht ein bisschen mehr erklären können ähm was die schüler da gemacht haben genau	2

Category	Explanation	Specifications	Anchor	Points
		providing examples, e.g., using the sample news report or giving correct example sentences Explanations do not have to be given consistently, e.g., when spelling mistakes are corrected, details or not necessary and KCR is appropriate.	[0:02:25] S: also das heißt vom inhalt her ist noch nicht alles drIN [0:02:28] TP: genau also der inhalt STIMMT aber vielleicht einfach ein bisschen mehr erklären dass man sich das ein bisschen vorstellen kann weil wir hatten ja hier zum beispiel die ähm das beispiel ne von prince william [0:02:39] S: ja [0:02:39] TP: ähm und wenn du dir das anguckst ist das ja schon ein bisschen länger als das zum beispiel also so ein zeitungsartikel sollte ja vielleicht schon ein bisschen ja so ein news report sollte vielleicht schon ein bisschen mehr informationen enthalten das ist jetzt die wichtigste information aber ähm genau einfach vielleicht ein bisschen mehr dass man sich da ein bisschen mehr darunter vorstellen kann genau (BTB01 Sim)	
	TP sometimes provides specific elaborated feedback, e.g., explains rules, gives reasons or provides examples. KCR and KR dominate, i.e., only the correct solution is given or it is said whether something	See above.	See TP27.	1

Category	Explanation	Specifications	Anchor	Points
	is (in)correct.			
	TP only provides KR or KCR without specific elaboration, i.e., only the correct solution is given or it is said whether something is (in)correct.	TP does not provide (correct) elaboration or is unspecific in the elaboration.	See TP17 [0:00:42] TP: ähm und zwar hast du bei einigen sätzen das simple past vergessen [0:00:48] S: mhmh [0:00:49] TP: ähm und du hast auch einige wörter falsch geschrieben [0:00:54] S: okay [0:00:55] TP: ähm aber ansonsten war es an sich nicht falsch also du hast nichts falsches reingeschrieben (CCF27_Sim)	0

Absolute frequency-based categories

Amount and balance of information

This category seeks to evaluate whether the kind of information that is discussed in the conversation is appropriate. This includes both quantity as well as quality. While too much information could potentially lead to a cognitive overload on the learner's side, too little information would not be enough for the learner to actually improve substantially in the text. As a guideline, 3-5 aspects are appropriate, although the individual depth and context in which each area is discussed must be considered in the evaluation.

Depending on the context, one aspect can include several pieces of information.

With regards to quality, both surface- and deep-level features should be addressed in the text in a balanced way, so that the learner can improve on both facets in the text.

Category	Explanation	Specifications	Anchor	Points
Amount and balance of information	TP provides an appropriate amount information on different aspects of the text and the writing process (about 3-5 aspects). Surface- and deep-level features are addressed in a balanced proportion.	One aspect of the text can contain multiple pieces of information if it is addressed as a pattern (e.g. irregular verbs, see anchor) or framed under an umbrella term (e.g. style).	Successful See TP49. [0:01:24] TP: du verwendest hier auf jeden fall schon hier ähm das ein oder andere mal äh das simple past das sieht man hier auch zum beispiel ähm du hast es nur falsch verwendet [0:01:34] S: ja [0:01:35] TP: aber generell sieht man an vielen momentan hast du noch dass du ähm in der gegenwartsform redest und du müssest hier in das simple past wechseln [0:01:45] S: okay [0:01:45] TP: also zum beispiel hier direkt am anfang ähm the heinz nixdorf gesamtschule WON und nicht win [0:01:52] S: ah okay [0:01:54] TP: here again äh where was it äh hier BOUGHT bought (MJO28 Sim)	2

Category	Explanation	Specifications	Anchor	Points
	TP provides an appropriate amount information on different aspects of the text and the writing process (about 3-5 aspects). However, surface- and deep-level features are not addressed in a balanced proportion.	See above. An unbalanced proportion is evident if both surface- and deep-level features are mentioned, however one level clearly dominates the conversation.	See TP32.	1
	TP provides an inappropriate amount of information (too little, too much), regardless of balance between surface- and deep-level features.		<u>Unsuccessful</u> See TP23. [0:03:45] TP: ähm da hast du das deutsche geschrieben das passt danach hast du geschrieben we become even thousand dollar äh euro prize money for plant more trees also da hätte ich geschrieben become da wäre receive besser gewesen planting anstatt plant also ist eine ing form weil du sagst ja für das unternehmen äh der des einplanzens da wäre eine ing form angemessen gewesen und trees hast du falsch geschrieben das wird mit doppel e und nicht doppel s geschrieben (NDA16_Sim)	0

Checking understanding

This category looks at the willingness of the teacher to adapt to learner's level of understanding by actively checking comprehension and adapting the content of their statement. This category explicitly excludes the willingness to adapt towards the learners' level of foreign language competence. It seeks to focus on the act of checking understanding as this is a requirement for potentially adapting to the learner. This category is therefore derived from the theory of the zone of proximal development by Vygotskij (1978)

Category	Explanation	Specifications	Anchor	Points
Checking understanding	TP assures themselves several times of learner's understanding and adapts to their needs, e.g., by re-phrasing or giving alternative explanations.	TP must actively check understanding. If the learner indicates that they did not understand, TP must adapt. Principal offers to the learners to ask questions are not included here and are an aspect of dialogue orientation. <u>Key phrases:</u> "Do you have questions on that?" "...okay?" "Do you understand that, or would you like me to explain?" "Do you know what that/... means?" In the sense of "Do you understand?" <u>Unsuccessful attempts:</u> "Can I go on in English?" (Language Adaptation)	<u>Successful</u> [0:07:10] TP: genau (.) dann ähm ich weiß nicht ob du vielleicht noch weitere fragen HAST [0:07:15] S: mh ne eigentlich nicht also ich würde mir den text dann nochmal angucken und dann genau gucken was ich dann verbessern kann (MMM07_Sim) [0:06:51] TP: [das sind] genau dazu kommen wir jetzt ähm und zwar ähm ähm hast du die neutralität vergessen also du als ähm schreiberin dieses textes bist ja neutral du schreibst ja nicht über uns als schule sondern über die schule aus einer äußeren perspektive wie es ein reporter machen [würde] [0:07:11] S: [mhm] [0:07:11] TP: verstehst du DAS [0:07:12] S: ja und was habe ich dann daran falsch geMACHT [0:07:14] TP: genau du hast oft ähm we become was in diesem fall ja we receive oder we get heißen sollte (RNM25_Sim) <u>Unsuccessful (general offer for questions)</u> [0:03:55] TP: ähm (.) des weiteren hattest du (5 Sek.) wir versuchen jetzt mal mit dem englischen wenn du etwas nicht	2

Category	Explanation	Specifications	Anchor	Points
			verstehst dann kannst du es ja sagen oKAY (APD10 Sim)	
	TP assures themselves a few times, but at least once, of learner's understanding and adapts to their needs, e.g., by rephrasing or giving alternative explanations.	See above.	See above.	1
	TP never assures themselves of learner's understanding and does not adapt to their needs, e.g., by rephrasing or giving alternative explanations.	See above.	-	0

Language accuracy

As a role model, the TP should speak English well and refer to linguistic aspects in the learner text correctly. Their performance in the English language, as well as their knowledge about the English language should be excellent.

More severe inaccuracies in performance and knowledge, especially when they are interfering with the learner's achievement of the learning objectives need to be taken into account and points deducted. When coding, the frequency and severity of the potentially harmful statement needs to be considered.

Minor slips are natural to oral language production and two points can still be coded if that is the case.

One point is coded if TP is predominantly correct when it comes to using and referring to the English language but shows rare inaccuracies.

Zero points are coded if the TP shows consistent inaccuracies.

Category	Explanation	Specifications	Anchor	Points
Language accuracy	TP uses the foreign language correctly almost always, with only a few minor inaccuracies that do not potentially hinder learning, e.g., providing incorrect corrections. This includes speaking English as well as when talking about the English language, e.g., providing corrections or rules on surface-level features.	TP makes only minor slips that are normal in oral speech.	Minor slip [0:02:02] TP: und ähm can we talk in english agAIN (.) and ähm (.) here you need to take a look on your tenses so if your write an article ähm (AHW03_Sim)	2
	TP predominantly uses the foreign language correctly. However, the TP also shows rare instances of inaccuracies and errors that potentially hinder learning, e.g., providing	The amount and quality of inaccuracies needs to be considered in the decision between one point and zero points. See below.	Error [0:04:26] TP: genau oder zum beispiel auch ähm hast du geschrieben we become even one hundred thousand euros prize money ähm das ist auch ein unregelmäßiges wort ähm (..) da kann ich dir jetzt auch sagen das bedeutet we äh das muss	1

Category	Explanation	Specifications	Anchor	Points
	incorrect corrections. This includes speaking English as well as when talking about the English language, e.g., providing corrections or rules on surface-level features.		heißen we became (MMM07_Sim)	
	TP uses consistent inaccuracies and errors in the English language that either in sum, or individually, potentially hinder learning. This includes speaking English as well as when talking about the English language, e.g., providing corrections or rules on surface-level features.	TP makes several mistakes that potentially hinder the learner because they are • central for achieving the learning objective; • display grave issues in or lack of TP's linguistic knowledge; • display grave issues in TP's linguistic performance.	Unsuccessful See TP13. [0:01:43] TP: ähm in the second sentence you wrote down it is for the action plant trees for bees first it's seems correct but you have to äh always choose äh nouns with äh (.) what do you call it ähm (..) maybe I say it in german äh der hat ähm (..) das ist klein geschrieben wird also genommen im englischen werden immer klein geschrieben [0:02:07] S: ach so okay aber das hier ist das aber okay weil das [0:02:11] TP: äh ja genau im deutschen ist das in ordnung weil es dann quasi äh eins zu eins übernommen wird weil es [0:02:16] S: mhm [0:02:17] TP: quasi äh man würde dann quasi glaube ich den kontext auch so gar nicht mehr hinterher dann verstehen [0:02:22] S: okay [0:02:23] TP: ähm maybe you could do this in an äh write this in an apostrophe but maybe it's not really necessary (NRS18_Sim)	0

Professional accuracy

This category looks at the broader correctness of TP's statements. TP's should not make any statements to the learner that could potentially hinder their learning, thus two points are only coded when no statement is potentially harmful. One point is still coded if the TP is predominantly correct. Zero points are coded when the TP shows consistent harmful statements. When coding, the frequency and severity of the potentially harmful statement needs to be taken into account as well

Any language-related statements are considered in the category "language accuracy" and are not relevant here.

Category	Explanation	Specifications	Anchor	Points
Professional accuracy	TP is always correct, e.g., does not potentially hinder learning when addressing professional non-language related aspects and providing information on them, e.g., deep-level features and learning strategies.	No mistakes are made. Other than with regards to language accuracy, these errors/mistakes are not potential slips but each false statement bears the potential to hinder learning.	<u>Not potentially hindering learning.</u> See TP2. [0:04:25] TP: [genau] (.) ja und ähm (.) also ich würde sagen das wenn du einen ja report auf englisch schreibst das du auch den satz (.) ja auf englisch übersetzen kannst [0:04:36] S: [mhmh] [0:04:36] TP: [also] nicht wort für wort sondern einfach (.) ja den sinn des satzes (ACG22 Sim)	2
	TP is predominantly correct when addressing professional non-language related aspects and providing information on them, e.g., deep-level features and learning strategies. However, TP also shows rare instances of inaccuracies and errors that potentially hinder	See below.	<u>Potentially hindering learning.</u> See TP43. [0:05:26] S: ja wie kann ich das denn dann dann üben [0:05:28] TP: tatsächlich kann man false friends NUR durch auswendig lernen [0:05:33] S: okay (BNP09_Sim)	1

Category	Explanation	Specifications	Anchor	Points
	learning, e.g., providing incorrect corrections.			
	TP shows consistent inaccuracies when addressing professional non-language related aspects and providing information on them, e.g., deep-level features and learning strategies. TP does not address non-language related, e.g., deep-level features and learning strategies.	(Common) events: - Arguing that the German quote must not be translated / not commenting on the lack of translation - Arguing that the final evaluative comment is appropriate/praising it - Arguing that, e.g., irregular verbs can only be learned by learning them by heart, excluding other viable communicatively oriented options - Providing incomplete feedback/corrections - Making false or contradicting statements - Claiming that not all the information	<u>Potentially hindering learning.</u> See TP23. [0:07:27] TP: aber wenn du es auf deutsch hingeschrieben hättest so wie du es hier auch geschrieben hast wäre das auch in Ordnung letztlich sagt er es ja auf deutsch [0:07:32] S: mhm (NDA16_Sim)	0

Category	Explanation	Specifications	Anchor	Points
		needs to be used - Providing false "corrections" for the learner - Not correcting/repeating mistakes or errors - Non-answers (not providing actual information on the question) - Inability to answer learner's questions - Arguing for simply taking out sentences or parts instead of improving them - Providing grades on the text without being asked to as this is in a formative setting - Arguing that surface-level features are more important than deep-level features - Correcting third person singular -s (tense not required in the report)		

Excluded categories

These categories have been excluded based on the findings on the internal reliability of the item scale.

Use of material

The material that is given should be used in order to visually enhance the feedback. This can be achieved by pointing out specific text fragments in the learner text or by using the task sheet or model text in order to compare them to the learner text. The spatial arrangement, however, should allow the learner to see what the TP is showing, thus it is not sufficient to merely point to the text, but it must be in a way that it is directed towards the learner.

Category	Explanation	Specifications	Anchor	Points
Use of material	TP includes the material in the conversation at least once explicitly and spatially, so that it is accessible to the learner, e.g., by pointing at specific aspects in the text.	The material should be arranged spatially in a way that it is accessible for the learner, e.g., they can potentially read it. <u>Key actions:</u> TP points to M1/M2 <u>Key phrases:</u> "In the task it says...." "Look at the text..." <u>Unsuccessful attempts:</u> TP looks at the material but does not use it to explain/emphasize/provide feedback. TP points to the text but does not spatially rearrange the material so that it is accessible to the learner, e.g., the text is upside down for the learner.	Successful Video [0:01:24] TP: richtet den Blick kurz in die Richtung der S und dann auf das Material [0:01:27] TP: berührt mit der linken Hand die untere Kante von M2 und hebt und senkt diese mehrmals [0:01:32] TP: zeigt auf M2 [0:01:34] TP: legt den Arm auf der Armlehne ab [0:01:37] TP: legt die rechte Hand flach auf M2 [0:01:38] TP: tippt mit dem Zeigefinger der rechten Hand mehrmals auf M2 (MMM07_Sim) Audio [0:01:26] TP: ähm (..) genau ähm (..) hier ist ja auch ähm steht ja auch dass ihr die die struktur benutzen sollTET [ähm würdest du] [0:01:36] S: [JA] [0:01:36] TP: sagen dass das hier ähm wie es hier aufgebaut ist die struktur ist die in einem news report gebraucht werden sollTE (...) (MMM07_Sim) Unsuccessful: TP29	2

Language adaptation

The aim of this category is to evaluate TP's ability to adapt to the learner's language level. As one paradigm of language teaching, it is expected to conduct the majority of the instruction in the foreign language. When learners indicate that they, however, cannot understand the language, the end of this should be to use the language the learners understand. Therefore, the best approach is to start in English and switch to German when required. Alternatively, but not adhering to the goal of monolingualism

is to start in German but providing reasons, indicating that the TP adapts. Zero points are coded if the TP does not show adaptivity by not addressing the language issue.

Category	Explanation	Specifications	Anchor	Points
Language adaptation	TP adapts to the learner's language needs, by starting the conversation in English and switching to German when required.	<u>Possible approaches:</u> 1) TP attempts to conduct the conversation in the foreign language (e.g. English) AND Potentially adapts to the learner's foreign language level in the foreign language AND continues the rest of the conversation in German, if asked to do so. 2) TP asks learners *in English* in which language they prefer the feedback conversation. 3) TP, at the start, offers *in English* to switch to German at any time, upon request. 4) Starts in English but switches to German and gives reasons for it. <u>Key phrases:</u> "In which language do you want the feedback?" "Let me know if you don't understand what I am	1) [0:00:26] TP: that takes a lot of courage ähm what do you think about your TEXT (.) itself [0:00:32] S: mh können sie die frage nochmal auf deutsch stellen [0:00:34] TP: natürlich ähm (..) wie ähm findest du denn selber den TEXT (räuspern) [0:00:49] TP: von mir aus können wir tatsächlich das ganze gespräch auch auf deutsch machen das ist immer ein bisschen persönlicher [0:00:54] S: ja gerne (lachen) [0:00:54] TP: englisch kreiert ja ein bisschen so mehr distanz [ähm (MMM07_Sim)] 2) [0:00:18] TP: good morning (..) take a seat (.) so do you want to get feedback in german or in english you can decide [0:00:27] S: hm i think in german [0:00:29] TP: okay gut (.) (CML16_Sim)	2

Category	Explanation	Specifications	Anchor	Points
		saying, I can explain it in German"		
	TP adapts to the learner's language needs, but starts the conversation in German.	<u>Possible approaches:</u> 1) TP starts the conversation in German and gives reasons for that. 2) TP offers *in German* to switch languages at any time upon learner's request. 3) TP asks learners *in German* in which language they prefer for the feedback conversation <u>Key phrases (in German):</u> "Let's do this in German because..."	1) [0:00:18] S: [good mor]ning [0:00:22] TP: ähm ich dachte wir machen dieses gespräch einmal in deutsch [0:00:26] S: mhmh [0:00:27] TP: es ist ja auch schon nach der sechsten stunde ähm da hat man wahrscheinlich auch nicht mehr so viel energie (BNP09_Sim)	1
	TP shows no adaptivity.	<u>Possible approaches:</u> 1) TP starts and conducts the conversation in German without giving reasons. (Does not show adaptive skills). 2) TP attempts to conduct the conversation in the foreign language (e.g. English) AND potentially adapts to the learner's foreign language level in the foreign language BUT does *not* continue the rest of the conversation in German, when asked.	1) [0:00:20] S: good morning [0:00:21] TP: good morning (8 Sek.) so genau ähm ich habe jetzt deinen text nochmal durchgelesen mia (AAS13_Sim)	0

Deep: Process

This category targets TP's diagnostic abilities and is based on the framework by Becker-Mrotzek (2014).

Writing process refers to the steps that the learner has taken in order to write the text. This can be done by identifying the lack of process elements in the text (lack of notes, copying from the task sheet, experimenting with language) or through asking the learner about their process. This category looks only at the diagnosis of process aspects and consequently only at feed back, not at feed up or feed forward.

Category	Explanation	Specifications	Anchor	Points
Deep: Process	TP refers to aspects of process, e.g., missing revision tracks, copying phrases or taking risks at least once.	Diagnosis must be correct and specific. Specificity can also be achieved through use of material. <u>Key phrases:</u> "It is great that you tried to use these new words." <u>Unsuccessful attempts:</u> "You need to improve your writing process."	Successful [0:05:24] TP: und das fällt etwas schwierig weil das sind auch sachen die man nicht vielleicht immer unbedingt in den texten verwendet aber du warst ja so mutig und hast es da wirklich getan und probiert damit zu arbeiten also da auf jeden fall gucken (GBP25_Sim) TP here provides positive feedback on the risk the learner took during the writing process (cf. Becker-Mrotzek, 2014)	2

C. Demographic and Academic Data Collected in the Written Survey

This is a translated version. The participants received the questions in German.

Personal Data

- 1 What is your gender?
 - a. Female
 - b. Male
 - c. Diverse
 - d. Not Statement
- 2 How old are you?
- 3 In which study program are you enrolled?
 - a. [List of Programs from Paderborn University]
- 4 In which phase of study are you?
 - a. Bachelor
 - b. Master
 - c. *Grundstudium*
 - d. *Hauptstudium*
- 5 In which semester did you start your teacher education studies?
 - a. Winter Term 2011/2012
 - b. Summer Term 2012
 - c. ...
 - d. Winter Term 2022/2023
 - e. Other
- 6 Which subjects do you study? [Up to four subjects possible]
 - a. [List of possible subjects of Paderborn University]
 - b. Other
- 7 Which is your highest degree before starting teacher education studies?
 - a. Graduation (*Abitur*)
 - b. Other University Degree
 - c. Vocational Apprenticeship
 - d. Vocational Diploma (*Fachabitur*)
 - e. Other
- 8 Please state your final graduation grade, i.e., (*Fach-*) *Abiturnote*, or grade of apprenticeship.
- 9 Did you have an intensive course (*Leistungskurs*) English-B?
 - a. Yes
 - b. No
- 10 In which federal state did you graduate (i.e., *Abitur*)?
 - a. List of German Federal States

Practical Teacher Education

11 I completed the following practical phases in my studies:

- a. AOI
 - i. Not yet started
 - ii. In the phase
 - iii. Completed the Phase
- b. OI
 - i. Not yet started
 - ii. In the phase
 - iii. Completed the Phase
- c. SIS
 - i. Not yet started
 - ii. In the phase
 - iii. Completed the Phase

Practical Experience

12 Did you work or are you currently working at a school?

- a. No
- b. Yes + Amount of Months

13 Did you complete your mandatory stay abroad?

- a. No
- b. Yes, in the year

D. Interview Guidelines

The guidelines will be presented as used in the study, i.e., in German.

D.1. Pre-pilot Phase: Expert Interviews

1. Alter/ Dienstzeit ohne Ref / Fächer
2. Welche Merkmale sehen Sie immer wieder in Texten von Lerner*innen?
3. Was sollte man SuS beibringen wenn sie einen Zeitungsbericht schreiben, also worauf muss man achten? Auch Makro Struktur neben Style erfragen

Aufgabe für die SuS geben

4. Für welche Jahrgangsstufe und Schulform wäre die Aufgabe geeignet?

Text geben – Zeit zum Lesen

2. Wer könnte diese Texte geschrieben haben? Jahrgangsstufe? Schulform?
3. Würden Sie die Texte als authentisch für durchschnittliche 10. Klässer betrachten? Wenn nein, welche nicht?
 - a. Authentisch: Typisch für Produkte von Sus?
4. Sortieren sie die Texte nach Authentizität. Und begründen sie ihre Entscheidung.
5. Typische Merkmale
 - a. Länge / Wortzahl
 - b. Zahl und Qualität der sprachlichen Fehler
 - i. Rechtschreibung
 - ii. Lexiko-Grammatik
 - iii. Syntax
 - iv. Ausdruck
 - c. Kohärenz
 - d. Textsortenspezifität
 - i. Headline ^ Lead ^ Lead development
 - ii. Angemessenheit des Registers
 - e. Qualität des Inhalts
 - f. Aufgabenerfüllung
 - i. Vollständigkeit
6. Präsentationsformat: Digital / Handschriftlich
7. Multimodalität

D.2. Pre-pilot Phase: PSETs Interviews

Demographische Daten

Alter

Geschlecht [w] [m] [d]

Studienfächer

Bachelor/Master; Fachsemester

Erfahrung an Schule/ mit Schüler*innen (Praktika, Nebenjobs etc.). Wenn ja, welche und ca. wie lang?

Vorbereitung: Lesen der Aufgabenstellung; Hinweis 8. Klasse

Fragen?

Stufe 1 – Eindrücke der Texte

- 1) Könnten das Texte aus der Schule sein, also geschrieben von Schülerinnen und Schülern?
- 2) Welcher der beiden Texte ist typischer für Schüler*innen der 9. Klasse. Warum?

Stufe 2a – Begründete Entscheidung Pärchen A,B,C,D (nur dichotome Pärchen da)

- 1) Eine der beiden Texte ist konstruiert und der andere ist so 1:1 von Schülerinnen und Schülern übernommen worden. Können sie entscheiden welcher konstruiert und welcher von Schüler*innen ist und die Entscheidung begründen?

Stufe 2b – Nicht mehr zwingend konträre Pärchen E,F,G (1 mal zwei konstruierte Texte)

- 1) Die Texte können entweder konstruiert oder 1:1 von Schüler*innen produziert worden sein. Können sie die Texte den Kategorien zuordnen und ihre Entscheidung begründen?

Wie schätzen sie sich ein?

Überrascht von Ergebnissen?

Wo Schwierigkeiten?

D.3. Pilot Survey

Einstiegsfrage

1. Wie war es für sie? Erzählen Sie.

Stimulated Recall

Anweisung: Es geht um Ihre Wahrnehmung in dem Moment, also was Sie in dem Moment gedacht haben. Es geht nicht um eine Reflektion/Erklärung ihres Verhaltens

2. In dem Moment, was haben sie gedacht/ wie haben Sie das wahrgenommen/ wie haben sie das aufgenommen?
3. Als sie geantwortet haben, was haben Sie gedacht?
4. Wussten Sie nach der Äußerung, was hier das Problem war?
5. Können Sie bitte Ihre Reaktion hier erläutern?
6. Wissen Sie, worauf hier auf einer „didaktisch/fachlichen“ Ebene abgezielt wurde?

Leitfaden - Nachbefragung

Einleitung

1. Welche Erfahrungen haben Sie bereits mit Rollenspielen gemacht? (Haben Sie bereits an der anderen Simulation teilgenommen?)

Authentizität

2. Wie haben Sie die Situation im Rollenspiel erlebt? Waren die Situationen so, wie Sie diese erwartet hätten?
3. Waren die Situationen für Sie authentisch? Warum? (Ähnlich wie in der Schule? Beeinflussung durch Kamera, Raum, Schauspielerinnen)
4. Haben sich die Schauspielerinnen für Sie authentisch verhalten?
 - Aussehen/Kleidung
 - Auftreten/Verhalten
 - kontinuierlich in der Rolle geblieben
 - wirkte wie eine natürliche Gesprächspartnerin

Überleitung Vorbereitung

5. Wie haben Sie sich mit dem Material auf die Situationen vorbereitet?
6. Fanden Sie das Material (Texte, Aufgabe, Reihenplanung, Beispiel) authentisch? Warum (nicht)?
7. Waren genügend Informationen und Zeit vorhanden, um sich angemessen auf die Situationen vorzubereiten? (Welche Informationen haben Ihnen gefehlt?) War der Modelltext zu ablenkend?
8. Sie haben auf dem Text mehrere Anmerkungen gemacht, die Sie im Gespräch nicht aufgegriffen haben. Warum?
9. Warum haben Sie das Gespräch auf Deutsch/Englisch begonnen? Warum mich vorher gefragt? /Sprache der Aufgabe relevant?

Überleitung Validität

10. Inwiefern wurden mit dem Rollenspiel Ihrer Meinung nach Kompetenzen erfragt, die für den Beruf als Lehrkraft notwendig sind? – Welche Kompetenzen, spezifisch für das Unterrichtsfach Englisch

11. Inwiefern halten sie den Fokus auf Feedback in so einem Rollenspiel für geeignet? Wichtig für den Beruf als Englischlehrkraft?

12. Haben Sie den Eindruck, dass Sie Ihre Fähigkeiten während der Situationen gut zeigen konnten? Warum? (Unwohl sein, Nervosität)

13. Was haben Sie bei dem Rollenspiel gelernt und können für Ihren weiteren Weg als Lehrkraft mitnehmen? – spezifisch auch für das Fach Englisch

Überleitung RobS als Prüfung

14. Wie authentisch war das Rollenspiel im Vergleich zu anderen Prüfungsformaten (Klausuren oder mündliche Prüfungen)? Unterrichtsbesuche?

15. Inwieweit können Sie sich solche Rollenspiele als Prüfungsformat im Lehramtsstudium vorstellen? – Spezifisch für Englisch / Zu welchem Zeitpunkt im Studium?

- Fairness,
- Praxisrelevanz,
- Vergleichbarkeit/Objektivität
- Praktikabel
- Sollten solche Rollenspiele benotet werden? Warum?
- Inwieweit sie sich Rollenspiele als Übungsformat vorstellen können?

16. Was für eine Art Feedback hätten Sie sich gewünscht?

17. Angenommen Sie hätten sich auf die Situationen als benotete Prüfung vorbereiten müssen – Wie hätten Sie sich vorbereitet?

18. Haben Sie noch Fragen?

D.4. Main Survey

Einstieg

1. Wie war es für Sie? Erzählen Sie.

Authentizität

2. Inwiefern war das für sie eine authentische (in der Schule so vorzufindende) Situation? Wie haben Sie die Situation im Rollenspiel erlebt?

3. Inwiefern haben sich die Schauspielerinnen Ihrer Meinung nach authentisch verhalten?

- kontinuierlich in der Rolle geblieben
- wirkte wie eine natürliche Gesprächspartnerin/ wie 9. Klässlerin

Validität

4. Inwiefern wurden mit dem Rollenspiel Ihrer Meinung nach Kompetenzen erfragt, die für den Beruf als Lehrkraft notwendig sind? – Welche Kompetenzen, spezifisch für das Unterrichtsfach Englisch

Überleitung RobS als Prüfung

5. Inwieweit können Sie sich solche Rollenspiele als Prüfungsformat im Lehramtsstudium vorstellen? – Spezifisch für Englisch.

- Fairness/Praxisrelevanz/Vergleichbarkeit/Objektivität/Praktikabel

6. Wie authentisch war das Rollenspiel im Vergleich zu anderen Prüfungsformaten (Klausuren oder mündliche Prüfungen)? Unterrichtsbesuche?

7. Haben Sie noch Fragen?

D.5. Expert Interview

The guideline for the expert interview will be present in the form of the relevant presentation slides used. The first slide includes the questions for the first part of the interview, the remaining slides show the scoring instrument of that time and the questions used to discuss it.



Questions

- 1) How would you evaluate the performance?
- 2) What grade would you assign the performance?

American Grading System	A	A-	B+	B	B-	C+	C	C-	D+	D	F
German Higher Ed. Grades	1,0	1,3	1,7	2,0	2,3	2,7	3,0	3,3	3,7	4,0	5,0

At the end, you will get the chance to adapt your grades and statements.



General aspects of feedback I

Points	Criterion	Explanation
3	Self-evaluation	TP asks the learner for their own evaluation of the performance e.g. writing of the text.
2	Positive feedback	TP mentions aspects that the learner has achieved successfully.
3	Learning objectives (Feed Up)	TP clearly states the learning objectives.
1	Feedback (Feed Back)	TP provides an evaluation of the level of performance against the learning objective.
3	Next Steps (Feed Forward)	TP shows how the learning objectives can be achieved.
1	Task	TP addresses aspects of the text.
2	Process	TP addresses aspects of the process of solving the task.
3	Self-regulation	TP addresses on a metacognitive level the use of different strategies.
1	Self	TP does not address the learner's personal level.
19		

1) Overall impression? 2) Particular attention? 3) Missing? 4) Redundant? 5) Dispensable? 6) Important?





General aspects of feedback II

	Dialog-orientation	Quality of feed back	Quality of feed forward
3	TP always involves the learner in the feedback conversation, e.g., through questions or prompts.	TP provides feedback that is always accurate, clear and on a descriptive level.	TP provides calls to action that are always positive, unambiguous and specific.
2	TP often involves the learner in the feedback conversation, e.g., through questions or prompts.	TP provides feedback that is often accurate, clear and on a descriptive level.	TP provides calls to action that are often positive, unambiguous and specific.
1	TP sometimes involves the learner in the feedback conversation, e.g., through questions or prompts.	TP provides feedback that is sometimes accurate, clear and on a descriptive level.	TP provides calls to action that are sometimes positive, unambiguous and specific.
0	TP never involves learners in the feedback conversation, e.g., through questions or prompts.	TP never provides feedback that is accurate, clear and on a descriptive level.	TP provides calls to action that are never positive, unambiguous and specific.
9			



1) Overall impression? 2) Particular attention? 3) Missing? 4) Redundant? 5) Dispensable? 6) Important?

Adaptivity

	Ensuring understanding / Knowledge-related adaptation	Language-related adaptation	Adaptive error correction
3	TP always assures themselves of learner's understanding and adapts to their needs, e.g., by giving alternative answers.	TP attempts to conduct the conversation in the foreign language (e.g. English) AND adapts to the learner's foreign language level in the foreign language AND continues the rest of the conversation in German if asked to do so. TP asks learners *in English* which language they prefer for the feedback conversation. TP, at the start, offers *In English* to switch to German at any time upon request.	TP provides opportunities for learner self-correction through implicit and indirect feedback, followed by more explicit and direct feedback.
2	TP often assures themselves of learner's understanding and adapts to their needs, e.g., by giving alternative answers.	TP asks learners *in German* which language they prefer for the feedback conversation. TP offers *in German* to switch languages at any time upon learner's request.	TP provides opportunities for learner self-correction, followed by explicit and direct feedback. (No implicit and indirect feedback)
1	TP sometimes assures themselves of learner's understanding and adapts to their needs, e.g., by giving alternative answers.	TP attempts to conduct the conversation in the foreign language (e.g. English) AND adapts to the learner's foreign language level in the foreign language BUT does *not* continue the rest of the conversation in German when asked. TP starts the conversation in German and gives reasons for that.	TP provides implicit and indirect feedback, followed by explicit and direct feedback. (No self-correction possible)
0	TP never assures themselves of learner's understanding and does not adapt to their needs, e.g., by giving alternative answers.	TP starts and conducts the conversation in German without giving reasons. (Does not show adaptive skills). TP attempts to conduct the conversation in the foreign language BUT never adapts to the learner's foreign language level in the foreign language AND never continues the conversation in German when asked.	TP provides only explicit and direct feedback.



Writing competences I

	Criterion	Explanation
3	Prioritization of content/structure	TP first provides feedback on content or structural features of the text.
3	Appropriate amount of information	TP provides an appropriate amount of information to improve the text's quality (Depending on the length of the conversation).
1	Language accuracy	TP refers to aspects of language accuracy e.g. forming of simple past.
2	Language appropriacy	TP refers to aspects of language appropriacy e.g. inappropriate use of tenses.
2	Content	TP refers to aspects of content e.g. missing information.
2	Structure	TP refers to aspects of structure e.g. disregard of genre conventions.
3	Process	TP refers to aspects of process e.g. missing revision tracks.
16		

1) Overall impression? 2) Particular attention? 3) Missing? 4) Redundant? 5) Dispensable? 6) Important?



Writing competences II

	Process orientation	Error pattern	Target-oriented error selection	Elaboration
3	TP clearly states that the text needs to be revised using the feedback.	TP always and exclusively addresses error patterns in the feedback conversation.	TP always selects errors and/or patterns and relates them to the learning objectives.	TP always provides elaborated feedback.
2	TP advises the learner to revise the text again.	TP often addresses error patterns in the feedback conversation. (Single errors are only marginally addressed.)	TP often selects errors and/or patterns and relates them to the learning objectives. (Errors that are not directly related to the learning objectives can be discussed marginally.)	TP often provides elaborated feedback. In some instances only KCR is provided.
1	TP advises the learner to use the feedback for future tasks.	TP sometimes addresses error patterns in the feedback conversation. Individual errors are addressed equally.	TP sometimes selects errors and/or patterns and relates them to the learning objectives. (Errors and/or patterns that are not directly related to the learning objectives are also discussed equally.)	TP sometimes provides elaborated feedback. KCR and KR dominate.
0	TP does not state that the feedback needs to be used for revision or for future tasks.	TP does not address errors patterns. Only individual errors are addressed.	TP never select errors and/or patterns and does not relate them to the learning objectives.	TP only provides KR or KCR without elaboration.

1) Overall impression? 2) Particular attention? 3) Missing? 4) Redundant? 5) Dispensable? 6) Important?



Professional accuracy

	Linguistic accuracy	Pedagogical accuracy
3	TP always uses the foreign language correctly as medium and object of learning. (Limitation: English must be spoken first).	TP is always correct in (subject-)learning specific aspects e.g., learning strategies.
2	TP often uses the foreign language correctly as medium and object of learning only a few individual exceptions.	TP is often correct in (subject-) learning specific aspects, e.g. learning strategies.
1	TP shows noticeable mistakes/errors in the foreign language, but not necessarily hindering learning.	TP shows noticeable mistakes/errors in (subject-) learning specific aspects, e.g. learning strategies, but not necessarily hindering learning.
0	TP shows grave mistakes/errors in the foreign language that might hinder learning.	TP shows grave mistakes/errors in (subject-) learning specific aspects that hinder learning. e.g., learning strategies.
6		



1) Overall impression? 2) Particular attention? 3) Missing? 4) Redundant? 5) Dispensable? 6) Important?



Overview

	General aspects of feedback	Adaptivity	Writing competences	Professional accuracy
Points	28	9	28	6
Sum: 71				





Questions

1. What is your overall impression of the scoring system?
2. Did you notice anything special in particular, which you would like to address?
3. In your opinion: Are any aspects missing?
4. In your opinion: Are there any redundant aspects?
5. In your opinion: Are there any dispensable aspects?
6. In your opinion: Which aspects would you consider to be of particular importance?



E. Transcription Guidelines

Transcription guidelines are translated, but the German examples kept as most interviews are also conducted in German.

E.1. Follow-Up Interviews and Simulations (Main Survey)

Extended content-semantic transcription (based on Dresing & Pehl, 2018, pp. 21–23)

1. The interview is transcribed verbatim, i.e., not phonetically or summarizing.
2. Word slurs are approximated to written German. "so'n buch" is transcribed as "so ein buch" and "hamma" becomes "haben wir". The sentence form is retained, even if it contains syntactical errors, e.g., : "Bin ich nach Kaufhaus gegangen."
3. Dialects are translated into High German as accurately as possible. If no clear translation is possible, the dialect is retained, e.g: "ich gehe heuer auf das oktoberfest".
4. Colloquial particles such as "gell, gelle, ne" are transcribed.
5. Stuttering, stopped words and word doublings are transcribed.
- 6 No punctuation.
7. Reception signals such as "hm, hmhm, mh, aha, yes, exactly" are transcribed.
8. Pauses of 1 second on are marked by (.), of 2 seconds on (..), of 3 seconds (...). Of 4 seconds, the length in seconds is noted in brackets (4 sec.).
9. Particularly emphasized words or utterances are marked by VERSALS. Questions are marked by the stress (capital letters) of syllables.
10. Emotional non-verbal expressions of the interviewee and the interviewer which interviewer that support or clarify the statement (such as laughing or sighing) are noted in brackets when used.
11. Unintelligible words are marked with "(unv.)" (inaudible). Longer unintelligible passages are, if possible, marked with the cause "(unv., microphone noise)". If a statement is suspected, the passage is marked with a question mark in brackets, e.g., "(axe?)". Incomprehensible passages are marked with a time stamp, if no further time marker is set within one minute.
12. Overlapping speech acts are marked by square brackets [...].
13. Abbreviations, spellings and numbers are transcribed literally: "das wird mit ed geschrieben". "five hundred trees were planted". If the abbreviations, spellings or numbers are emphasized, they are written in VERSALS.

E.2. Stakeholder and Expert Interviews

Content-semantic transcription for expert interviews (cf. guideline interviews)
(based on Dresing & Pehl, 2018, pp. 21–22)

1. The interview is transcribed semantically.
2. The abbreviations are "I" for interviewer, "S" for student and "E" for expert.
3. Video excerpts are marked as follows: "Video von TP[NR] wird abgespielt"
("Video of TP[No] is playing")
4. If there are interruptions, the time marker from the video is inserted.

"Interruption at [h:mm:ss]

E:

I:

Continuation of the video at [h:mm:ss]"

F. Coding Manuals

F.1. Stimulated Recall

Code		Explanation	Example
Revision			
	"Muss ich den Text nochmal neu schreiben?" ("Do I have to write the text again?")		
	Learner's motivation & cognition		Participants discuss the learner's motivation and cognition.
	Deficient		TP: Ähm, dass sie sich ein bisschen jetzt vor Arbeit irgendwie drücken will. Also, wenn es jetzt schon eine Herausforderung darstellt, den Text jetzt nochmal zu überarbeiten oder da nur abzuschreiben, sagt sie jetzt ja eigentlich dann, ja, habe ich mich so ein bisschen gefragt so: „Wieso hast du es dann eigentlich abgegeben, ne? Wenn du nicht bereit bist, jetzt da irgendwie noch was dran zu ändern oder so. (IEH20)
	Positive		TP: Hm. Ja muss. Ähm. Also ich habe mir halt den Verlaufsplan der Unterrichtsreihe angeguckt und da stand ja drin, dass die halt als nächstes / in der nächsten Stunde in der Klasse zusammen gemacht wird. Deshalb muss sie das nicht jetzt machen, zu dem Zeitpunkt. Ähm, genau. Und halt auch wieder, sie zeigt Arbeitsbereitschaft. Also sie ist willig da dran zu arbeiten. (MJZ08)
	Development of writing competence		Participants discuss the necessity of supporting the learner in developing their writing competence.
	Confirming the necessity		TP: Ich habe gedacht natürlich muss man einen Text mit so vielen Fehlern nochmal neu schreiben gerade ähm mit diesem ähm das man etwas üben kann (RMS24)

Code				Explanation	Example
				that rewriting is necessary.	
			No necessity	Participants discuss the necessity of supporting the learner in developing their writing competence while confirming that rewriting is not necessary.	TP: Da habe ich mir gedacht, dass es nicht so viel Sinn machen würde, wenn sie es / Also, was ich alles dran geschrieben habe, das war ja nicht / Das war ja nur so grob, was so was falsch war, aber alleine diese ganzen Sätze nochmal neu formulieren das hätte sie auch alleine hätte sie das nicht hinbekommen. (AYW07)
			General discussion	Participants discuss the necessity of supporting the learner in developing their writing competence in general.	TP: Eine gute, authentische Frage, ähm, weil das ist immer die Angst von Schülern, dass das ganz schlecht ist und dann das alles immer neu machen muss, aber, ähm, meine Antwort war ja auch: „Ja“, ähm, aber es ist ja ein Entwurf und deswegen ist das / diese Angst muss man finde ich Schülern dann nehmen, um / weil es nicht schlecht ist etwas neu zu schreiben und es muss ja auch nicht alles neu formuliert werden, aber nur halt verbessert werden. (GTM21)
Self-regulation / Feed Forward					
		“Wie kann ich das üben?” (“How can I practice that?”)			
		Participants’ competence	Participants discuss how they did not know how to answer this question.	TP: Ähm. Ja, da geht es ja sozusagen so ein bisschen darum irgendwie Lernstrategien irgendwie aufzubauen und wenn sie das jetzt noch nicht hat in der neunten Klasse / also da wusste ich auch nicht wirklich, wie ich darauf jetzt reagieren soll, weil, ja. (IEH20)	
		Options for support	Participants discuss the necessity to provide possible actions for the learner to support them.	TP: Dass ich dann der Schülerin einige Handlungsmöglichkeiten aufzeigen kann. Also, wie sie tatsächlich diese Vokabeln dann am besten lernen kann. Vielleicht verschiedene Methoden. Einem fehlt da natürlich / also ich kannte sie ja jetzt nicht in dieser Simulation, aber teilweise kann man ja dann schon mal auf, ähm, vorheriges Verhalten oder generell die, ähm, wie sie lernt darauf zurückgreifen und dann schauen, okay, was passt am besten auf diese Schülerin. (ESG19)	
		Learner’s motivation / cognition	Participants address the learners’ motivation and	TP: Das ist, das ehrt sie, dass sie das üben möchte, aber an der Stelle wusste ich nicht wie man üben soll bzw. was sie genau meint (RMS24)	

Code		Explanation	Example
		their reflective abilities.	
Adaptation of content			
	"Was habe ich hier falsch gemacht?" ("What is wrong here?")		
	Learner's motivation/cognition.	Participants address the learner's motivation or cognition.	TP: Was sie falsch gemacht hat in dem Text. Weil ich sie jetzt nicht gleich komplett demotivieren möchte oder so. Weil da waren schon viele Fehler drin, aber das wollte ich nicht so deutlich sagen. (MJZ08)
	Participants' competence	Participants discuss how they did not know how to answer this question.	TP: Äh. Ich war generell so ein bisschen überfordert, wie ich mich ausdrücken soll. Was soll ich sagen? Dass sie jetzt nicht so äh verunsichert ist oder so, weil das war wirklich nicht so ein guter Text. Und dann wo sie gefragt hat, was genau der Fehler war konnte ich es ehrlich gesagt / Ich mache das auch erstmal nur so das vom Gefühl her und genau konnte ich jetzt auch nicht sagen, was falsch war. Außer ja, dass der Satz nicht in der richtigen Reihenfolge war oder so.(AYW07)
	Ist das hier richtig? ("Is this correct here?")		
	Learner's motivation/cognition.	Participants address the learner's motivation or cognition.	TP: Ähm. So als ob sie sich jetzt sehr unsicher bezüglich der Vergangenheitsform. Ähm. Also, ja doch, eigentlich schon. Hm. (...) Ja, ich habe das in dem Moment halt erstmal noch ein bisschen ignoriert und einfach ein bisschen Sicherheit nochmal gegeben: ‚Das hast du an der Stelle richtig gemacht.‘ Ja, ich glaube nachher haben wir dann nochmal kurz drüber gesprochen dann und über irgendwelche Zettel, um die unregelmäßigen Formen zu lernen und so was. Ja, ich glaube das kommt später. (SAS05)
	Participants' competence	Participants discuss how they reacted.	TP: Da habe ich mir wieder gedacht, jetzt muss ich was positives sagen. [lachen] Also, obwohl der Satz gar nicht richtig war, ja. I: Und warum haben Sie dann gesagt, dass es richtig war? TP: Hab ich das gesagt danach? I: Wir können nochmal Reinhören. [...] TP: [lachen] Oh. [lachen] Eigentlich war es nicht okay da, aber ich dachte, ich sag ihr das jetzt einfach. (AYW07)

Code		Explanation	Example
	"Das habe ich nicht verstanden." (("I did not understand that."))		
	Learner's motivation/cognition.	Participants address the learner's motivation or cognition.	TP: Ja, ist jetzt schwierig. Ähm, sie hat wahrscheinlich gar nicht verstanden, was überhaupt eine „conjunction“ ist und dann hätte ich ihr das eigentlich nochmal erklären müssen (IEH20)
	Participants' competence	Participants discuss their actions.	TP: Ja, da ist mir aufgefallen, dass ich da eine Sache nicht verbessert habe und das die Schülerin ähm schon vorher geahnt hat, dass sie da irgendwas nicht richtig gemacht hat und dass es im Englischen die Anführungsstriche anders ähm verwendet werden als im Deutschen und dass sie das selber weiß, aber die Regel nicht genau wusste, genau und ähm da dachte ich mir, äh ja schade dass mir das äh dass ich das nicht, also es ist mir aufgefallen, aber da ich das alles auf die Schnelle gemacht habe, dachte ich mir, schade dass ich die Zeit nicht dazu hatte das schnell noch zu verbessern. (HKH07)
Diagnosis			
	Deep-level: Content		
	"Ist vom Inhalt her alles drin?" (("Is everything included with regards to content?"))		
	Learner's motivation/cognition.	Participants address the learner's motivation or cognition.	TP: Ähm. Da habe ich mir dann gedacht als sie nach dem Inhalt gefragt hat so, dass es jetzt wahrscheinlich auch mit den sprachlichen Aspekten genug ist. Ähm. Ich habe mich dann in dem Moment gefragt, ob das in die Richtung geht, so, dass ähm, dass so langsam von der Zeit her ausreicht, dass es ihr zu lang wurde. Ähm. Und dann sagte sie so, dass wir dann zum Inhalt gehen, damit wir schneller fertig sind dann. Ähm. Oder, ob ihr in dem Moment halt der Inhalt so wichtig war, dass sie äh eben auch unbedingt wissen wollte, ob es da jetzt inhaltlich gestimmt hat. (VJD15)
	Participants' lack of knowledge	Participants discuss how they did not know how to answer this question.	TP: Ja da hatte ich keine Ahnung, da hatte ich echt keine Ahnung. Da hätte ich mich viel mehr, also da hätte ich ein paar Minuten länger äh haben müssen äh, damit ich das vergleichen kann. (HKH07)

Code			Explanation	Example
	Deep-level: Structure			
		"Ist das vom Aufbau her in Ordnung?" ("Is the structure okay?")		
		Learner's motivation/cognition.	Participants address the learner's motivation or cognition.	TP: Hm. Ja, da habe ich mir überlegt: „Wie kann ich ihr eine hilfreiche Antwort geben, die sie halt auch nicht entmutigt?“ Weil es war ja eigentlich sehr viel / war relativ viel verkehrt. Aber, ähm, ich wollte sie halt nicht so entmutigen, weil sie halt wirklich auf mich zugekommen ist. Hätte sie ja nicht machen müssen. (MJZ08)
		Text Structure	Participants address the structure of the text.	TP: Ja das ist mir wieder eingefallen, dass ich beim Lesen gedacht habe, wer ist das We? Wenn man einen Zeitungsartikel schreibt und dann sagt so wir haben das und das gemacht, dann sollte irgendwann am Anfang klargestellt werden, dass wir eine Schülergruppe sind und ähm insgesamt war es natürlich ein sehr kurzer Text, aber Schülertexte sind häufig kürzer als sie sein müssen, also die werden nie einen Artikel der zwei Seiten in der Zeit einnimmt schreiben über, für ihre Hausaufgabe über ein Projekt. Also abgesehen davon, dass der Text kurz war, fand ich, dass er dem Genre genug entsprach. Ich habe überlegt, naja es gab eine Überschrift, es gab ähm die wichtigen Informationen waren untergebracht und dann als sie meinte so ja es ist aber ansonsten entspricht es dem was es sein soll, ist mir dieses mit dem We ingefallen und das war das einzige was ich in dem Moment noch ganz wichtig fand äh zu ähm korrigieren oder noch anzumerken. Man muss ja auch eigentlich, es war ja sowieso ein sehr langes Feedback, man muss ja auch eigentlich gucken, dass das Feedback kurz genug ist, damit die Sachen auch umgesetzt werden können und man nicht einen zu langen Katalog hat an Sachen, die noch äh, ja zu kommentieren sind. (RMS24)
Adaptation of language				
	"Good morning"			
		Language	Participants discuss the topic of language.	
		Conversation is supposed to take place in English	Participants state that they thought that the	TP: Ja, das sie erwartet, dass das Feedback jetzt auf Englisch ist. (IEH20)

Code			Explanation	Example
			conversation is supposed to take place in English.	
		Insecurity with regard to language but choose German	Participants are insecure which language to use and decide for German.	TP: Ich habe mich gewundert, dass sie mich auf Englisch angesprochen hat. Und dann war ich so am überlegen, soll ich jetzt auf Englisch oder nicht. Und dann habe ich mir gedacht, ähm „Wir machen das lieber auf Deutsch.“ Das habe ich mir da gedacht. (AYW07)
		Surprise	Participants are surprised about the learner speaking English.	TP: Ich war überrascht davon, dass das schon der Anfang äh der Situation generell war, darauf war ich nicht ganz vorbereitet ähm und dachte dann ah wenn das schon der Anfang der Situation ist dann äh müsste ich mich eigentlich anders platzieren, weil ich saß ja an diesem Tisch und das war jetzt irgendwie nicht so authentisch für mich. (CAH06)
		Test Situation	Participants utter thoughts about the test situation.	TP: Eigentlich nicht so viel, weil ich es gar nicht erwartet habe, dass sie überhaupt reinkommt. Ich habe halt gedacht das findet irgendwie zwischen einer Tür oder auf dem Flur irgendwo statt. Aber jetzt im Nachhinein ist es ja schön, dass sie gleich Englisch gesprochen hat (MJZ08)
		"Können Sie das auf Deutsch sagen?" / "Can you say that in German?"		
		Learner's language competence	Participants consider the learner's needs with regard to language.	TP: Dass ich das vielleicht etwas zu kompliziert ausgedrückt habe (lachen). Ähm, ja, ich ne/ Ich glaube, das war mein Fehler. Das war sehr kompliziert vielleicht. Und dementsprechend habe ich dann auch / Da bin ich einmal zum Deutschen gewechselt. Und habe das versucht nochmal auf Deutsch etwas unkomplizierter zu sagen (SAS05)

F.2. Guideline Interview

Authenticity**Actors**

Code		Explanation	Example
Behavior			
	Rather Authentic	Behavior of actors was considered generally authentic	
	Authentic but with limitations (limitations coded individually)	Behavior was perceived generally authentic, but limiting reasons were given	
	Communicative behavior (verbal)	The communicative behavior was perceived as limiting the actors' authenticity	[0:03:00] TP: Ja, schon, aber ich glaube sie hat mich einmal auch auf die Struktur von diesem äh Artikel hingewiesen, wo ich gerade irgendwie nicht mehr so ganz war, weil ich mit anderen Dingen beschäftigt war und das wäre natürlich vielleicht unauthentisch. Außer die Schülerin ist sehr motiviert und sagt auch hier.
	no limitations		
	General authenticity	The authenticity of the actors is addressed positively without any specification	[0:01:47] TP: Ich fand, sie hat sich sehr authentisch verhalten. Sie hat auch immer wieder nachgefragt, äh was ich genau damit meine, also wirklich auch ja dem Kenntnisstand eines Schülers zuzuordnen so zusagen. Ähm ja ich habe am Anfang vor allem auch erstmal Antworten auf Englisch gegeben, wo ich war mir nicht mehr sicher, auf welcher Sprache wir das machen sollten. Und dann hat sie mich auf ja immer hingewiesen, dass sie es gerne auf Deutsch nochmal erklärt haben ja wollen würde. Und ich habe es dann gecheckt (lachen) und dann habe ich es auch auf Deutsch erklärt. (ACG22: 25 - 25)
	Highlights		
	Body language (non-verbal)	The authenticity of the actors is addressed positively and the body language is mentioned specifically.	[0:03:31] TP: Ähm das ist jetzt die Frage. Ähm also einerseits an ihrem schauspielerischen Leistung, also sie hat auch äh vielleicht auch den Blick zu mir auch immer gesucht äh vielleicht wie Schüler das machen würden zu Lehrern, wenn sie da die Antwort äh ja, dass die denken, da kommt die Antwort vom Lehrer jetzt und dann gucken die sehr erwartungsvoll. Das hat sie, glaube ich, sehr gut gemacht. Ähm ja und auch ja einfach die richtigen Sachen gesagt, also sie hat auch gesagt ‚Ja, ich muss mich da verbessern‘ und ‚Ja ja ja‘, also die Rolle der

Code			Explanation	Example
				Schülerin hat sie sehr gut eingenommen. (AOS13: 33 - 33)
		Communicative behavior (verbal)	The authenticity of the actors is addressed positively and the communicative behavior is mentioned specifically.	[0:00:54] TP: Hm dadurch, dass die Schülerin nach der Unterrichtsstunde reinkam mit dem Klopfen, erst ein bisschen schüchtern und dann bin ich auf sie zugegangen und dann ging es und sie hat auch selber Fragen gestellt, wie sie sich verbessern könnte und hat die Rolle ganz gut übernommen. Ähm und dass ich dann als Lehrerin auch das Gespräch geführt habe, also die Leitung übernommen habe. (BJP10: 22 - 25)
	Rather Inauthentic		Behavior of actors was considered generally inauthentic	
		Communicative behavior (verbal)	Actors' communicative behavior is specified as being inauthentic.	[0:01:47] TP: Also es kam schon so rüber als würde sie gar nichts wissen und ich kenne das auch von, äh, meinem Nachhilfeschüler, dass wenn ich mit dem, ich probiere mit dem hauptsächlich auf Englisch zu reden, aber bei vielen Sachen fragt er halt nach, ob man das auf Deutsch erklären kann und da habe ich das jetzt auch in dem Kontext sehr stark gemerkt, dass das wahrscheinlich dann bei Schülern in dem Alter dann genau so wäre in der Schule und vor allem, was mir sehr stark aufgefallen ist, dass ich nicht wusste, wie man in einer solchen Situation dann handelt, ob man dann das Englische einfach strikt durchzieht oder ob man dann in, in Englisch wieder rüberwechselt und, ähm, aber vom / das einzige was mir stark noch aufgefallen ist, dass es glaube ich, was ich auch schon eben gesagt hatte, dass dieses dumm stellen das kam mir zu stark rüber, dass sie wirklich stur fast alles, was ich gesagt habe, nochmal hinterfragt hat und das kam mehr so rüber, dass es war, um mich zu ärgern, sag ich mal, ärgern in Anführungsstrichen, als dass sie wirklich interessiert daran war mehr zu lernen. (BJL19: 25)
	Unclear		Statements could not be put into dichotomous categories because of, e.g., differentiated answers, misunderstanding of question	
		Communicative behavior (verbal)	Unclear statement revolved around actors' communicative behavior	[0:03:32] TP: Hm (...) also einmal in dem, das sie zwischendurch auch Rückfragen gestellt hat so: „Ja, wie hätte ich das denn besser machen können?“ Also vor allem dann nicht einfach nur alles abzunicken, okay. Da habe ich mich aber auch

Code	Explanation	Example
		zwischen durch gefragt, inwiefern würde ich, hätte ich das als Schülerin gemacht, dann auch so zu fragen: ‚Was würden Sie denn sagen?‘ oder ‚Wie hätte ich das denn besser machen können?‘ Ich habe eher das Gefühl, dass Schülerinnen und Schüler meistens eher so dann so sitzen und dann auch so bisschen einfach so empfangen anstatt dann auch in dem Feedback nochmal Fragen zu stellen. Aber von dem her wie sie dann selbst überlegt hat und wie sie das dann auch ähm angenommen hat, hatte ich das Gefühl/ Das wirkte auch so wie jemand, der halt wirklich nicht genau weiß: ‚Habe ich jetzt einen guten Text geschrieben, oder nicht? Und ich brauche das gerade wirklich als Feedback von Ihnen.‘ Genau, aber ich habe mich zwischen durch so gefragt: ‚Okay, inwiefern hätte man wirklich den Mut dann auch nochmal zu sagen ähm an die Lehrkraft selbst diese aktiven Rückfragen auch zu stellen?‘ [0:04:37] I: Hm du hast gerade gesagt, das hast du dich in der Situation gefragt. Hat dich das irgendwie irritiert in der Situation? Hat dich das irgendwie rausgebracht? [0:04:43] TP: Einmal hat es mich kurz ein bisschen irritiert ähm als es darum ging wie man die Überschrift hätte formulieren können und dann hat sie gefragt: ‚Wie hätte ich das denn machen können?‘ Dadurch, dass ich dann auch selbst ja gar nicht in dem Thema drin war, war es ein bisschen so ja muss ich mir jetzt auch ein bisschen aus den Fingern saugen, wie man jetzt eine gute Überschrift für einen Zeitungsartikel finden kann. Und dann habe ich ihr auch einfach wieder die Rückfrage gestellt: ‚Ja, was würdest du denn sagen?‘, weil theoretisch ist, sie auch nicht jetzt in der Situation, aber wäre sie ja auch in dem Thema drin. Ähm das hat mich einmal ganz kurz rausgebracht so diese aktive/ [0:05:17] I: Aber hat dich das rausgebracht in dem Sinne, dass du ein bisschen gestockt hast und überlegt hast, oder weil die Frage für dich ähm nicht eine typische Frage von Schülerinnen und Schülerin ist? [0:05:28] TP: Beides ein bisschen, also es gab zwischen durch Situationen, wo ich mich gefragt habe, also dann auch reflektiert habe in dem Moment selbst. Ähm hätte ich damit jetzt gerechnet, obwohl man gleichzeitig nicht so richtig mit was rechnen kann, weil wenn man dann in so einem Gespräch sitzt, glaube ich, kann gefühlt alles und Nichts

Code		Explanation	Example
			passieren. Deswegen ist auch die Frage: ‚Ist das erwartbar, oder nicht?‘, kann, glaube ich, sehr unterschiedlich sein. Ähm genau, aber auch, weil ich für mich selbst drüber nachdenken musste: ‚Okay, was würde ich jetzt sagen?‘ und ‚Wie soll ich oder kann ich jetzt gut darauf reagieren?‘ (EME03_M: 23 - 27)
Natural conversation partner			
	Rather positive	Actors were perceived as natural conversation partners.	
	No limitations	No limitations were mentioned	[0:04:20] I: Inwiefern wirkte sie denn wie eine natürliche Gesprächspartnerin, also wie eine natürliche Neuntklässlerin? [0:04:26] TP: Ähm (lachen) erstmal, dass sie auf Englisch ‚Hallo‘ gesagt hat am Anfang. Ich glaube, das machen viele Schülerinnen und Schüler. Und ähm ja dann einfach, dass sie, ich glaube, sehr motiviert war, mehr über ihren Text zu erfahren und das dann nochmal paraphrasiert hat, was ich so als Feedback gegeben habe. Und das fand ich ja das war so recht authentisch. (ALA02, Pos. 34-35)
	With limitations	Awareness of question aims	[0:02:53] I: Mhm und inwiefern wirkte sie wie eine natürliche Gesprächspartnerin oder wirkte sie in irgendeiner Weise künstlich? [0:03:02] TP: Hm, ja das ist so ein bisschen der Punkt, also ich meine, wenn ich jetzt ähm da ein bisschen in dem Feedbackgespräch bin, ich erkläre ein paar Sachen und dann kommen Fragen wie ähm ja wie kann ich denn die Verben üben ähm natürlich weiß ich dann okay sie fragt mich jetzt (unv.) um zu gucken was ich da jetzt auch als Feedback geben kann ähm, ob das jetzt unnatürlich ist ähm würde ich so direkt jetzt nicht sagen, weil ich mir schon gut vorstellen kann, dass es in der Schule eben auch genau so ablaufen kann. [0:03:30] I: Mhm. [0:03:30] TP: Ja. [0:03:31] I: Also die Aussage an sich würden Sie sagen Sie wussten worauf damit abgezielt wird, aber Sie würden schon sagen, dass das eine Aussage ist, die Schülerinnen und Schüler tätigen würden. [0:03:41] TP: Genau, ja. [0:03:42] I: Mhm. Ähm hat Sie das irgendwie aus der Simulation rausgebracht kurz oder waren Sie trotzdem immer in der Situation drin? [0:03:50] TP: Ne, also rausgebracht hat es mich jetzt nicht ähm, ich wusste jetzt natürlich nicht was noch

Code		Explanation	Example
			auf mich zukommt und inwiefern jetzt auch Fragen auf mich zukommen, ähm aber rausgebracht hat mich das jetzt ähm nicht so wirklich. [0:04:02] I: Wirkte sie denn wie eine natürliche 9.Klässlerin? [0:04:06] TP: Ähm ja also ich muss sagen jetzt in meinem Fall ich studiere ja fürs Berufskolleg heißt mit 9.Klässlern habe ich jetzt äh nicht viel zu tun gehabt, aber so aus der Nachhilfe jetzt in meinem Fall die ich so habe würde ich schon sagen, ja das war schon natürlich, ja. (FCP15: 28 - 37)
	Negative	Actors were not perceived as natural conversation partners	n.a.
	Unclear	It is unclear whether TP perceived the actor as natural or not.	[0:05:59] I: Wirkte sie denn trotzdem oder inwiefern wirkte sie denn dann überhaupt wie eine natürliche Gesprächspartnerin wie eine natürliche Neuntklässlerin? (...) [0:06:07] TP: Hm (...) ich glaube in dem, dass sie selbst halt auch auf die Klassenarbeit zum Beispiel nochmal eingegangen ist und so. Und auch reingekommen ist ne mit dieser Begrüßung, die man dann auch irgendwie hat. Das hat natürlich den, das setzt ja dieses Setting irgendwie, nicht einfach nur so ein man begrüßt sich jetzt als würden wir uns gerade eben in einem außerschulischen Setting begegnen. Und schon eine gewisse Unsicherheit, die sie dann halt auch an gezeigt hat. Von dem her ähm ‚Habe ich das denn richtig gemacht?‘ ‚Wann habe ich das denn nicht so richtig gemacht?‘ und so dann halt auch schon wirklich dieses hm das muss ich jetzt von Ihnen hören. Das kann ich selbst nicht beurteilen. [0:06:43] I: Also so ein bisschen ambivalent das Bild? [0:06:45] TP: Ja, zwischendurch genau. (EME03_M: 28 - 31)
Continuity			
	Rather positive		
	No exceptions	Actors were perceived to consistently have stayed in their role	[0:03:45] I: Ist sie denn auch kontinuierlich in der Rolle der Schülerin geblieben in der Situation eben? [0:03:50] TP: Ich fand schon, ja. (AJB12, Pos. 32-33)
	Exceptions	Exceptions to consistency are addressed.	[0:04:17] I: Ja, okay. Ähm Inwiefern ist die Schauspielerin kontinuierlich in der Rolle der Schülerin geblieben?

Code		Explanation	Example
			[0:04:24] TP: Ähm eigentlich die ganze Zeit, außer zwischendurch da ging irgendwie der Kulli nicht, da meinte sie: ‚Ach alles gut.‘ [0:04:30] I: (lachen) [0:04:31] TP: Ähm aber sonst die ganze Zeit. (BNW18: 38 - 41)
	Rather negative	Actors were not perceived to consistently have stayed in their role	[0:07:38] TP: Hm man hatte das Gefühl, dass sie am Ende ein bisschen erwachsener, also so mehr die Rückfragen kamen, umso mehr würde ich sagen ist sie so ein bisschen aus der Rolle gefallen. Ähm warum sie diese Rückfragen gestellt hat, ob das Teil der Studie war oder ob das geplant war, weiß ich natürlich nicht, aber ähm meistens/ Also ich habe noch nicht so viel Feedback gegeben, aber das, was ich geben konnte im Rahmen meines Praxissemesters hat mir gezeigt, dass ähm sie häufig, also umso älter sie werden, umso weniger Rückfragen kommen eigentlich bei diesen Feedbackrunden, also die kleineren oder ähm jüngeren Schüler innen, die fragen nochmal häufiger auch rein tatsächlich. Umso älter sie sind, umso zurückhaltender werden die, was das angeht. (MJO28: 41 - 41)

Situation

Code		Explanation	Example
Perception of the situation			
	Rather Authentic		
	Authentic but with limitations	Situation is considered to be authentic, but certain limitations are mentioned	
	Limitations		
	Text	Text is mentioned as limiting authenticity.	[0:01:45] TP: Sie war sehr gut da drin, das zu simulieren eine Schülerin den. Also ich weiß nicht, ob sie das einfach von sich aus gut konnte oder gut vorbereitet war, aber sie hat das sehr gut rüber gebracht, das authentisch rüberzubringen vor allem ihre Englischkenntnisse, wie weit sie damit ist, was sie versteht, was sie nicht versteht. Inwiefern der Text authentisch war/ Ich weiß nicht wie weit Schüler in dem Jahrgang sind. Mittlere Schüler ja im Dreier-, Viererbereich in dem Jahrgang wie gut deren Englischkenntnisse sind. Deswegen weiß ich nicht wie sehr authentisch der Text war, weil ich kann das nur mit meinen eigenen Kenntnissen/ Anhand von den Kenntnissen abschätzen und da waren manche Fehler doch schon sehr fraglich, wo ich mir dachte ‚Ist das

Code			Explanation	Example
				wirklich ein Fehler, den eine Neuntklässlerin, die Drei, Vier schreibt, machen würde?' Aber ansonsten war es gut. Ja. (AFS29: 19 - 21)
		Amount of material	The amount of material is mentioned as limiting the situation's authenticity.	[0:01:45] I: Mhm. Inwiefern war das denn eben eine authentische Situation? [0:01:50] TP: Also, wie gesagt die viertel Stunde finde ich zu viel, aber ähm, wenn ich mir vorstelle, jetzt war alles für mich neu gerade, aber wenn ich/ wenn ich im Thema wäre, bräuchte ich auch weniger Zeit von daher finde ich es gut, dass diese viertel Stunde gegeben wird, weil man dann im Thema noch mehr reinkommt, wenn man jetzt schon im Thema drin ist, weil man/ das war jetzt die 7. Glaube ich, die 7. Stunde oder 6. Stunde, wenn man da schon so tief drin ist, dann ist es was anderes. Da braucht man vielleicht auch nur dieser zwei, drei Minuten. [0:02:20] I: [Mhm.] [0:02:20] TP: Um ein vernünftiges, gutes Feedback geben zu können. [0:02:22] I: Hattest du denn das Gefühl, dass die Situation wie sie eben war, dass die sich authentisch angefühlt hat? [0:02:29] TP: Ja schon, schon. Also mit den ganzen Zetteln, die vor einem liegen, das war schon ein bisschen unauthentisch. Ich habe immer so gedacht okay hier, hier muss ich gucken, hier muss ich gucken und man vergisst schnell mal was. [0:02:40] I: Mhm. [0:02:41] TP: Wenn man natürlich total im Thema drin ist, weiß man auch, worauf man achten muss. [0:02:45] I: Mhm. [0:02:46] TP: Aber da, ich finde, dass diese viertel Stunde da geholfen hat, sich dann da reinzudenken. [0:02:51] I: Und jetzt abseits dieser Vorbereitungszeit, sondern einfach nur diese zehn Minuten und das Gespräch mit der Schülerin ähm, hattest du da das Gefühl, dass das authentisch ist? [0:03:00] TP: Ja, schon, aber ich glaube sie hat mich einmal auch auf die Struktur von diesem äh Artikel hingewiesen, wo ich gerade irgendwie nicht mehr so ganz war, weil ich mit anderen Dingen beschäftigt war und das wäre natürlich vielleicht unauthentisch. Außer die Schülerin ist sehr motiviert und sagt auch hier. [0:03:17] I: Mhm, das ist aber Teil der Aufgabestellung, da steht ja explizit drin beurteile die Struktur eines Zeitungsberichtes.

Code			Explanation	Example
				[0:03:21] TP: Ja, genau, genau. [0:03:22] I: Mhm. [0:03:23] TP: Ja, aber ich war irgendwie nicht mehr da. (SJH28: 22 - 37)
		Time	Time is mentioned as a limiting factor for the situation's authenticity.	[0:01:45] I: Mhm. Inwiefern war das denn eben eine authentische Situation? [0:01:50] TP: Also, wie gesagt die viertel Stunde finde ich zu viel, aber ähm, wenn ich mir vorstelle, jetzt war alles für mich neu gerade, aber wenn ich/ wenn ich im Thema wäre, bräuchte ich auch weniger Zeit von daher finde ich es gut, dass diese viertel Stunde gegeben wird, weil man dann im Thema noch mehr reinkommt, wenn man jetzt schon im Thema drin ist, weil man/ das war jetzt die 7. Glaube ich, die 7. Stunde oder 6. Stunde, wenn man da schon so tief drin ist, dann ist es was anderes. Da braucht man vielleicht auch nur dieser zwei, drei Minuten. [0:02:20] I: [Mhm.] [0:02:20] TP: Um ein vernünftiges, gutes Feedback geben zu können. (SJH28: 22 - 25)
		Lack of experience	Lack of experience is mentioned as a limiting factor in evaluating the authenticity.	[0:01:23] TP: Hm war die authentisch? (...) Also grundsätzlich weiß ich ja, dass es solche Situationen in der Schule gibt, dass die Schüler dann kommen und sagen: ‚Hier, können Sie mal drüber gucken? Was kann ich besser machen? Was ist nicht so gut? Können Sie mir Tipps geben?‘ Also in der Hinsicht hielt ich sie für authentisch, aber dadurch, dass ich noch keine Lehrperson bin, noch nicht richtig, ähm ja/ Es ist wahrscheinlich mein Mindset gewesen und nicht die Situation an sich. [0:01:53] I: Was genau mit ihrem Mindset? Was meinen Sie damit? [0:01:56] TP: Ähm ich hatte ich/ Natürlich hatte ich im Kopf, dass ich keine Lehrerin bin, ne. Und ich wollte mich eigentlich wie eine fühlen, weil ich ja anscheinend eine Schülerin vor mir sitzen hatte, aber ich konnte das noch nicht. (ADH08: 23 - 25)
		Actors' age	Actors' age was mentioned as limiting the situation's authenticity.	[0:00:42] TP: Hm schon sehr bis auf natürlich, dass die Dame gerade nicht in dem Alter war, mit dem man normalerweise dann so ein Feedback ähm -gespräch führen würde, aber ähm ich fand es schon relativ authentisch, weil sie halt auch oft so getan hat [0:01:00] I: Hm und ähm hat Sie das irgendwie gestört, dass das Alter nicht stimmte? [0:01:05] TP: Nein. (CCF27: 21 - 23)

Code			Explanation	Example
		Actors' communicative behavior	Actors' communicative behavior was mentioned as limiting the situation's authenticity.	[0:00:49] TP: Hm in dem Part, wo ich dann am Anfang erstmal meine Erkenntnisse von dem Text erklärt habe, war es ziemlich authentisch. Nur, ich glaube, dass Schüler innen in der richtigen Situation dann nicht noch Nachfragen stellen würden, wo sie denn die Fehler genau gemacht hätten, sondern dass sie da eher ein bisschen zurückhaltender sind, aber sonst war es recht nah an der Realität, finde ich. (SLG21_M: 21 - 21)
		Lack of context knowledge	Limited knowledge about the learner and teaching context was mentioned as limiting the situation's authenticity.	[0:00:54] TP: Hm also ich glaube, es war auf jeden Fall authentisch, dass man ähm dass man erstmal authentisches Material vor sich liegen hatte. Das fand ich, was ganz gut gewählt ähm und auch die Schülerin hat eigentlich ganz ganz gut gespielt, wenn man das so sagen kann, also es hat sich wirklich so angefühlt als wäre man gerade beim bei einer Gesprächssituation. Ja genau, das Einzige, was so gefehlt, was man noch, glaube ich, schwer äh nachkonstruieren kann ist, dass man ja das war jetzt eine fremde Person für mich und äh in der Schule hätte ich da natürlich schon mehr Vorwissen, was so die anderen Stunden angeht, wie die Leistung oder wie sich die Schülerin einfach verhält im Unterricht. (ALA02: 21 - 21)
		Lack of work/school context	Lack of school context, e.g., stress was mentioned as limiting the situation's authenticity.	[0:01:34] TP: Hm es war authen/ Also die Aufgabenstellung war natürlich authentisch und der Text erschien mir auch sehr authentisch. Ähm dann fand ich, dass die Schauspielerin das auch ganz gut gemacht hat. Ähm ja, also ich bin jetzt natürlich nicht so in diesem/ Okay, ich komme gerade aus dem Unterricht, bin gestresst und muss eigentlich den nächsten Unterricht vorbereiten. Das wird wahrscheinlich meine, so wie ich das machen würde, vielleicht auch beeinflussen, denke ich mal. Ähm das ist dann wahrscheinlich weniger authentisch, aber ja äh an sich schon. Also es ist auf jeden Fall es hilft mir, glaube ich. (BNW18: 21 - 21)
		Laboratory	The "laboratory" setting was mentioned as limiting the situation's authenticity.	[0:02:47] I: Inwiefern war das für Sie denn eben eine authentische Situation? [0:02:52] TP: Also den Umständen entsprechend fand ich es schon sehr authentisch, weil ich denke schon, dass es vorkommen wird, dass wenn eine Schülerin schlechter abschneidet, dass man ihr dann die Möglichkeit gibt ähm, was einzureichen und dann persönlich mit der darüber zu sprechen. Deshalb kann ich mir das so sehr gut vorstellen in der Schule auch. [0:03:09] I: Und welche Umstände meinen Sie? [0:03:11] TP: Ähm dadurch, dass es halt immer noch eine Laborforschung ist mit ähm den Gegebenheiten, dass alles nochmal erklärt wird, dass sie halt eben nicht

Code			Explanation	Example
				auf dem Wissensstand einer Schülerin ist, sondern, denke ich mal, deutlich weiter und so weiter. [0:03:25] I: Und das hat für Sie auch in der Situation Sie irgendwie anders handeln lassen oder? [0:03:32] TP: Ich denke nicht, aber das ist halt Hintergrundwissen und dann schwierig zu sagen, wie einen das beeinflusst hat. Ich denke eher wenig, aber ist immer noch vorhanden. (MAP27: 32 - 37)
		Authentic without limitations	Situation is considered authentic.	
		General authenticity	Situation is considered authentic without giving any specifications.	[0:00:24] TP: Hm, die war durchaus authentisch und ja, authentisch, also war authentisch. (AAS13: 19 - 19)
		Specifications	Aspects are mentioned that contributed to an authentic situation.	
		Actors	The actor is specified as contributing to authentic perception.	[0:00:33] TP: Hm, ja, also ich glaube auch durch die Leistung der, ähm, Schauspielerin, sie hat sehr authentisch gewirkt, ähm, hat Nachfragen gestellt, die ich auch von, ähm, normalen Schülern erwarten würde. Dementsprechend, ähm, und das beiliegende Material hat dazu beigeholfen. (AEM23: 19 - 19)
		Material	The material is specified as contributing to authentic perception.	[0:00:56] TP: Ich würde sagen, äh sehr authentisch ähm auch was die Anzahl der Fehler betrifft ähm in diesem Text. Also äh das Thema newspaper äh report äh ist ja das Thema auch Jahrgang Neun. Und ähm diese Fehler vor allem in der Grammatik die äh kommen häufiger vor als man denkt auch in Texten. Äh vor allem das sind noch nicht mal Themenfelder, die werden von Lehrern nicht wieder besprochen, weil man erwartet ab einem bestimmten Zeitpunkt, dass die Schüler das einfach können sollen. Zum Beispiel auch irreguläre Verben selber lernen. Äh auch, dass die Schüler das ,he, she, it, das S muss mit' das sieht man bis zum Abitur. Und äh das erzählt eben auch, das bringt einem auch kein Lehrer mehr bei. Das deswegen. (AOS13: 19 - 19)
		Timeframe	The timeframe is specified as contributing to authentic perception.	[0:01:04] TP: Ja ziemlich würde ich sagen, weil ähm/ also ja nicht nur in/ bei der Nachhilfe, sondern auch in der Schule, die Schüler kommen ja immer mal wieder zu einem, fragen nach Feedback ähm und auch dass das mal zwischen den Pausen oder zwischen den Stunden dann geschieht und wo man dann eben nur diese paar Minuten Zeit hat, also ich glaube das ist schon/ das kommt dem schon sehr nahe. (SSM23: 23 - 23)
		Frequency	The frequency of the situation at	[0:00:21] TP: Naja sowas wird ja im Lehrerberuf mehrfach vorkommen, ne, dass Schüler nach den Stunden zu

Code			Explanation	Example
			school is specified as contributing to authentic perception.	einem kommen und Feedback sich einholen, deshalb fand ich das auch schön. Macht Spaß sowas auszuprobieren. (FLA04: 19)
		Own performance	TP's own performance is specified as contributing to an authentic situation.	[0:01:28] TP: Ich kann die Frage so nicht beantworten. Also ich sage mal so, ich bin jetzt zurzeit als Lehrer tätig. Ich äh als DAZ-Lehrer an einer Schule und dort läuft das ja ungefähr auch so ab. Ich äh lasse die Schüler Texte schreiben teilweise und lese die Texte durch und gebe denen anhand des der Texte Feedback. Ähm ich denke, dass es doch schon sehr authentisch ist, vor allem weil man anhand der Fehler arbeitet. Also ich habe mich wirklich auf die Fehler konzentriert, hab aber auch im Endeffekt gesagt, also ich habe immer versucht einen positiven Ton zu halten. Also es war weniger Kritik. Ich habe wirklich gesagt: ‚Ne, das ist falsch. Das ist richtig.‘ Und ähm versucht am Ende trotzdem zu loben und zu sagen: ‚Ich habe es verstanden am Ende des Tages.‘ Solange ich den Text verstehe, ist es für mich eine gute Arbeit gewesen. Und das habe ich dann auch so wiedergespiegelt. Und ich denke mal, dass das dann auch also von meiner Seite aus, doch war schon auch authentisch. (NDA16: 21 – 21)
	Rather Inauthentic		Situation is considered inauthentic.	n.a.
	Limitations			
		Lack of experience	Lack of experience lead to an inauthentic situation.	[0:00:46] TP. Ähm, ich würde sagen von, von dem Aspekt her, dass es überhaupt nicht authentisch rüberkam, aber das kann auch, das kann auch hauptsächlich daran liegen, dass das halt eine ungewohnte Situation für mich war, weil ich so eine solche Simulation noch nicht gemacht habe und ich da auch kein Vergleichswert für habe. (BJL19: 19 - 19)
	Unclear		Position does not become clear.	[00:00:52] TP: Also gut sie ist jetzt nicht 15 gewesen das muss man natürlich klar sagen (lacht). Ähm, es war jetzt natürlich von der Räumlichkeit auch anders so, das war dann, da musste man sich ein bisschen einfügen natürlich ähm, ja gut vielleicht dann halbrealistisch so gesehen, also so dass es dann also am besten wäre es natürlich gewesen, wenn sie dann auch wirklich 15 gewesen wäre, aber dann auch vielleicht auch wegen Datenschutz und so keine Ahnung ein bisschen schwierig gewesen. [0:01:26] I: Mhm, hat Sie das irgendwie rausgebracht, als sie nicht 15 war? Hat es Sie beeinflusst, indem was Sie getan haben?

Code		Explanation	Example
			[0:01:33] TP: Ähm ja, schon also ich arbeite auch außerhalb vom Studium halt auch bei der Nachhilfe auch mit Grundschülern auch und mit Sek1 auch, das wäre dann vielleicht auch vom Tonfall ein kleines bisschen anders gewesen oder (MMK21: 21 – 23)
Situation found in school			
	Rather Positive	Situation is regarded as one that is typical to school	
	Positive with limitations	Situation is regarded as one that is typical to school but with limitations	
	Limitations		
	Frequency	The situation does not occur often	[0:00:51] TP: Äh, es ist natürlich schon eher selten, dass eine Schülerin oder ein Schüler direkt zehn Minuten in der Pause opfert, um über Englisch zu reden, aber trotzdem glaube ich, dass genau diese Form von Feedback dann doch sehr häufig anfällt, vielleicht nicht unbedingt in der Pause, aber zum Beispiel an Elternsprechtagen oder bei Rückgaben der äh Klassenarbeiten oder auch zwischendurch dass man mal Feedback gibt, weil ich glaube es ist relativ wahrscheinlich, da ja Schüler auch mal irgendwie Fragen haben was ihre Noten und ihre Leistungen anbelangt, also ich glaube das ist schon wahrscheinlich das man/ also man muss eigentlich sehr viel Feedback immer geben (AJB12: 19 - 19)
	Timeframe	The timeframe (less OR more time at school) is mentioned as a limiting factor.	[0:01:05] I: Mhm. Und inwiefern war das für Sie eine Situation, wie man sie so in der Schule vorfinden könnte? [0:01:11] TP: Ähm, weiß ich / also mit 15-jährigen, ähm, vielleicht schon eher. Bei jüngeren Sch-, äh, Schülerinnen und Schülern vielleicht weniger, dass sie weniger Nachfragen stellen. Das kommt auf den Typ an, aber ich glaube, dass wenn die jünger sind, dass sie halt weniger Nachfragen würden, sondern eher alles zunehm-, äh, annehmen, was die Lehrkraft sagt. [0:01:33] I: Mhm und so generell diese Situation, dass man an einer Schülerin individuell Feedback gibt, ist das etwas, was in der Schule vorkommt? [0:01:39] TP: Ähm, in der Schule würde ich glaube ich sagen zwischen Tür und Angel. Das man extra so ein Gesprächs-, äh, Zeitfenster hat eher weniger. Ich denke mal eher, also nicht in der Ausführlichkeit, 10 Minuten. Ist, glaube ich, auch nicht umsetzbar.

Code				Explanation	Example
					[0:01:52] I: Also tendenziell ja, aber nicht in dieser Ausführlichkeit? [0:01:54] TP: Genau. (CHS10: 26 - 31)
			Age appropriate-ness	The student's, i.e., participants, (!) age is mentioned as a limiting factor as such a situation would preferably occur with older learners.	[0:00:54] TP: Also ich finde, das war sehr authentisch meiner Meinung nach, weil ich halt solche Situationen damals auch erlebt habe. Ähm nicht halt direkt schon in dem frühen Alter, sage ich mal, so dass ich äh bisschen älter halt war mit Englisch. Äh und erst dann quasi meine Lehrerin mit mir Englisch gesprochen hat. (NRS18: 19 - 19)
			Preparation with material	The preparation with material is mentioned as a limiting factor.	[0:02:49] TP: Hm, ich glaube, was ich nicht so vorfinden würde, wäre das halt, dass ich halt alle Materialien vorher erst lesen müsste. Das ist ja natürlich / ich würde mir dann die, ähm, die Unterrichtsreihe zum Beispiel selber überlegt, das nicht, aber die Situation, dass nach, also gerade wenn's um ein Schreibprozess, ein Erarbeitungsprozess geht, dass die Schüler*innen Feedback erwarten und ich auch nur eine begrenzte Zeit zwischen Stunden habe, das empfinde ich als, ähm, durchaus authentisch oder hilfreich, auch mit der kurzen Vorbereitungszeit vorher, weil ich werde ja auch, wenn ich im normalen Unterrichtsgeschehen bin, werde ich nicht für jeden Entwurf, den die Schüler*innen mir geben super viel Zeit haben, denen Feedback zu geben. Ja, deswegen habe ich das zumindest, ja, insofern als authentisch empfunden, weil's einfach, ähm, ja, einfach eine gute Geschichte drum rum war und einfach sehr kurze Zeit. (ICW08(_M): 23 - 23)
			Context	The context of students seeking feedback on a learning task is a limiting factor.	[0:01:36] I: Hm okay. Inwiefern würdest du denn sagen, dass das eine Situation war wie man sie so in der Schule vorfindet? [0:01:44] TP: Hm also ehrlicherweise würde ich sagen ähm nicht so oft beziehungsweise nicht in der Ausführlichkeit so. Ähm zum Einen aus Zeitgründen auch oder vor allem aus Zeitgründen eigentlich. Und es ist halt auch die Frage, inwiefern ähm sind die Schüler jetzt so engagiert, dass sie ähm ja so gezielt und so detailliert Feedback wollen äh für ihre äh ja für ihre Ausarbeitungen. Jetzt klar bei äh Klassenarbeiten und so weiter ist es ähm, glaube ich, was anderes, da wird es wahrscheinlich gängiger sein, dass äh Schüler und Schülerinnen kommen und wirklich Feedback wollen oder eine Rückmeldung wollen oder Dinge nicht verstanden haben. Aber so wie ich das verstanden habe, war das jetzt wahrscheinlich im Rahmen einer Hausaufgabe oder im Rahmen des Unterrichts ähm einfach Verfassung eines news reports und ähm ja ich weiß nicht, ob es so realistisch ist ähm, dass Schüler dann ja im Zuge so einer

Code				Explanation	Example
					Aufgabe zum einem kommen. Natürlich kann es auch mal passieren, aber dann wahrscheinlich eher am Ende des Unterrichts ähm, wo man sagen muss, das ist halt, ja, oftmals nicht so, ja, nicht so gut, weil man meistens unter Stress ist und zum nächsten Unterricht muss und so weiter, also da äh kann man sich wirklich nicht so die Zeit nehmen ähm ja, Feedback zu geben. [0:03:04] I: Hm ähm in der Vignette steht ja auch drinnen, dass es am Ende der Stunde quasi abgegeben wurde. [0:03:08] TP: Ja. [0:03:09] I: Also das würde ja irgendwie so passen, oder? [0:03:11] TP: Ja. [0:03:11] I: Ja und ähm lag es jetzt dann eher an der Ausführlichkeit des Gesprächs, dass das nicht so wie in der Schule ist, aber tendenziell die Situation an sich, dass man Feedback geben muss ja oder habe ich das jetzt falsch verstanden? [0:03:21] TP: Doch, das ja, auf jeden Fall. Ja, auf jeden Fall. (SAP01: 18 - 25)
			School environment	Lack of a real school environment, e.g., disturbances are a limiting actor.	[0:03:26] I: Mhm. War das denn eine Situation, wie man sie in der Schule vorfinden würde oder könnte? [0:03:34] TP: Es war jetzt ja ruhig ne, also ich glaube wenn man so in der Pause vielleicht nochmal jemand reinkommt, seinen Schal vergessen hat oder so irgendwie so Sachen, die hetzt hier quasi so in einer simulierten Situation so nicht auftreten, aber sonst gehe ich schon davon aus, dass man so eine Situation mal in der Schule findet. (MJH23: 26-27)
			Lack of experience	Lack of experience is mentioned as a limiting factor in answering the question.	[0:01:23] TP: Hm war die authentisch? (...) Also grundsätzlich weiß ich ja, dass es solche Situationen in der Schule gibt, dass die Schüler dann kommen und sagen: ‚Hier, können Sie mal drüber gucken? Was kann ich besser machen? Was ist nicht so gut? Können Sie mir Tipps geben?‘ Also in der Hinsicht hielt ich sie für authentisch, aber dadurch, dass ich noch keine Lehrperson bin, noch nicht richtig, ähm ja/ Es ist wahrscheinlich mein Mindset gewesen und nicht die Situation an sich. (ADH08: 23 - 23)
			Specifications despite limitations		
			Time pressure	The time pressure is mentioned as something that	[0:02:49] TP: Hm, ich glaube, was ich nicht so vorfinden würde, wäre das halt, dass ich halt alle Materialien vorher erst lesen müsste. Das ist ja natürlich / ich würde mir dann die, ähm, die Unterrichtsreihe zum Beispiel

Code				Explanation	Example
				happens at school as well.	selber überlegt, das nicht, aber die Situation, dass nach, also gerade wenn's um ein Schreibprozess, ein Erarbeitungsprozess geht, dass die Schüler*innen Feedback erwarten und ich auch nur eine begrenzte Zeit zwischen Stunden habe, das empfinde ich als, ähm, durchaus authentisch oder hilfreich, auch mit der kurzen Vorbereitungszeit vorher, weil ich werde ja auch, wenn ich im normalen Unterrichtsgeschehen bin, werde ich nicht für jeden Entwurf, den die Schüler*innen mir geben super viel Zeit haben, denen Feedback zu geben. Ja, deswegen habe ich das zumindest, ja, insofern als authentisch empfunden, weil's einfach, ähm, ja, einfach eine gute Geschichte drum rum war und einfach sehr kurze Zeit. (ICW08(_M): 23 - 23)
			Positive without limitations	Situation is regarded as one that is typical to school without limitations	
			General agreement	General agreement is stated.	[0:05:46] TP: Ja, ja also das ist ja auch so ein Ziel, das ist auch etwas, was ich später als Lehrer möchte, dass die Klasse weiß, dass man auch mal, dass ich, wenn es wichtig ist, mir mal Zeit dafür nehme und dementsprechend ist das schon realistisch. Da denke ich, dass man sich mal die Zeit und die Energie dafür, also finde ich schon wichtig und auf jeden Fall, das kann schon so vorkommen. (MMK21: 51 - 51)
			Specifications		
			Frequency	Situation is described as happening frequently at school.	[0:01:18] TP: Ja, auf jeden Fall. Also ich glaube das kommt sehr regelmäßig vor, dass man solche Gespräche führen muss später. (BJP10: 27 - 27)
			Rather Negative	Situation is not regarded as one that is typical to school	[0:01:13] TP: Ich habe das nicht oft erlebt. Also ähm ich habe das wirklich nicht oft erlebt leider. Ähm eigentlich kam das immer nur beim Elternsprechtag und dann auch mit den Eltern, ne. (SAD06: 23 - 23)
			Unclear	Position does not become clear.	[0:03:00] TP: Nach meinem persönlichen Erfahrung eher seltener, aber es ist natürlich nur meine eigene persönliche Erfahrung. Da ist es bei jedem unterschiedlich, aber ich glaube schon, dass das eine Situation ist, die sehr oft kommen würde im Unterrichtsalltag. (AFS29: 27 - 27)

Fairness

Code	Explanation	Example
Rather Positive		
No limitations	No limitations or emphases are made with regards to fairness.	[0:07:31] TP: Ich glaube schon, weil ich glaube genau das, also genau das sind sollen ja Lehrer am Ende machen. Und ich glaube für viele wäre das mal praxisbezogen. Es beschweren sich ja viele darüber, dass das Studium sehr theoretisch ist und nicht praxis- äh nicht so viel Praxis enthält. Und ähm ja da kann ich mir vorstellen, dass das viele gut finden würden. Ja. (BAH29, Pos. 57)
Emphases		
Fairer than written exams	It is emphasized that the RobS is fairer than a written exam.	[0:07:25] TP: Hm ich denke schon, also da das weiß ich jetzt nicht. Also ähm ich meine klar vielleicht müsste man da ein bisschen mehr Forschen, aber ich denke, dass das doch fair bewertet werden kann. Vielleicht sogar/ Vielleicht ist es sogar am Ende des Tages fairer als wenn man eine Klausur schreibt. Da eine Note meistens nicht das widerspiegelt, was eine Person kann oder halt nicht kann. (NDA16, Pos. 43)
Similar to oral exams	It is emphasized that a RobS is as fair as an oral exam.	[0:08:40] I: Kannst du dir denn vorstellen, dass man so eine Prüfung fair bewerten kann? Und damit meine ich kann man in so einer Prüfung das zeigen, was man eigentlich wirklich kann? [0:08:47] TP: Ich würde sagen, viel eher als in anderen Prüfungen. [0:08:50] I: Und warum? [0:08:51] TP: Weil ähm also klar kann man auch in so einer Prüfung irgendwie ein Blackout haben, aber wenn dann / also ich finde es viel schwieriger unter Zeitdruck in einer Klausur irgendwelche Sachen ankreuzen zu müssen als mich da jetzt 15 Minuten vorzubereiten und den Rest sollte ich ja eigentlich sowieso mitbringen, also „Wie rede ich mit einer eingeschüchterten Schülerin?“, zum Beispiel und deshalb finde ich, dass man das gerade gut kann und man wird eben auch mal als Person gesehen und als ja Person im Jetzt in der Prüfungssituation und nicht nur als Matrikelnummer irgendwas, wo man dann gar nicht weiß: „Okay, was war jetzt mit der Person während der Klausur?“. Ja, deshalb würde ich sagen ähnlich wie in einer mündlichen Prüfung, die ich auch generell angenehmer finde oder irgendwie authentischer, äh, könnte man da das gut bewerten. (UMB26_M, Pos. 76-79)
Time pressure like in	It is emphasized that the time	[0:07:51] TP: Ich denke schon, ähm, weil schlussendlich sind Lehrsituation auch immer Stresssituationen, also man

Code		Explanation	Example
	a real-life situation	pressure is similar to real life and contributes to a fair perception.	hat dann nicht nochmal fünf Minuten Zeit, um nochmal irgendwas anzugucken, also ich fand die fünfzehn Minuten waren auch ein guter Zeitraum, ähm, weil man hat als Lehrperson nicht viel Zeit und ich glaube, was man nicht unter Druck in der Prüfung kann, kann man auch sonst nicht, ähm, auch wenn man vielleicht sagt: „Ja, okay ich könnte diese Sachen vielleicht noch, aber halt nur theoretisch und nicht praktisch.“ (SCP15, Pos. 67)
	Formative Function	The formative function is addressed as contributing to fairness	[0:02:38] TP: Ähm, ich glaube es ist, es dient ja eher zur Verbesserung der eigenen, der eigenen Leistung, was einen ja später zum besseren Lehrer machen soll. Ähm, dementsprechend finde ich es sehr sinnvoll. (AEM23: 39 - 39)
	With limitations		
	General Challenge of assessment	Fairness is addressed as a general challenge for assessment.	[0:08:08] TP: Das kann man ja nie. (lachen) Ähm ich denke schon, dass man da relativ gut zeigen kann, wie viel Erfahrung man schon hat und welches Hintergrundwissen ähm, deshalb denke ich, dass man da das wahrscheinlich eher zeigen kann als in einer schriftlichen Überprüfung. Ja. (MAP27: 59 - 59)
	High-pressure situation of assessment	The high-pressure that is baked into summative assessment is addressed.	[0:07:22] TP: Ja also ich meine/ ich meine es ist und bleibt dann eine Prüfungssituation, das heißt ähm man sitzt dann jetzt als Studierender mit gewissem Druck vielleicht oder ein bisschen Nervosität dann auch da und ich denke das ist dann jetzt auch nicht dasselbe wie wenn man jetzt in der Schule also wirklich in der Schule sitzt und dann eine Schülerin nach der Stunde mal kommt ähm also das ist dann natürlich so ein bisschen schwierig aber ich denke im großen und ganzen, wenn man dann auch so darauf vorbereitet wird in den Seminaren ähm dann äh ja doch dann denke ich schon dass man dann auch das zeigen kann. (FCP15: 49 - 49)
	Lack of assessor's prior knowledge	Assessor's lack of knowledge about the participant is addressed.	[0:07:21] TP: Glaube ich eigentlich schon. Natürlich kennen jetzt die Prüfer quasi den Prüfenden nicht und wissen nicht, was sind die, ähm, Gegebenheiten, war der schon mal im Ausland oder hat er vielleicht Englisch als Muttersprache oder sowas, aber ich glaube an für sich kann man schon in dieser Zeit, ähm, feststellen, was quasi, ähm, die Schwächen oder was auch die Stärken der Person sind und worauf man vielleicht auch, ähm, ja, selber dann für sich im Praxissemester oder jetzt auch im Anschluss als / im Referendariat oder sowas, ähm, ja, achten kann, dass man da schon mal so eine erste Rückmeldung bekommt. (AHW03, Pos. 41)
	Fidelity	Lack of fidelity is addressed as potentially affecting fairness.	[0:06:27] TP: Äh, ich denke ich gewisser Weise schon äh, natürlich ist es immer eine Prüfungssituation die niemals hundertprozentig eine realistische Reaktion im Schulalltag darstellt, weil man dann natürlich nicht den Stress hat, weil

Code		Explanation	Example
			man die ganzen anderen Bedingungen, die natürlich im Schulkontext auftreten nicht hat, aber trotzdem werden ja diese essenziellen Skills äh abgefragt und äh, ja dann auch getestet. (AJB12: 43 - 43)
	More time	The need for more time is addressed.	[0:16:12] I: Okay. Ähm, kann man in so einer Prüfung das zeigen, was man wirklich kann? [0:16:18] TP: Das braucht, da braucht man, das ist ja das Ding, man braucht glaube ich dann vielleicht ein paar Minuten mehr als eine viertel Stunde (lacht) ähm, ist vielleicht jetzt auch aus meiner Perspektive nur, ich weiß nicht wie die anderen das gesehen haben. Hm, ja ich habe halt noch relativ wenig Erfahrung mit Feedback geben muss ich sagen. [0:16:42] I: Alles gut. Aber tendenziell würden Sie das eher bejahen oder eher verneinen? [0:16:48] TP: Doch in so einem, doch ja. (MMK21, Pos. 110-113)
	Preparation	Adequate preparation for the assessment is addressed as a requirement for fairness.	[0:09:34] TP: Hm, ich denke schon, dass man da viel zeigen kann, also natürlich muss man sich erstmal so drauf einlassen, ähm, aber / und es ist halt glaube ich gut, wenn man vorher in einem Seminar oder so dann auch nochmal darauf vorbereitet wird, also was gibt es überhaupt für Feedbackmöglichkeiten, weil bei mir war das jetzt zum Beispiel schon wieder total lange her alles, ähm, und was wichtig ist in so einem Gespräch, auch mit in so Elterngesprächen und so, aber dann, ähm, also wenn man da gut drauf vorbereitet ist, glaube ich, kann man das schon ganz gut zeigen, ja.
	Criteria for assessment	Criteria are mentioned as a requirement for fair assessment.	[0:03:57] TP: Hm, es ist halt immer noch eine Prüfungssituation und ich bin zum Beispiel jemand ich bin immer sehr, sehr aufgeregt vor Prüfungen ähm, das ist aber in allen Prüfungen so würde ich sagen/ also ich würde sagen, wenn man recht ähm Richtlinien hat, woran man etwas misst und das eben auch transparent ist, dann hätte ich da jetzt keine [0:04:16] I: Mhm. Also man kann in so einer Prüfung das zeigen, was man wirklich kann. [0:04:18] TP: Genau, ich denke schon. (SSM23, Pos. 49-51)
	Comparability of scenarios	The different scenarios of assessment must be comparable.	[0:09:52] TP: Das ist eine gute Frage. Kann man das fair beantworten? Es bleibt im Ermessensspielraum in einer gewissen Hinsicht natürlich im Ermessensspielraum des Prüfers. Äh der hat natürlich seine ähm auch er oder sie haben ja seine ähm äh ja sein Arbeitsblatt mit den Anforderungen. Sind sie erfüllt worden oder sind sie nicht erfüllt worden? Das könnte sowas sein wie ‚Der zu Prüfende hat nur rumgestottert und hat keinen Satz rausgebracht.‘ Dann ist natürlich die Leistung konnte sie nicht erfolgen. Es müsste fair/ Wie könnte es fair sein? Man müsste dafür sorgen,

Code		Explanation	Example
			dass die Schauspieler, bei denen, die gerade geprüft werden, immer die gleichen sind, dass die sich möglichst gleich verhalten mit den Fragen. Das wäre ein Kriterium. Aufgabenstellung muss natürlich die gleiche sein. Ja, sonst die ganzen Rahmenbedingungen, ne. Aufgabenstellung, Zeit. Ich hoffe, das kann man fair gestalten, aber ich denke schon. (AOS13: 49 - 49)
	Dependent on the individual.	Fairness of the assessment format is addressed positively, but dependent on individuals' strengths and weaknesses.	[0:02:42] I: Mhm und haben Sie auch das Gefühl, dass man in so einer Prüfung das zeigen kann, was man wirklich kann? [0:02:47] TP: Hm ja, es kommt dabei auch immer auf die Situation drauf an, die man dann hat. [0:02:52] I: Ich habe Sie gerade akustisch nicht verstanden, ja oder nein? [0:02:53] TP: Ja, aber kommt auch immer auf die Situation an, die man hat. [0:02:56] I: Mhm, also man kann es in manchen Situationen zeigen und in manchen dann nicht? [0:03:00] TP: Ja, also wie gesagt es kommt auf die Situation an, weil jeder auch unterschiedliche Vorgehensweisen hat und Stärken und Schwächen. (GJS22: 40 - 45)
	Rather Negative		
	Limitations		
	Complexity of the scenario		[0:08:21] I: Mhm. Kann man denn in so einem Rollenspiel oder in so einer Rollenspielprüfung tatsächlich das zeigen, was man auch wirklich kann? [0:08:29] TP: Hm, ne nicht alles, also dafür hat man auch zu wenig Zeit finde ich. Man vergisst viel, also auch wenn man sich die Notizen gemacht hat ist es dieses, dieses man ist im Gespräch man/ dann ist man auch so auf Augenkontakt, man will ja auch schauen und auf diesen Zetteln alles zeigen und dann noch auf die Notizen zu gucken, man switched so sehr, dass man vielleicht ein bisschen von seiner Gehirnleistung braucht, um äh alles unter einen Hut zu kriegen und man ist nicht zu hundertprozent konzentriert und ruft das wieder was man eigentlich gelernt hat. [0:09:03] I: Ist das anders als in anderen Prüfungen? Oder ist das von der Belastung her ähnlich? [0:09:09] TP: Also in einer normalen mündlichen Prüfung sitzt man da nicht mit Notizen da [0:09:12] I: Mhm. [0:09:13] TP: Und für mich selbst, wenn ich Notizen geschrieben habe, muss ich dann darauf gucken und dann überlege ich über diese Notizen und dann kommt man so

Code		Explanation	Example
			ins Stocken. Also so eine normale mündliche Prüfung ist für mich so man redet aus dem Kopf(unv.) [0:09:23] I: Also vom, von dem was man zeigen kann was man wirklich kann, hast du eben gesagt man kann nicht alles zeigen, weil so viel einen auch ablenkt. [0:09:30] TP: Mhm. [0:09:30] I: Und weil es ähm/ man durch die Situation auch dann vielleicht Sachen nicht zeigen kann, weil sie sich vielleicht nicht ergeben, habe ich das richtig verstanden? [0:09:37] TP: Ja, das auch. [0:09:39] I: Mhm. Also oder habe ich das gerade da hineininterpretiert? [0:09:42] TP: Ähm, also ich hatte da jetzt noch nicht so drüber nachgedacht, aber ich finde schon. (SJH28, Pos. 70-81)
	Similar to other assessment formats	Fairness is considered similarly negatively to other assessment formats.	[0:06:49] I: Kann man denn bei so einer Prüfung wirklich das zeigen, was man kann? Oder inwieweit kann man das? [0:06:56] TP: Ich als Lehrerin jetzt? [0:06:58] I: Mhm. [0:06:59] TP: Ähm, naja Prüfungen sind ja immer auch irgendwo tagesformabhängig deswegen würde ich eher sagen nein, natürlich kann man nicht immer alles zeigen, was man wirklich kann. Ähm, dennoch ähm kann man sich ja auf so eine Prüfung vorbereiten und man kann ja auch darauf hinarbeiten, wenn man dann gut vorbereitet ist, dann bin ich mir sicher, dass man schon einiges an Leistung abrufen kann (MWL05: 62 - 65)
	Dependent on the individual	Fairness of the assessment format is addressed as negatively dependent on individuals' strengths and weaknesses.	[0:06:38] TP: Ähm ich sage mal so fair wird halt schwierig, weil halt jeder eine eigene Persönlichkeit hat für sich, wie man mit so einer Situation umgeht. Ich nehme mich mal jetzt als Beispiel. Ich war halt da relativ nervös gewesen. Bin auch nicht so der richtige Sprecher in einer Gruppe. Bin halt quasi auf dem Weg erstmal dahin. Deswegen äh hätte ich auch halt diese Idee halt als Vorschlag, dass man halt so ein Vorher- und Nachherprüfung macht, um halt zu schauen, wie die Entwicklung von einem ist. Und je nach dem kann man dann halt auch etwas fairer, sage ich mal, das Ganze beantworten. Und ja. (NRS18, Pos. 37)
Partially		A rather neutral position is taken.	[0:05:57] TP: Teilweise. Also man ist/ man hat natürlich schon das Gefühl man wird beobachtet und versucht sich dann natürlich zu präsentieren auf eine bestimmte Art und Weise und die Aufregung spielt da dann vielleicht ein bisschen mit rein, dass man nicht alles zeigen kann oder so sagt wie man es gerne sagen möchte. (BNP09: 49 - 49)
Unclear			

Code	Explanation	Example
GenerallyUnclear	No clear position is taken	[0:13:49] I: Denkst du denn, dass man in so einer Prüfung tatsächlich das zeigen kann, was man glaubt selber auch zu können? [0:13:56] TP: Habe ich mich bei mir auch gefragt. Ich war so ein bisschen: ‚Okay, weiß ich gar nicht, ob das jetzt kompetent ist, was ich da hier gerade mache.‘ Ich glaube, es zeigt halt nicht nur das, was ich in der Situation sage, sondern man es kommt halt auch drauf an. Nimmt man dann auch so Faktoren mit rein. Die Person redet super schnell und super undeutlich und man wirkt total nervös und so. Also auch so wie kann ich umgehen mit dieser Prüfungssituation. Deswegen habe ich mich auch gefragt. Weiß ich nicht, ob das jetzt so repräsentativ ist. Gleichzeitig halte ich ähm eine Klausur oder eine Hausarbeit auch nicht für besonders repräsentativ, weil da sehr wenig, finde ich, von dem, was mich als Person als Lehrkraft ausmacht, zum Tragen kommt. (EME03_M: 48 - 49)
Comparable to other assessment formats	The fairness of the simulation is compared to other assessment formats	[0:08:38] TP: Hm ja, also ich sehe jetzt eine Simulation nicht anders als eine mündliche Prüfung, auch die es in Englisch gibt. Die gibt es ja schon lange und oft und da ist ja auch immer die Diskussion, inwiefern hängt dein Ergebnis mit den Dozenten ab, die dich prüfen, mit der Prüfungssituation, mit den Inhalten, die da spontan abgefragt werden und da sehe ich jetzt keinen Unterschied zu einer Simulation, nur dass halt keine lineare mündliche Prüfung ist, wo mir Fragen gestellt werden, sondern dass ich halt interagieren muss mit einer Person. [0:09:08] I: Hm und sind so verschiedene Leistungen, die in so einer Simulation dann auch gezeigt werden können, auch mit einander vergleichbar? Also wenn man jetzt vielleicht immer die gleiche Situation hat, aber da sitzen zum Beispiel verschiedene Schauspielerinnen vor einem. Ist das trotzdem dann mit verschiedenen Personen immer noch eine vergleichbare Leistung? [0:09:24] TP: Hm ja, würde ich eigentlich schon sagen, also genau wieder, was ich gerade gesagt habe, mit den verschiedenen Dozenten, die einen überprüfen ähm, die reagieren ja in einer mündlichen Prüfung auch immer verschieden und ich denke, das gehört auch mit dazu später im Lehreralltag mit verschiedenen Persönlichkeiten klarzukommen. Und zudem glaube ich auch, dass die, wenn jetzt äh andere Studierende dafür auserwählt werden mit in der Prüfung mitzuspielen oder studentische Hilfskräfte oder wie auch immer ähm, werden die ja schon einen gewissen Standard erfüllen, was so/ Das wird ja nicht komplett wahllos sein, was die da sagen. (ALA02: 45 - 47)

Tested Competences

Code		Explanation	Example
Tested competences general			
	Feedback		
	General: Feedback competence	A general referral to feedback is made, without being specific.	[0:01:53] TP: Hm, inwiefern man Feedback geben kann und ob es auch konstruktiv ist. (GJS22: 33 - 33)
	Dealing with errors and mistakes	The appropriate handling of errors and mistakes is addressed.	[0:01:10] TP: Ähm, ich glaube Rückmeldung geben von, ähm, äh, der, hm, den Text, den die Schülerin bearbeitet hat, ähm, das Ganze zu erläutern, wo die Fehler liegen, was man verbessern kann, ähm, wie man sich selber verbessern kann. Ich glaube, das war sehr authentisch auch. (AEM23: 27 - 27)
	Feed Up	Formulating learning objectives is mentioned as a required competency.	[0:03:48] TP: Ähm auf jeden Fall sehr viele, weil ähm es ist es geht ja nicht nur um Unterricht, sondern auch darum, dass man die Schüler innen so gut wie es geht individuell fördert. Ähm das heißt auch also ich als Lehrkraft habe auch die Möglichkeit angeboten, dass die Schüler innen Texte einreichen und ich Feedback gebe. Ähm deswegen ist das auf jeden Fall ähm eine Kompetenz, die auch wichtig ist, ähm auch dass man so in gewisser Art und Weise ähm kontrolliert man damit ja auch das Lernziel, was ja auch auf diesem Unterrichtsreihen-Zettel stand. Ähm das heißt, diese Kompetenz der Lernzielformulierung kann man damit vielleicht auch irgendwie noch so ein bisschen fördern, würde ich behaupten. Und ähm generell das Sozialverhalten ähm wird damit ja auch gefördert, weil man, ich glaube man wächst da dran auch, man wird immer besser. Ähm und auch generell so dieses Feingefühl und ich finde auch so diese das kann auch die Atmosphäre in der Klasse positiv beeinflussen, weil man ja außerhalb des Unterrichts Schüler auch nochmal motivieren kann. Ähm genau, das wäre dann insofern, dass man ja auch nochmal eine andere Bindung aufbaut. Es ist natürlich trotzdem wichtig ähm so eine gewisse Distanz immer zu wahren. Ja. (MMM07: 31 - 31)
	Feed Back	Providing information to the learner is mentioned as a required competency.	[0:12:26] TP: Also Feedbackkompetenz ist ja ganz wichtig. Mir war wichtig als ich das Feedback gegeben habe, erstmal positiv zu sein ähm, um die Schülerin nicht zu überwältigen. Also erstmal auf das eingehen, was sie

Code			Explanation	Example
				gut gemacht hat. Wie ich fand in dem Fall, dass sie viele der Informationen schon mitreingenommen hat. Erst dann die Fehler besprechen. Äh da jetzt auch nicht zu sehr den Knüppel raushauen. Ich habe auch nicht jeden einzelnen Fehler angesprochen, den sie gemacht hat. Ich habe sie aber angestrichen, weil sie sie ja verbessern muss. Ähm und eben nicht nur sagen, was hat nicht geklappt, sondern was kann ich tun, das zu verbessern. Da hatte ich mich ja Mia auch proaktiv drauf angesprochen, dass sie gesagt hatte, dass ihr das Simple Past schwerfällt, was mir auch aufgefallen ist. Das hatte ich auch angesprochen. Da habe ich ihr dann die Empfehlung gegeben: ‚Okay, dann arbeite nochmal die Regeln auf. Schau dir das nochmal an.‘ Äh also nicht nur Feedback geben, sondern auch eben dieses Feedforward, also was, dass die etwas haben, an dem sie arbeiten können, das sie anwenden können, um es besser zu machen. (MJO28: 59 - 59)
		Feed Forward	The support of the learner (i.e., how to improve) is mentioned as a required competency.	[0:12:26] TP: Also Feedbackkompetenz ist ja ganz wichtig. Mir war wichtig als ich das Feedback gegeben habe, erstmal positiv zu sein ähm, um die Schülerin nicht zu überwältigen. Also erstmal auf das eingehen, was sie gut gemacht hat. Wie ich fand in dem Fall, dass sie viele der Informationen schon mitreingenommen hat. Erst dann die Fehler besprechen. Äh da jetzt auch nicht zu sehr den Knüppel raushauen. Ich habe auch nicht jeden einzelnen Fehler angesprochen, den sie gemacht hat. Ich habe sie aber angestrichen, weil sie sie ja verbessern muss. Ähm und eben nicht nur sagen, was hat nicht geklappt, sondern was kann ich tun, das zu verbessern. Da hatte ich mich ja Mia auch proaktiv drauf angesprochen, dass sie gesagt hatte, dass ihr das Simple Past schwerfällt, was mir auch aufgefallen ist. Das hatte ich auch angesprochen. Da habe ich ihr dann die Empfehlung gegeben: ‚Okay, dann arbeite nochmal die Regeln auf. Schau dir das nochmal an.‘ Äh also nicht nur Feedback geben, sondern auch eben dieses Feedforward, also was, dass die etwas haben, an dem sie arbeiten können, das sie anwenden können, um es besser zu machen. (MJO28: 59 - 59)

Code		Explanation	Example
	Positive Feedback	Ability to mention aspects that the learner has mastered successfully is mentioned as a required competency.	[0:12:26] TP: Also Feedbackkompetenz ist ja ganz wichtig. Mir war wichtig als ich das Feedback gegeben habe, erstmal positiv zu sein ähm, um die Schülerin nicht zu überwältigen. Also erstmal auf das eingehen, was sie gut gemacht hat. Wie ich fand in dem Fall, dass sie viele der Informationen schon mitreingenommen hat. Erst dann die Fehler besprechen. Äh da jetzt auch nicht zu sehr den Knüppel raushauen. Ich habe auch nicht jeden einzelnen Fehler angesprochen, den sie gemacht hat. Ich habe sie aber angestrichen, weil sie sie ja verbessern muss. Ähm und eben nicht nur sagen, was hat nicht geklappt, sondern was kann ich tun, das zu verbessern. Da hatte ich mich ja Mia auch proaktiv drauf angesprochen, dass sie gesagt hatte, dass ihr das Simple Past schwerfällt, was mir auch aufgefallen ist. Das hatte ich auch angesprochen. Da habe ich ihr dann die Empfehlung gegeben: ‚Okay, dann arbeite nochmal die Regeln auf. Schau dir das nochmal an.‘ Äh also nicht nur Feedback geben, sondern auch eben dieses Feedforward, also was, dass die etwas haben, an dem sie arbeiten können, das sie anwenden können, um es besser zu machen. (MJO28: 59 - 59)
	Assessment	Aspects are mentioned that focus on the teacher and their ability to assess/diagnose/evaluate.	
	Evaluation	Assessment/evaluation skills are mentioned. TP need to be able to make a judgement of the learners' performance.	[0:01:56] TP: Also ich musste schon echt äh manchmal bisschen nachdenken, so es war jetzt nicht so, dass ich mich da einfach hingestellt habe und äh gesagt habe ‚Ja, ich mache das jetzt mal eben schnell.‘ Also ich musste schon deutlich nachdenken, wie kann ich ihr denn helfen und äh auch spontan äh dachte ich erst so ‚Okay, durchatmen, nicht überfordert sein und äh einfach das machen, was man so als Lehrer macht, ne.‘ Und äh ja. [0:02:20] I: Und was ist das genau, was man da machen müsste und was so wichtig war für Lehrkräfte? [0:02:23] TP: Ähm Evaluieren einfach, ne. Was ist gegeben? Wie kann ich damit arbeiten? (SAD06: 31 - 33)
	Correction	Correction/ corrective competence is mentioned.	[0:02:17] TP: Hm auf jeden Fall einmal das Korrigieren von Schülertexten und dann das

Code			Explanation	Example
			TP need to be able to correct false statements made by the learner.	Erfassen von den Kompetenzen der Schüler, ähm das dann einmal zu benennen und auch Förderungsmöglichkeiten zu finden und die dann auch weiterzugeben. (BJP10: 37 - 37)
		Diagnosis	The diagnosis of textual features and/or the learner level is mentioned. TP need to be able to diagnose the learners' level of competence by detecting successful as well as failed aspects of performance.	[0:03:26] TP: Ähm, ich glaube vor allem, vielleicht Kompetenzen zur Lernstandserhebung in dem Sinne, dass ich jetzt erstmal schauen muss, okay, wo steht die Schülerin gerade ähm aber wie kann ich auch meinen oder ja wie passe ich meine zukünftigen Unterrichtsstunden dahingehend an, dass ich diese Fehler von Mia ähm ja verhindern kann, weil offensichtlich hat sie das mit dem Simple Past noch nicht so ganz verstanden, dann ähm die Struktur des Reports war jetzt auch nicht unbedingt gut obwohl das ja erst ich glaube eine Doppelstunde davor gemacht wurde, ja. Deswegen und natürlich auch Flexibilität. Das ich halt wirklich in dem Moment sagen kann: „Okay, ich lese mir das eben schnell durch und schaue wo liegen deine Probleme“ (MWL05: 33 - 33)
	Teacher-student-interaction			
		General: Dealing with learners	Ability to appropriately deal with learners is mentioned.	0:02:44] TP: Ähm, allgemein erstmal dieses spontane Reagieren, mit Schülerinnen und Schülern umgehen, auf Fragen und Wünsche eingehen und allgemein das man eben halt auch Feedback geben kann in dem Moment, wo es gebraucht wird oder gefordert wird, ja. (SSM23: 43 - 43)
		Emotional-motivational competences		
		Social competence	General social competence, e.g., ability to appropriately engage in social interactions and see the learner as a person is mentioned.	[0:04:49] TP: Ähm, also auf jeden Fall so Feedback geben, dabei dann natürlich auch so soziale Kompetenzen zu beachten, dass man da irgendwie gutes Maß findet, um die, um nicht demotivierend zu sein, gerade bei schwachen Schülern oder Schülerinnen, ähm, aber halt auch eben zu vermitteln, dass da auf jeden Fall was gemacht werden muss. Das fand ich nicht so einfach und ähm, ja, auch hm, klar auch so einfach die Kompetenz Fehler zu finden in Texten und äh auch dann zu erkennen, was für Fehler das sind, was die Ursachen dafür sind, waren das Flüchtigkeitsfehler oder fehlt da einfach wirklich das

Code			Explanation	Example
				Grundverständnis von der Grammatik, ähm, ja. (MNA19_M: 43 - 43)
		Emotions	Ability to adequately cope with and anticipate learner emotions.	[0:03:50] TP: Ja, das auf jeden Fall auch dieses auf Emotionen eingehen, Motivieren und gerade diese individuelle Bezugsnorm. Die es halt gerade für schwächere Schüler sehr wichtig, dass man da als Schüler oder Lehrer eher realisiert, okay, es bringt jetzt nichts, wenn ich hier die normalen Kriterien heranziehe und sage: ‚Das ist immer noch nicht das, was man sich erhofft‘, sondern dass man dann individuell draufgeht, okay, das ist schon ein sehr guter Start, noch ein bisschen mehr Arbeit, dann kannst du dich verbessern. Also dieser Aspekt ist denke ich sehr gut geübt. (MAP27: 39 - 39)
		Empathy	Ability to understand learner's emotions and empathize is mentioned.	[0:04:27] TP: Gute Frage. Ähm, ja einmal die Fähigkeit sich in, in andere hinzusetzen, also einfach Empathie zu entwickeln, warum gewisse Fehler gemacht werden oder, ähm, ja Fehler klingt immer / ist immer sehr negativ konnotiert, aber zum Beispiel das mit dem englischen Wort „become“ wird ja oft von Schüler*innen falsch verwendet, weil es einfach wie so ein falscher Freund ist, es ist der deutschen Muttersprache einfach eben sehr ähnlich und ich finde, dass die Kompetenz die da gefordert wurde, die man eigentlich auch ganz gut umsetzen kann, beziehungsweise die auch ganz gut nachgewiesen werden kann, eben dass man versucht das möglichst, ähm, auf der Basis der fünfzehnjährigen Schülerin zu vermitteln und eben nicht wie in einem, ja, Professionsgespräch nur unter Lehrkräften, äh, Fachtermini da reinzuhauen. (ELH30: 33 - 33)
		Motivation	Ability to motivate / not demotivate learner is mentioned.	[0:04:49] TP: Ähm, also auf jeden Fall so Feedback geben, dabei dann natürlich auch so soziale Kompetenzen zu beachten, dass man da irgendwie gutes Maß findet, um die, um nicht demotivierend zu sein, gerade bei schwachen Schülern oder Schülerinnen, ähm, aber halt auch eben zu vermitteln, dass da auf jeden Fall was gemacht werden muss. Das fand ich nicht so einfach und ähm, ja, auch hm, klar auch so einfach die Kompetenz Fehler zu finden in Texten und äh auch dann zu erkennen, was für Fehler das sind, was die Ursachen dafür sind, waren das

Code			Explanation	Example
				Flüchtigkeitsfehler oder fehlt da einfach wirklich das Grundverständnis von der Grammatik, ähm, ja. (MNA19_M: 43 - 43)
		Relationship	The ability to create an appropriate relationship with a learner is mentioned.	[0:03:48] TP: Ähm auf jeden Fall sehr viele, weil ähm es ist es geht ja nicht nur um Unterricht, sondern auch darum, dass man die Schüler innen so gut wie es geht individuell fördert. Ähm das heißt auch also ich als Lehrkraft habe auch die Möglichkeit angeboten, dass die Schüler innen Texte einreichen und ich Feedback gebe. Ähm deswegen ist das auf jeden Fall ähm eine Kompetenz, die auch wichtig ist, ähm auch dass man so in gewisser Art und Weise ähm kontrolliert man damit ja auch das Lernziel, was ja auch auf diesem Unterrichtsreihen-Zettel stand. Ähm das heißt, diese Kompetenz der Lernzielformulierung kann man damit vielleicht auch irgendwie noch so ein bisschen fördern, würde ich behaupten. Und ähm generell das Sozialverhalten ähm wird damit ja auch gefördert, weil man, ich glaube man wächst da dran auch, man wird immer besser. Ähm und auch generell so dieses Feingefühl und ich finde auch so diese das kann auch die Atmosphäre in der Klasse positiv beeinflussen, weil man ja außerhalb des Unterrichts Schüler auch nochmal motivieren kann. Ähm genau, das wäre dann insofern, dass man ja auch nochmal eine andere Bindung aufbaut. Es ist natürlich trotzdem wichtig ähm so eine gewisse Distanz immer zu wahren. Ja. (MMM07: 31 - 31)
		Conversational competences	Aspects with regards to how to conduct conversations are mentioned	
		Communication	Ability to communicate adequately is mentioned.	[0:03:59] TP: Also ich finde das man als Lehrkraft/ das die zentralste Kompetenz als Lehrkraft eigentlich eine Kommunikationskompetenz ist, weil in jedem Bereich wo wir wirklich als Lehrer in Erscheinung treten kommunizieren wir und ich glaube das ist die zentralste Kompetenz die man haben kann als Lehrer und Feedbackgespräche sind natürlich super wichtige Situationen wo es halt/ wo Kommunikation halt sehr wichtig ist und auch richtige Kommunikation weil man spricht ja über Fehler von einem Menschen, aber es muss halt klar werden das man hier eine (unv.) was

Code			Explanation	Example
				den Menschen an sich betrifft, sondern quasi eine Leistung eines Menschen und eine Leistung, die man verbessern kann und man nicht den Menschen an sich bewertet und ich glaube da muss man halt äh gut kommunizieren um auch das so ein bisschen rüber zu bringen. (AJB12: 35 - 35)
		Structuring	The ability to structure the feedback conversation is mentioned.	[0:04:45] TP: Ähm, also es hat auf jeden Fall Erklärkompetenz aufgegriffen, also wie erkläre ich etwas. Ähm, auch wie strukturiere ich Feedback, also da musste ich mir echt überlegen, was will ich / was für Punkte will ich ansprechen? Ähm, ich glaube auch einfach generell Kommunikation mit Schülern. Wie spreche ich die Schüler an? Wie kann ich die an mich ran holen, dass sie mir auch zuhören und auch gewillt sind / gut, die Schülerin ist natürlich, hat den ersten Schritt gemacht und hat nach Feedback gefragt, ähm und ich glaube es zeigt einem halt eben auch auf, wo also was, was für Wissen man hat im eigenen Bereich. Also, weiß ich jetzt die ganzen Regeln, wann man Passiv benutzt? Wann man / ähm wie man Present Simple oder was auch immer bildet? Ähm, ich glaube das ist auch einfach die Wissenskompetenz, die vor allen Dingen dann auch selber nochmal abgefragt wird. (SCP15: 49 - 49)
		Subject-related interaction		
		Learner orientation	Aspects are mentioned regarding adapting to learner's individual needs.	[0:02:33] TP: Hm ich würde sagen, so genau auf den Schüler einzugehen, also auf seine ja individuellen Lernstände, so nicht nur grob zu unterrichten, sondern wirklich auch zusagen, 'Okay, hier hast du das und das falsch gemacht'. Und auch ja Tipps weiterzugeben und vielleicht auch Strategien, wie man diese Defizite dann beheben kann. Ich glaube, das wird relativ oft vorkommen im Lehrersein. (ACG22: 29 - 29)
		Prioritization	The ability to select, reduce and prioritize features for the feedback conversation is mentioned.	[0:04:49] TP: Ähm (...) ich denke, erstmal im Sinne von Feedback geben, dass man, weil ich es gerade schon gesagt habe, Feedback in reduzierter Form gibt. Dann, dass man ähm sein Sprachniveau oder nicht nur das Sprachniveau, aber auch das geschriebene Englisch, auf das Level der Schülerinnen und Schüler anpasst, also, dass ich jetzt nicht so korrigiere als würde ich einen Q2-Schüler

Code			Explanation	Example
				korrigieren oder Schülerin. Ähm genau und, dass man halt so dieses positive Feedback trotzdem noch beibehält so. (ALA02: 37 - 37)
		Explanation	TP's ability to explain is mentioned.	[0:04:45] TP: Ähm, also es hat auf jeden Fall Erklärkompetenz aufgegriffen, also wie erkläre ich etwas. Ähm, auch wie strukturiere ich Feedback, also da musste ich mir echt überlegen, was will ich / was für Punkte will ich ansprechen? Ähm, ich glaube auch einfach generell Kommunikation mit Schülern. Wie spreche ich die Schüler an? Wie kann ich die an mich ran holen, dass sie mir auch zuhören und auch gewillt sind / gut, die Schülerin ist natürlich, hat den ersten Schritt gemacht und hat nach Feedback gefragt, ähm und ich glaube es zeigt einem halt eben auch auf, wo also was, was für Wissen man hat im eigenen Bereich. Also, weiß ich jetzt die ganzen Regeln, wann man Passiv benutzt? Wann man / ähm wie man Present Simple oder was auch immer bildet? Ähm, ich glaube das ist auch einfach die Wissenskompetenz, die vor allen Dingen dann auch selber nochmal abgefragt wird. (SCP15: 49 - 49)
		Appropriate use of language as medium of instruction	Use of language as a transmitter/ as medium of instruction is mentioned as a competence.	[0:04:49] TP: Ähm (...) ich denke, erstmal im Sinne von Feedback geben, dass man, weil ich es gerade schon gesagt habe, Feedback in reduzierter Form gibt. Dann, dass man ähm sein Sprachniveau oder nicht nur das Sprachniveau, aber auch das geschriebene Englisch, auf das Level der Schülerinnen und Schüler anpasst, also, dass ich jetzt nicht so korrigiere als würde ich einen Q2-Schüler korrigieren oder Schülerin. Ähm genau und, dass man halt so dieses positive Feedback trotzdem noch beibehält so. (ALA02: 37 - 37)
	Situation-related			
	Organiza- tion/time man- agement		The management and handling of time resources is mentioned as a competence.	[0:01:45] TP: Hm einmal Stressmanagement, also man hat nur zehn Minuten Zeit direkt nach dem Unterricht ähm, dass man das ganz schnell und möglichst also möglichst professionell, dass man auch nichts vergisst, dass man da alle Aspekte ähm ja mitgibt. Und zum anderen ähm, dass das auch alles ankommt, also man gibt das ab, aber das muss dann auch alles bei dem Schüler, bei der Schülerin ankommen. Und ja. (AAS13: 29 - 29)

Code		Explanation	Example
	Spontaneity/flexibility	Ability to react spontaneously and flexibly in the situation is mentioned.	[0:03:02] I: Hm, ich würde sagen halt dass man so auf/ in so einem Feedbackgespräch auf so, ja spontan reagieren muss ähm dass man natürlich auch einen Gegenüber hat, wo man aufpassen muss, wie man das formuliert, dass man die Schüler nicht ähm demotiviert, also dass man ein bisschen auch Einfühlungsvermögen ähm da übt, ähm das würde ich sagen hat man schon da gehabt, natürlich kannte ich jetzt nicht den Hintergrund von der Person außer diese zwei Informationen, die da vorher standen ähm dass sie halt äh zum Beispiel eine schlechte Schülerin oder nicht schlecht, aber ja (lacht) zwischen drei und vier steht, dass man halt aufpassen muss dass man sie auch nicht zu sehr demotiviert ähm, dass man das halt im Hintergrund hat ähm ja sonst hatte ich zwar kein Hintergrundwissen, aber ich denke schon, dass die Kompetenzen äh dann erfasst wurden, ja. (BNP09: 39 - 39)
	Adaptation to the situation	Ability to adapt to the situation is mentioned.	[0:03:17] TP: Also es ist wurden die Kompetenzen erfasst, dass man halt schaut wie man mit Schüler innen reden kann, wie man ein vernünftiges Feedback geben kann, dass man das an die Situation auch anpasst. Ähm inwiefern das dann da ja das. Passt das? (lachen) Ich weiß nicht. (lachen) (SLG21_M: 39 - 39)
	Self-reflection	Ability to use the information to improve one's own teaching is mentioned.	[0:02:58] I: Mhm. Okay, gut. Inwiefern wurden mit dem Rollenspielen denn eben Ihrer Meinung nach Kompetenzen erfragt oder welche Kompetenzen musste man anwenden, die für den Beruf als Lehrkraft notwendig sind? [0:03:10] TP: Ähm, einige. Einmal generell den Text nochmal durchgehen, dann auch mal sagen, was man genau markiert hat und auch begründen. Das Begründen ist, glaube ich, ganz, ganz wichtig und hilfreich für die Schülerinnen und Schüler und das war in der Situation halt auch für mich halt auch immer wieder hilfreich zu gucken, was hilft, äh, der Person und wie begründe ich das, wie bringe ich es rüber. [0:03:31] I: Was genau meinen Sie mit „begründen“? [0:03:33] TP: Ähm, warum das, ähm, (unverständlich, ca. 2 Sekunden), ähm

Code		Explanation	Example
			[0:03:37] I: Also warum etwas falsch ist, oder? [0:03:38] TP: Ähm, warum etwas falsch ist und warum man es anders schreiben würde. Das mit dem „Tuesday“ habe ich ja erläutert, dass das einfach ein Dienstag ist, wenn man sagt welcher Tag mit Datum oder vielleicht auch den Ort, dann kann man es zuordnen. [0:03:50] I: Mhm. [0:03:50] TP: Ähm / [0:03:51] I: Also auch erklären? [0:03:51] TP: Ja erklären, [0:03:52] I: Mhm. [0:03:53] TP: dass man es nachvollziehen kann fürs nächste Mal es anders gestalten kann. [0:03:55] I: Mhm. Noch weitere Kompetenzen? [0:03:58] TP: Hm, ja einmal die Selbstreflexion für die Lehrer, warum sage ich das und äh nicht davon ausgehen, dass ich es einmal gelernt habe und es immer kann. [0:04:07] I: Mhm. [0:04:08] TP: Das ist wichtig und äh ja generell Kompetenzen, also Fachwissen, dann, ja, dieses Begründen ist glaube ich ganz, ganz wichtig. (CHS10: 44 - 59)
	Subject-related knowledge		
	Professional Knowledge (General)	Professional knowledge is addressed in General terms.	[0:04:54] TP: Ähm also natürlich einmal das Fachliche, dass man diese Texte beurteilen kann, dass man selbst halt dieses Wissen hat. Ähm also das wird ja jetzt auch im Studium schon relativ gut vermittelt, aber so diese ganzen, ich sage mal, diese Softskills: Wie formuliere ich Feedback? Ähm wie gehe ich mit Schülern um? Wie/ Da hatte ich das Gefühl, dass äh ja das wurde bis jetzt irgendwie gar nicht vermittelt. Also ich bin ja noch im Bachelor und also sowas dann im Bachelor auch als Prüfung einzusetzen, natürlich, wenn man vorher drauf vorbereitet wird, finde ich sehr sinnvoll. (KMP05: 51 - 51)
	Linguistic knowledge	Knowledge of explicit linguistic knowledge.	[0:06:25] TP: Für so ein Feedback ähm, ja also die ähm, was ich gerade so ein bisschen gesagt habe, dass man sag ich mal irgendwie auch den Raum dafür hat, dass man sich die

Code		Explanation	Example
			Zeit nimmt und die Energie irgendwie, also auch Geduld mitbringt auf jeden Fall, Empathie sowieso ähm und auch so die didaktische Kompetenz bzw. die, wie nennt man das, die also die Sprachwissenschaft die dahinter steckt, dass man das halt, dass das auch so eine der Sachen wo ich sagen kann, das habe ich im Bachelor gelernt so, ich bin zwar eher so am Ende vom Studium aber das ist sowas, dass konnte ich vorher nicht, dass ich das in Kategorien einteile und sage okay das ist das und das, wenn Sie wissen was ich meine. (MMK21: 55 - 55)
	Mastering of the English language	Ability to speak English is mentioned.	[0:04:21] TP: Ähm zum Einen war es ja wichtig, dass man der englischen Sprache mächtig ist und die Fehler korrigiert. Das heißt in Schrift, Wort, also da waren Syntaxfehler. Da waren Rechtschreibfehler. Ähm also die wichtigsten Kompetenzen waren, sage ich mal, vorausgesetzt. Ja. Gleichzeitig geht es auch darum, es ist ja eine Art Feedbackgespräch. Es ist ein Feedbackgespräch. Und man muss ja auch Feedback geben und ich denke, dass das sozusagen die wichtigsten Eigenschaften einer Lehrkraft sind. Ich könnte mich da auch hinstellen und sagen: ‚Ja, ne alles falsch. Mach es neu.‘ Äh ist ja auch nicht Sinn der Sache. Dann weiß sie ja auch nicht äh, also weiß sie ja nicht, was genau wirklich falsch gemacht wurde. Man hat sich hingesetzt. Man hat/ Ich hatte äh Platz für Notizen. Ich musste Notizen machen. Diese Notizen musste ich dann der Schülerin erklären. Ähm ich denke, die wichtigsten Kompetenzen wurden doch äh vorausgesetzt und berücksichtigt für einen Lehrer. (NDA16: 35 - 35)
	Other	Aspects mentioned only once.	
	Action-oriented competence	Action-oriented competences, e.g., ability to use information in the situation is mentioned.	[0:04:33] TP: Ähm, jetzt müsste ich die Kompetenzen finden. Ich würde auf jeden Fall behaupten, dass es eine Handlungskompetenz ist auf jeden Fall, weil das ein sehr natürliches Setting ist das kann passieren. Ähm, dann Methodenkompetenz (SEL31: 33 - 33)
	Solution-orientation	Ability to focus on solutions is mentioned.	[0:04:19] TP: Ähm, naja es wurden Kompetenzen/ Ja, also (lachen) ich musste ja quasi so tun als wäre ich Lehrerin und es wurden die Kompetenzen einer/ Also erstens halt

Code		Explanation	Example
			Zeitmanagement. Ich hatte die Zeit überhaupt nicht im äh im Blick. Äh sie hat mich eher darauf hingewiesen indem sie auf die Uhr geguckt hat und indem sie auch gesagt hat ‚Ich muss zum Bus.‘ Und äh da habe ich gemerkt, dass ich das gar nicht drauf habe. Und ähm ja die Kompetenz, ich weiß nicht, wie man das nennt, aber etwas korrigieren zu können, den Fehler zu entdecken und zu gucken, was also worum handelt es sich hier überhaupt. Äh weil, wenn ich jetzt einen Grammatikfehler sehe und den als Grammatikfehler erkenne, dann heißt es ja nicht direkt, dass ich der Schülerin erklären kann, was jetzt genau ihr Fehler war, woran das liegt und was sie machen muss. Also diese, ich weiß nicht genau, wie man das nennt. [0:05:13] I: Diagnose vielleicht. [0:05:14] TP: Ja, ja. Solche Sachen und äh vielleicht es auf den Punkt bringen einfach, weil ich habe ganz viele ähm gesagt und es überhaupt nicht verknüpft und dann hat sie, also ich glaube, sie hatte am Ende auch immer noch viele Fragen. Also ich glaube nicht, dass ich das gut auf den Punkt gebracht habe, woran genau, sie arbeiten muss und ähm was sie tun muss. Also ich habe keine lösungsorientierte ja Herangehensweise gehabt. (BAH29: 43 - 45)
	Efficiency	Ability to be precise is mentioned.	[0:04:19] TP: Ähm, naja es wurden Kompetenzen/ Ja, also (lachen) ich musste ja quasi so tun als wäre ich Lehrerin und es wurden die Kompetenzen einer/ Also erstens halt Zeitmanagement. Ich hatte die Zeit überhaupt nicht im äh im Blick. Äh sie hat mich eher darauf hingewiesen indem sie auf die Uhr geguckt hat und indem sie auch gesagt hat ‚Ich muss zum Bus.‘ Und äh da habe ich gemerkt, dass ich das gar nicht drauf habe. Und ähm ja die Kompetenz, ich weiß nicht, wie man das nennt, aber etwas korrigieren zu können, den Fehler zu entdecken und zu gucken, was also worum handelt es sich hier überhaupt. Äh weil, wenn ich jetzt einen Grammatikfehler sehe und den als Grammatikfehler erkenne, dann heißt es ja nicht direkt, dass ich der Schülerin erklären kann, was jetzt genau ihr Fehler war, woran das

Code		Explanation	Example
			liegt und was sie machen muss. Also diese, ich weiß nicht genau, wie man das nennt. [0:05:13] I: Diagnose vielleicht. [0:05:14] TP: Ja, ja. Solche Sachen und äh vielleicht es auf den Punkt bringen einfach, weil ich habe ganz viele ähm gesagt und es überhaupt nicht verknüpft und dann hat sie, also ich glaube, sie hatte am Ende auch immer noch viele Fragen. Also ich glaube nicht, dass ich das gut auf den Punkt gebracht habe, woran genau, sie arbeiten muss und ähm was sie tun muss. Also ich habe keine lösungsorientierte ja Herangehensweise gehabt. (BAH29: 43 - 45)
	Provision of resources	Ability to provide the (timely, mental) resources is mentioned.	TP: Für so ein Feedback ähm, ja also die ähm, was ich gerade so ein bisschen gesagt habe, dass man sag ich mal irgendwie auch den Raum dafür hat, dass man sich die Zeit nimmt und die Energie irgendwie, also auch Geduld mitbringt auf jeden Fall, Empathie sowieso ähm und auch so die didaktische Kompetenz bzw. die, wie nennt man das, die also die Sprachwissenschaft die dahinter steckt, dass man das halt, dass das auch so eine der Sachen wo ich sagen kann, das habe ich im Bachelor gelernt so, ich bin zwar eher so am Ende vom Studium aber das ist sowas, dass konnte ich vorher nicht, dass ich das in Kategorien einteile und sage okay das ist das und das, wenn Sie wissen was ich meine. (MMK21: 55 - 55)
	Methodic competence	Methodic competence is mentioned.	[0:04:33] TP: Ähm, jetzt müsste ich die Kompetenzen finden. Ich würde auf jeden Fall behaupten, dass es eine Handlungskompetenz ist auf jeden Fall, weil das ein sehr natürliches Setting ist das kann passieren. Ähm, dann Methodenkompetenz (SEL31: 33 - 33)
	Honesty	Ability to be honest in the conversation is mentioned.	[0:06:52] TP: Hm äh the intentional äh for the perception. How ähm my feedback äh to her is äh in fact ähm understood by her because actually if the feedback is not understood äh, the next essay will be the same. And the essay after next it will be the same and it will be something like the vicious circle. But if äh the feedback is properly done and empathetically done ähm, I was also monitoring her emotional reaction äh, because actually it is ähm the ähm on the one hand ähm it is äh important to reflect what you have to reflect.

Code		Explanation	Example
			But on the other hand, it is äh quite good to give not the äh objective for fact as it is, but I mean something like äh the äh just the fact, just the fact without any äh actually (unv.), just as it is. But it is important to äh actually give äh the äh truth which is correctable because äh if I had äh just äh like äh the tons of the trash like I had said 'you have many, many grammar mistakes', 'you have problems with this tile' or 'this essay is quite not good at all', I would destroy her. But ähm with the äh giving of the feedback äh even regarding the contest I think that the contest jury is to be more empathetical. (LAT16: 33 - 33)
	Counseling	Counseling competence is mentioned.	[0:02:12] TP: Hm also ich finde eigentlich, dass diese Feedbackgespräche, also dass ist ja, was man auch wirklich im Lehralltag später machen muss, ähm was man aber in der Uni sehr wenig lernt. Ähm also eigentlich gar nicht lernt. Deshalb finde ich, dass so eine Feedback- beziehungsweise Beratungskompetenz dadurch geschult wird. (SNA28: 33 - 33)
	Separation between person and product	Ability to direct feedback not at the self-level.	[0:03:59] TP: Also ich finde das man als Lehrkraft/ das die zentralste Kompetenz als Lehrkraft eigentlich eine Kommunikationskompetenz ist, weil in jedem Bereich wo wir wirklich als Lehrer in Erscheinung treten kommunizieren wir und ich glaube das ist die zentralste Kompetenz die man haben kann als Lehrer und Feedbackgespräche sind natürlich super wichtige Situationen wo es halt/ wo Kommunikation halt sehr wichtig ist und auch richtige Kommunikation weil man spricht ja über Fehler von einem Menschen, aber es muss halt klar werden das man hier an etwas arbeitet was nicht den Menschen an sich betrifft, sondern quasi eine Leistung eines Menschen und eine Leistung, die man verbessern kann und man nicht den Menschen an sich bewertet und ich glaube da muss man halt äh gut kommunizieren um auch das so ein bisschen rüber zu bringen. (AJB12: 35 - 35)
EFL-specific competences			
	Feedback		
	Feedback in general	General aspects of feedback are mentioned	[0:06:03] TP: Ähm (...) gleich, was ich gerade auch schon gesagt habe: Feedback geben in Bezug auf die Zielsprache, allerdings in

Code		Explanation	Example
			reduzierter Form. Ähm ansonsten würde ich da tatsächlich gar nicht so viele Unterschiede sehen, weil das Gespräch ja auf Deutsch stattgefunden hat, finde ich tatsächlich in dem Zusammenhang auch wichtig, da gibt es ja verschiedene Ansätze, aber ich glaube, dass das gerade bei Feedback wichtig ist, das auch in der Muttersprache zu machen, dass äh Fehler verstanden werden. (ALA02: 39 - 39)
	Feed Forward	The support of the learner (i.e., how to improve) and/or formative function of the feedback conversation is mentioned.	[0:07:51] TP: Mhm, mhm. Hm, da finde ich vor allem den sprachlichen Aspekt, ähm, genau weil das spielt wieder ins Feedback geben rein, einfach so wie, also unabhängig von dem „Wie gebe ich generell Feedback?“, „Wie gebe ich konkretes fachliches Feedback, was auf eine Sprachverbesserung im Prinzip hinzielt?“ und ähm an Beispiel Vokabeln ist es dann auch immer so bedingt: „Okay, ich muss nicht nur Feedback geben, sondern ich muss vielleicht auch so ein bisschen weiterhelfen können, so ein bisschen diese fachliche Perspektive haben. Wie kann sich die Schülerin jetzt auch weiterlernen, weiterverbessern?“ Ähm, und das finde ich gerade als Englischlehrkraft einfach sehr schwierig, weil es, keine Ahnung, bei Vokabeln anfängt sowohl die Wörter an sich als auch die Aussprache wäre noch so ein weiterer Punkt gewesen und wenn ich das ganze jetzt in Englisch, also in der englischen Sprache geführt hätte, ähm, dann wäre auch einfach quasi die, äh, „Welche Vokabeln kann ich in einem Feedbackgespräch nutzen, dass beide Parteien verstanden werden?“ Ähm genau das habe ich jetzt nicht ausprobiert, aber so das könnte ich mir auch noch vorstellen. (ICW08(_M): 51 - 51)
	Dealing with errors & mistakes	Ability to act upon diagnosed errors/mistakes is mentioned.	[0:04:55] TP: Äh, ich denke als Englischlehrkräfte ist es wichtig ähm, dass man weiß welche Fehler häufig auftreten äh und wie man mit den Fehlern dann umgehen kann, also es waren ja jetzt zum Beispiel diese false friends oder die Zeitenfehler und halt so ein bisschen spelling mistakes. (AJB12: 37 - 37)
	Positive Feedback	Ability to mention aspects that the learner has mastered successfully is mentioned.	[0:04:34] TP: Mhm. Ähm, also ich glaube da vor allem auch Motivation, dass ich den Schüler*innen Freude auch gebe am Englisch schreiben, Englisch sprechen, wie auch

Code		Explanation	Example
			immer. Das habe ich versucht, indem ich ihr gesagt habe: „Hey komm, versuch doch nochmal auf Englisch“, aber das hat nicht so ganz so geklappt. Ähm, und dann aber eben auch durch positives Feedback, also es war ja eine 5+ dieser Text, ähm, und trotzdem habe ich gesagt: „Hey, well done! Du hast irgendwie die ganzen Informationen und so, ähm, reingebracht und so.“ Ja. (UMB26_M: 45 - 45)
	Assessment		
	Evaluation	Assessment/evaluation skills are mentioned.	[0:04:26] I: Mhm. Und welche Kompetenzen sind besonders wichtig für Englischlehrkräfte? Das können jetzt die Gleichen sein, aber das können auch jetzt noch andere sein. [0:04:34] TP: Mh, ich würde sagen, dass die Kompetenzen auf jeden Fall ähnlich sind, aber dass dann natürlich auch noch die äh Sprachkompetenzen dazukommen. [0:04:44] I: Mhm. Was [0:04:45] TP: Ja. [0:04:45] I: was genau meinen Sie mit Sprachkompetenzen? [0:04:48] TP: Äh [0:04:48] I: Die Eigenen, oder? [0:04:49] TP: Ja genau, also was heißt meine Eigenen, aber natürlich es können meine eigenen, die müssen natürlich gut sein damit ich überhaupt überprüfen kann, ob und was Mia da schreibt grammatikalisch richtig ist, [0:04:59] I: Mhm. [0:05:00] TP: wie der Satzbau ist, ja. (MWL05: 36 - 45)
	Diagnosis	The diagnosis of textual features and/or the learner level is mentioned.	[0:03:08] TP: Ähm, Fachkompetenzen würde ich jetzt erstmal behaupten, also allgemein auch erstmal zu sehen, wo die Fehler, die sprachlichen Fehler sind jetzt die grammar tenses was da jetzt an Fehlern zum Beispiel war, genau. (SSM23: 44 - 45)
	Correction	Correction/ corrective competence is mentioned.	[0:04:11] TP: Hm, für Englisch würde ich dann eben sagen, dass Korrigieren eines englischen Textes und ähm das Verständnis zwischen ähm/ dass das Feedback auch immer dieses Sprachenlernen miteinbeziehen muss, zum Beispiel habe ich ein Beispiel mit den false friends gegeben, dass ist dann ja das was sich speziell auf Englisch-Deutsch-

Code		Explanation	Example
			Beziehungen bezieht und nicht einfach generell auf/ auf generelles Feedback zum Text oder so (RNM25_M: 47 - 47)
	Learner-Teacher-interaction		
	Emotional-motivational		
	Comfort with the language	The comfort of the learner with regards to the language is mentioned.	TP: ‚Wozu gebe ich Feedback?‘, glaube ich ganz doll, also ‚Gebe ich jetzt inhaltliches oder sprachliches Feedback?‘ Vor allem wenn das dann auch eine äh Schülerin ist, die halt tendenziell nicht so gut in der Sprache ist, wo man vielleicht auch berücksichtigen muss, warum fühlt sie sich nicht so wohl in der Sprache. Ähm keine Ahnung wie das in anderen Fächern aussieht, aber das ist, glaube ich, dann dieses eben ‚Wozu gebe ich jetzt gerade in diesem Moment Feedback?‘, ‚Muss ich jeden kleinen Grammatikfehler da irgendwie gerade ankreiden oder geht es um ein generelles Sprachgefühl, was ich ihr da vermitteln möchte?‘ Ähm das, glaube ich, ist da besonders wichtig und überhaupt dieses sich zu trauen, Sprache oder eine Fremdsprache zu gebrauchen irgendwie. Ähm da, finde ich, kann man in Englisch dann als stellvertretende Fremdsprache ähm, glaube ich, viel falsch machen in dem Sinne, dass es diesen Rückzug dann sonst geben kann, dass man sich gar nicht mehr traut dann vielleicht irgendwas auszuprobieren, weil man das Gefühl hat: ‚Okay, ich mache die ganze Zeit nur irgendwelche Fehler. (EME03_M: 37 - 37)
	Empathy	Ability to understand learner's emotions and empathize is mentioned.	[0:02:45] TP: Achso ähm ja vielleicht so Sachen wie soziale Kompetenzen, also richtiges Zuhören und ähm Empathie. Ich weiß nicht auf was Sie genau. (AAS13: 35 - 35)
	Motivation	Ability to motivate / not demotivate learner is mentioned.	[0:04:34] TP: Mhm. Ähm, also ich glaube da vor allem auch Motivation, dass ich den Schüler*innen Freude auch gebe am Englisch schreiben, Englisch sprechen, wie auch immer. Das habe ich versucht, indem ich ihr gesagt habe: „Hey komm, versuch doch nochmal auf Englisch“, aber das hat nicht so ganz so geklappt. Ähm, und dann aber eben auch durch positives Feedback, also es war ja eine 5+ dieser Text, ähm, und trotzdem habe ich gesagt: „Hey, well done! Du hast irgendwie

Code			Explanation	Example
				die ganzen Informationen und so, ähm, rein-gebracht und so." Ja. (UMB26_M: 45 - 45)
		Social com- petence	General social competence, e.g., ability to appropriately engage in social interactions and see the learner as a person is mentioned.	[0:13:29] TP: Hm also ich würde auf jeden Fall sagen ein wichtiger Teil davon ist Sozialkompetenz, also der Umgang ähm mit Schüler innen an sich. Also es ist einfach wichtig, dass ich nicht nur, wenn ich vor einer ganzen Klasse stehe, dann repräsentiere ich eine andere Ruhe als die, die ich gebe, wenn ich vor einem Schüler einer Schülerin individuell stehe äh, weil das oft intimere Situationen sind ähm und das merkt auch dem Verhalten oft der Schüler innen an. Also wie gesagt, die werden dann eher kleinlicher, ziehen sich etwas zurück, hören mehr zu, werden ruhiger und ich finde das super wichtig, dass Lehrer da auch eine Sozialkompetenz entwickeln: Wie gehe dann damit um? Also das zu händeln ähm. (MJO28: 61 - 61)
		Explanation	Aspects of explanation are mentioned with a focus on teaching English.	[0:01:33] TP: Ähm, Erläuterung von Grammatik nochmal. Ähm, auch, meine ich, wie man, ähm, solche Texte besser verfassen kann im Englischen, ähm, und auch, dass man den Schülern auch eine gewisse Sicherheit im Englischen gibt. (AEM23: 29 - 29)
		Conversational competence		
		Communica- tive compe- tence	Ability to communicate adequately is mentioned.	[0:02:45] TP: Achso ähm ja vielleicht so Sachen wie soziale Kompetenzen, also richtiges Zuhören und ähm Empathie. Ich weiß nicht auf was Sie genau. (AAS13: 35 - 35)
		Subject-specific aspects		
		Use of lan- guage as medium	Use of language as a transmitter/ as medium of instruction adapted to the learners' needs is mentioned. Focus is on the medium (how) of the dialogue.	[0:05:29] TP: Ähm, vielleicht ein gutes Einfühlungsvermögen dafür, inwieweit man sich jetzt doch zu sehr in seinen eigenen englischen Sprachschatz vertieft und dann vielleicht mehr in eine angepasste vereinfachte Sprache gehen, das ist mir auch schwergefallen. (SEL31: 37 - 37)
		Learner ori- entation	The orientation and/or adaptation towards the learner's individual level of competence is addressed (subject-specifically). Focus is on the content (what) of the conversation, not the medium.	[0:04:22] TP: Ja. Also halt ähm dieses, dass man halt ein bisschen sensibel ist ja sprachsensibel zum Beispiel, dass man guckt was ist jetzt wichtig, ist die Grammatik wichtig, die man erfasst oder ähm geht man jetzt eher auf den Inhalt ein, dass man diese Balance ähm halt finden muss in dem Gespräch. Dann halt, dass man auch mit dem/ ja mit

Code			Explanation	Example
				den false friends also, dass man diesem äh Bezug zwischen Muttersprache und äh Fremdsprache äh erkennen muss und äh berücksichtigen muss und dass man halt generell so kulturelle oder ja den Hintergrund von den Schülern und den Leistungsstand irgendwie berücksichtigen muss. (BNP09: 45 - 45)
		Selection of feedback content	The ability to select and/or prioritize topics for the feedback conversation is mentioned.	0:02:24] TP: Boah das ist eine gute Frage. Ähm speziell für Englisch. Ja so einen Text schnell durchzugehen, schnell Fehler zu finden und vielleicht ähm den Schülern nicht alle Fehler wild um die Ohren zu hauen, sondern zu gucken, okay was ist jetzt hier gerade das Schlimmste, was passiert, also das ne, dass man nicht so jeden Fehler rauspicks, weil irgendwann haben die auch einfach keinen Bock mehr. Ich glaube, das ist auch eine Kompetenz, die man über längere Zeit lernen muss. (FLA04: 35 - 35)
		Explanation	TP's ability to explain is mentioned	[0:01:33] TP: Ähm, Erläuterung von Grammatik nochmal. Ähm, auch, meine ich, wie man, ähm, solche Texte besser verfassen kann im Englischen, ähm, und auch, dass man den Schülern auch eine gewisse Sicherheit im Englischen gibt. (AEM23: 29 - 29)
		Subject-related aspects		
		Knowledge	TP mention knowledge aspects.	
		Typical errors / mistakes	Knowledge about typical errors/mistakes from learners is mentioned.	[0:04:55] TP: Äh, ich denke als Englischlehrkräfte ist es wichtig ähm, dass man weiß welche Fehler häufig auftreten äh und wie man mit den Fehlern dann umgehen kann, also es waren ja jetzt zum Beispiel diese false friends oder die Zeitenfehler und halt so ein bisschen spelling mistakes. (AJB12: 37 - 37)
		Specific Issues of German L1 learners of English	Aspects regarding the specifics of German L1 speakers learning English are mentioned.	[0:04:06] TP: Für Englischlehrkräfte war es halt besonders wichtig, ähm so darauf einzugehen, dass ja halt jemand neben einem sitzt, der Englisch nicht wirklich gut besitzt, dass es halt eine Fremdsprache ist, dass wir darauf auch als Lehrer dann quasi achten müssen, dass wir nicht zu sehr Englisch mit denen sprechen so wie ich es am Anfang getan hatte. Ähm und halt auch darauf achten müssen, dass halt Vieles, was im Deutschen geschrieben wird, äh nicht so eins zu eins ins

Code			Explanation	Example
				Englische übersetzt werden kann, weil sonst würde Vieles keinen Sinn ergeben, äh da auch die Satzreihenfolge einfach anders sein soll. Auch Wörter ähm, die Vokabeln waren zum Beispiel halt so immer gesetzt, dass sie eben manchmal ähm quasi das so eins zu eins übersetzt hätte. Sprich äh bekommen hat sie als become übersetzt, was halt äh in dem Satz dann gar keinen Sinn mehr ergeben hat. Also es gibt halt ähm die Vokabeln, auf die man eingehen müsste, das Schreiben an sich äh, auch das Verstehen und das Sprechen. (NRS18: 29 - 29)
		Genre knowledge	Knowledge about the demands of the required text genre (structure, register) is mentioned.	[0:07:02] TP: Ja zum Einen die ja, also sprachliche Kompetenz in dem Sinne, dass es ja viel um Grammatik geht auch ähm um grammatikalische Konstrukte und auch grammatikalische Inhalte, die halt die Schüler zu einem gewissen Grad drauf haben müssten. Ähm ja natürlich geht es da um Vokabular und so weiter ähm. Ja, aber auch um Formales und Strukturelles jetzt ja in Punkto Thematik news report anfertigen. So ja. (SAP01: 45 - 45)
		Cultural context	TP mentions specific knowledge from the cultural context.	[0:05:47] TP: Ähm, also eben, ich glaube vor allen Dingen halt diese Wissensfähigkeit, also Grammatik, das Ganze, was dazugehört beim Texte schreiben. Ähm, ich glaube aber auch ähm die sprachlichen und äh kulturellen Kontexte, die da zum Teil mit drinnen sind. Das war jetzt bei dem Text nicht so stark, aber ähm, wenn man sich ja diesen, ähm, äh, den anderen Artikel durchgelesen hat da waren natürlich gewisse kulturelle Sachen, die es in Deutschland halt so nicht gibt. Ähm, da muss man dann natürlich immer wieder drauf eingehen und das äh auch mit / den Schülern da auch Feedback geben, wenn sie in ihrem Text etwas erwähnen, was halt super Deutsch ist und jemand, der nur Englisch spricht, nicht verstehen könnte. (SCP15: 51 - 51)
		Language		
		General language competence	Aspects of language are mentioned, either general or specific, but without directly referring to knowledge or mastering of the language.	[0:02:59] TP: Ähm vielleicht ähm Sprachkompetenz ähm und das Vokabular, was man da anwendet ähm in einem Gespräch.

Code			Explanation	Example
		Mastering of language	TP's ability to master English in general (e.g. ability to speak and write fluently and correctly; comprehend English) is mentioned.	[0:03:11] TP: Hm naja man als Lehrer oder als Lehrerin muss man ja auf jeden Fall die Sprache irgendwo beherrschen können und halt auch ähm bei dem Text ist mir ja auch aufgefallen, dass ähm sie jetzt ähm viele viele ähm Formen falsch angewendet an und ähm oder ähm Wörter falsch geschrieben hat ganz einfach. Und das muss man erkennen können, finde ich, als Lehrerin oder als Lehrer. (CCF27: 41 - 41)
		Explicit knowledge of language	TP's knowledge of explicit linguistic features of the English language (grammar, lexis, syntax etc.) is mentioned.	[0:08:59] TP: Ähm, also für Englischlehrkräfte ist besonders bei solchen Texten super wichtig, dass wir so eine ähm/ also wir sind ja irgendwo ein sprachliches Vorbild und es muss halt dementsprechend eine sehr hohe sprachliche Richtigkeit vorhanden sein und dementsprechend Wissen bei Grammatik, Textgenre ähm ja stilistische ähm, stilistische oder Rhetorik an sich, was die Schüler zu verschiedenen Textarten halt wissen müssen äh, wichtig Wortschatz, alles was in diesem Feld dann rein ähm ja reinfallen könnte ist halt auf jeden Fall wichtig und ähm halt darüber hinaus zu wissen, dass/ also wir hatten das ja zum Beispiel hier auch, dass zum Beispiel deutsche Strukturen auch oft verwendet worden sind, dass wir auch irgendwo wissen, dass Schüler dazu neigen halt dieses Germanism praktisch attraktiv anzuwenden und da dann halt auch viele Kompetenzen, wie man da Strategien heranzuführen kann. (GBP25: 45 - 45)
		Translation skills	Aspects of translation are mentioned.	[0:05:07] TP: Genau ähm also quasi einmal die Übersetzungsfähigkeit bzw. die Sprachkenntnisse im Allgemeinen ähm ich musste ja verstehen was sie schreibt und was sie mir sagen möchte auch wenn es manchmal ein bisschen schwerer war zu verstehen, weil es eben kein perfektes Englisch war ähm das wird glaube ich ganz gut gefördert auch ähm mein Allgemeinwissen wird irgendwie oder mein Vorwissen wird irgendwie angereizt, ich ähm muss ja irgendwie verstehen was damit gemeint ist und ich muss das irgendwie an das anknüpfen was ich schon im Voraus weiß ähm und didaktisch auch, ich musste ja auch irgendwie Hinweise geben, wie sie denn damit noch mehr lernen kann und was sie daraus mitnehmen kann ähm ja ich denke mal

Code			Explanation	Example
				im Englischen braucht man das irgendwie genauso. (LSG02: 37 - 37)
		Other		
		Pedagogical implementation	Aspects are mentioned regarding the "teaching" of subject matter.	[0:02:36] TP: Ähm, dass man auf jeden Fall die englische Sprache beherrscht, die Grammatik und auch die Sprachkompetenz mitbringt. Ähm und das dann aber auch didaktisch weitergeben kann, also es bringt nichts, wenn man die Kompetenz beherrscht, aber niemanden dabei helfen kann, diese selbst zu erwerben. (BJP10: 39 - 39)
		Fostering oral competences	Fostering oral competence is mentioned.	[0:05:22] TP: Ja, ähm also an sich ja natürlich die ähm/ Was in erster Linie unterschiedlich ist, ist die Sprache. Ähm und die Kompetenz da zu fördern, dass man vielleicht auch nochmal so ein bisschen mündliche Sprachkompetenz entwickelt als ähm Schüler oder Schülerin ähm. In meinem Fall jetzt so im Nachhinein bin ich mir nicht sicher, ob ich das bereue, das auf Deutsch gemacht zu haben oder auf Englisch. Ich glaube, dass das gut war für das erste Gespräch, weil Englisch kriert halt immer eine gewisse Distanz ähm. Deswegen würde ich dann sagen, dass man im Laufe der Feedbackgespräche, wenn der Text besser wird, das Gespräch auch ein bisschen konstruktiver gestalten kann im Sinne der Sprache, dass man vielleicht die jetzt in dem Fall die Schülerin nicht ähm direkt überfordert, weil man gibt ihr auch eine Korrektur und dann noch ähm so zu überfordern, dass sie nicht alles versteht äh. Ich glaube, das ist ein Prozess, der sich langsam entwickeln muss ähm (...). Ja Sprachkompetenz würde ich jetzt einfach mal stehen lassen. Da gehört ja viel zu. (MMM07: 33 - 33)
		General knowledge	General knowledge is mentioned.	[0:05:07] TP: Genau ähm also quasi einmal die Übersetzungsfähigkeit bzw. die Sprachkenntnisse im Allgemeinen ähm ich musste ja verstehen was sie schreibt und was sie mir sagen möchte auch wenn es manchmal ein bisschen schwerer war zu verstehen, weil es eben kein perfektes Englisch war ähm das wird glaube ich ganz gut gefördert auch ähm mein Allgemeinwissen wird irgendwie oder mein Vorwissen wird irgendwie angereizt, ich ähm muss ja irgendwie verstehen was damit gemeint ist und ich muss das irgendwie an das anknüpfen was ich schon im Voraus weiß

Code		Explanation	Example
			ähm und didaktisch auch, ich musste ja auch irgendwie Hinweise geben, wie sie denn damit noch mehr lernen kann und was sie daraus mitnehmen kann ähm ja ich denke mal im Englischen braucht man das irgendwie genauso. (LSG02: 37 - 37)
	No amendments to competences for teachers in general.	No additional competences specific for English teachers are mentioned.	[0:05:46] TP: Hm ja eigentlich alles, was ich gerade gesagt habe auch für Englischlehrkräfte, ne. Ja. (BAH29: 47 - 47)
	Providing safety	Safety in the meaning of "comfort" or in the meaning of speaking English well reliably.	[0:01:33] TP: Ähm, Erläuterung von Grammatik nochmal. Ähm, auch, meine ich, wie man, ähm, solche Texte besser verfassen kann im Englischen, ähm, und auch, dass man den Schülern auch eine gewisse Sicherheit im Englischen gibt. (AEM23: 29 - 29)

Use

Code		Explanation	Example		
Teacher Education in general					
	Rather positive				
		Positive with limitations			
		Limitations			
			Constructive Alignment/ Preparation for the exam	Preparation for such an assessment format is mentioned.	[0:07:04] TP: Ja ähm ich hatte ja gesagt, ich fände das sehr sinnvoll. Natürlich müsste man dann auch besser drauf vorbereitet werden durch die ja Prüfungsordnung oder die Module oder sowas. Ähm aber das fände ich sehr sinnvoll, also bereits im Bachelor sowas anzusetzen. (KMP05: 59 - 59)
			Necessity of evaluation criteria	The necessity of evaluation criteria for such an assessment is mentioned.	[0:07:13] TP: Hm ich kann mir das ganz gut vorstellen auf jeden Fall, weil das ja wirklich Sachen sind, mit denen ich im Lehreralltag konfrontiert bin. Und ich glaube schon, dass man das ganz gut abprüfen kann. Also klar, man müsste dann irgendwie ähm ja irgendwie halt Aspekte entwickeln, woran bewerte ich das wie, aber ich kann mir das gut vorstellen. (BNW18: 59 - 59)
			Actors' abilities	The dependence on the actors' abilities is mentioned.	[0:07:45] TP: Also ich könnte es mir ehrlich gesagt schon vorstellen ähm. Wenn es so gut gemacht ist wie hier, dann auf jeden Fall, also wirklich ähm. Ich stelle mir jetzt vor, wenn die Person, die einem gegenüber sitzt ähm das Ganze schauspielerisch in Anführungszeichen nicht so gut rübergebracht hätte, dann wäre es, glaube ich, schon ein Ticken schwieriger, sich wirklich darauf einzulassen und ähm ja diese Spannung, die ich vorhin, oder diesen Spannungsbogen, den ich vorhin beschrieben habe, den aufrecht zu halten. Ähm aber wenn sowohl die Inhalte, die jetzt gepasst haben als auch dann die in Anführungszeichen schauspielerischen Rahmenbedingungen, wenn die passen, dann wüsste ich nicht, was dem Wege stehen sollte. Ja. (SAP01: 47 - 47)
			Organization of test	The complexity of the organization (setup, actors etc.) is mentioned.	[0:04:24] TP: Ich könnte mir das sehr gut vorstellen als Ersatz für eine mündliche Prüfung zum Beispiel. Es wäre vielleicht ein bisschen schwerer umzusetzen wegen so einer, weil die Situation an sich sehr aufwands- äh weil da ein hoher Aufwand hinter steckt. Aber ich glaube, dass das einem mehr bringen würde als wenn ich zum Beispiel vorne stehe und

					vier Dozenten erkläre, was die Geschichte vom Englisch ist oder sowas. (AFS29: 33 - 33)
				Comparability of TPs	The difficulty of compare TPs' behavior is mentioned. [0:02:52] TP: Finde ich schwierig, also ich fände es schon. Ich fände es besser als eine Klausur, wesentlich besser als eine Hausarbeit, aber gleichzeitig finde ich es schwierig, denn ähm egal wie gut die Schauspieler sind und egal wie oft man durchgeht in welche Richtung das Gespräch gehen soll, wenn ich in eine ganz andere Richtung steuere, ähm dann irgendwie können die da ja auch nicht mehr gegen halten. Das/ Also ich find bei einer Klausur kriegen halt hundert Leute die gleiche Klausur und auch wenn ich hier die gleiche Aufgabe kriege, kann das ganz anders sein. Und deshalb finde ich das schwierig, dann das zu bewerten und das mit anderen Prüfenden praktisch zu vergleichen. Aber an sich das Format finde ich toll, nur ich kann mir nicht vorstellen, ähm wie man es macht, dass es für alle identisch ist. [0:03:24] I: Hm also für Sie ist die Vergleichbarkeit so ein bisschen ein Problem. [0:03:27] TP: Ja tatsächlich, aber an sich ist es super. Also ich finde es toll, wenn das ein Prüfungsformat wäre. (FLA04: 37 - 39)
				Only as part of a larger assessment context	Using the assessment only as one part of a larger assessment context is mentioned. [0:05:55] TP: Ähm, ich könnte es mir sehr gut vorstellen, aber eher als, nicht als kompletter Teil einer Prüfung, sondern eher sozusagen entweder als Extrateil, um eine potentielle Note aufzubessern oder als Teil einer Prüfung, dass man die Hälfte der Prüfung theoretisch hat und dass man das theoretische praktisch danach anwendet und an sich finde ich sowas praktisches gut, weil das hat ja auch dann mit den Praktika zu tun, die man noch machen muss, weil dann kann man das sozusagen einmal durchsimulieren, wie das überhaupt ist, ob man da überhaupt gut drin ist und, ähm, was ich jetzt auch durch dieses erste Semester / man, man lernt viel theoretisches, aber kaum wie man das Theoretische wirklich anwendet und damit könnte das, wenn man das schon von Anfang an mit einbringt, sehr hilfreich sein. (BJL19: 39 - 39)
				Available information on the learner	[0:05:11] TP: Ich könnte es mir schon vorstellen, weil ähm ja. Ja, also ich denke schon, dass es eine gute Sache ist, weil es halt sehr praxisnah ist und sowas kann man einfach nicht in der Theorie äh üben und deswegen

				braucht man da schon so eine Simulation für, aber natürlich hat die Simulation auch Begrenzungen ähm, wie zum Beispiel halt das Zeitlimit, da ist halt die Frage, ob ein Lehrer/der würde natürlich mehr Informationen über den Schüler haben und der wird vielleicht auch mehr Zeit haben da reinzugucken und sich zu überlegen, was er sagt. Deswegen ähm ist vielleicht dieser Zeitpunkt oder das Zeitfenster so ein Limit, aber sonst kann ich es mir gut als Prüfungsformat vorstellen. (BNP09: 47 - 47)
	Similarity to classroom observation	The assessment is compared to classroom observation.		[0:05:59] TP: Also ich kann es mir vorstellen. Das ist schon eine Situation, also es hat sich gerade auch teilweise wie eine Prüfung angefühlt, muss ich sagen. Ich kann es mir sehr gut vorstellen, aber ich wüsste nicht, worin sich das von einem ganz normalen Unterrichtsbesuch unterscheidet. Also warum man das jetzt unbedingt machen sollte. (BAH29: 49 - 49)
	Time constraints	Time constraints are mentioned as a limitation.		[0:05:11] TP: Ich könnte es mir schon vorstellen, weil ähm ja. Ja, also ich denke schon, dass es eine gute Sache ist, weil es halt sehr praxisnah ist und sowas kann man einfach nicht in der Theorie äh üben und deswegen braucht man da schon so eine Simulation für, aber natürlich hat die Simulation auch Begrenzungen ähm, wie zum Beispiel halt das Zeitlimit, da ist halt die Frage, ob ein Lehrer/der würde natürlich mehr Informationen über den Schüler haben und der wird vielleicht auch mehr Zeit haben da reinzugucken und sich zu überlegen, was er sagt. Deswegen ähm ist vielleicht dieser Zeitpunkt oder das Zeitfenster so ein Limit, aber sonst kann ich es mir gut als Prüfungsformat vorstellen. (BNP09: 47 - 47)
	Anxiety	Anxiety is mentioned as a potential obstacle for TPs to empathize with the learner.		[0:05:03] TP: Ich kann mir das gut vorstellen, obwohl ich auch weiß von vielen ähm Mitstudierenden, dass sie sehr nervös sind bei mündlichen Prüfungen. Das war für mich nie ein Problem, da habe ich immer sehr positive Erfahrungen gehabt. Und das gerade, wenn dann bei den Studierenden viele Emotionen auch mit reinkommen, dass sie halt sich bewusster sind, das ist eine Prüfung und die selber dann nervös sind, dann ist man denke ich nicht so aufnahmefähig, was so dieses Emotionale angeht, dass man dann auch als Studierender darauf achtet, okay, in einer

						echten Situation versuche ich hier den Schüler zu beruhigen, wenn man erstmal versucht sich selber zu beruhigen in der Situation. (MAP27: 43 - 43)
				Specifications despite limitations		
				Higher value for professionalization		
				Better preparation for the profession	The assessment is regarded as an approach to better prepare for the later profession.	[0:09:58] TP: Ich kann es mir an sich sehr gut vorstellen, weil ich habe das Gefühl, das hat viel mehr als sehr sehr viel andere Uniinhalte damit eben zutun, was später mal auf uns zukommt, und Dinge, um die es halt wirklich geht, weil grade in solchen Situationen kommt zum Vorschein einmal die soziale Kompetenz, aber auch wirklich die Kompetenz als Lehrkraft: ‚Inwiefern kann ich die mir anvertrauten Schülerinnen und Schüler wirklich anleiten und wirklich irgendwie weiterbringen in ihrem Lernprozess und auf den nächsten Schritt hin?‘ Ähm und ich habe das Gefühl, das ist was, was im Studium sehr sehr viel untergeht indem, dass man vor jeder Prüfung so ein bisschen da sitzt: ‚Okay, was hat das jetzt mit dem zu tun, was ich mal machen werde?‘ Das finde ich sehr viel näher an unserem Alltag dran. Ich frage mich ein bisschen, glaube ich, ähm was das beurteilen dann vielleicht nachher angeht, die Auswertung das halte ich für schwierig, weil das ist ja trotzdem in der Ansicht der Person, die das dann mal auswertet vielleicht: ‚Okay, das halte ich jetzt für eine tolle Kompetenz und da sehe ich die Person.‘ Da also inwiefern man dann sagt: ‚Okay, das hat die Person jetzt besonders gut gemacht‘ Kann man natürlich verschiedene Bewertungsskalen und -bögen entwerfen zu, aber das stelle ich mir schon auch sehr schwierig vor oder auch zeitintensiver als jetzt: ‚Ich lege euch mal eben eine Klausur dahin, die ihr schreibt.‘ (EME03_M: 39 - 39)
				Practice-orientation	The assessment is describe as being practice-oriented.	[0:07:53] TP: Ja, gerade so mit Hinblick auf den Bachelor war bei uns viel von den Prüfungen sehr gleich, egal welches Lehramt man hat, das war alles immer gleich und ich finde es schon gut, wenn dann mal so etwas Praktisches ist. (SJH28, Pos. 63)

				Replacement of oral exams	Using the assessment as a replacement of oral exams is mentioned.	[0:02:31] TP: Also ich kann mir zum Beispiel vorstellen, dass man das im Sinne eines Seminars macht, wo dann Rollenspiel immer gemacht werden und dass es eventuell auch als mündliche Prüfung eine Prüfungsleistung wird. (GJS22: 39 - 39)
				Increased motivation	More motivation for such an assessment is mentioned.	[0:08:16] TP: Also ich fände das eine sehr, ich hätte da, also rein subjektiv als Student hätte ich da mehr Bock drauf als auf eine Fachdidaktik, also Klausur sage ich mal. (MJH23: 43 - 43)
				Positive without limitations		
				Generally rather positive	The format is generally evaluated positively but no further specifications are made.	[0:05:47] I: Mhm. Können Sie sich den Einsatz von solchen Rollenspielen im Lehramtsstudium vorstellen? [0:05:52] TP: Definitiv und ich finde es auch wichtig. (SEL31: 38 - 39)
				Specifications		
				Higher value for professionalization		
				Better preparation for the profession	The assessment is regarded as an approach to better prepare for the later profession.	[0:03:13] TP: Hm ich kann mir das ähm, ich habe das ja vorhin schon erwähnt, sehr gut als Prüfung äh vorstellen. Da man vor allem auch nicht so in das kalte Wasser geschmissen wird, wenn man dann wirklich in die Situation kommt. Und äh deswegen würde ich mir auch gerne solche Prüfungen als mündliche Prüfung zum Beispiel ähm wünschen. Und äh dann auch nicht nur als Leistung so als qualifizierte Teilnahme, sondern als richtige Prüfung, damit man das auch richtig ernst nimmt. Und ja. [0:03:41] I: Meinen Sie mit ‚in das kalte Wasser‘ dann in den zukünftigen Beruf oder? [0:03:46] TP: Ja, genau (AAS13: 39 - 41)
				Personal development	Using the assessment for one's own personal development, e.g., reflectively is mentioned.	[0:07:01] TP: Also ich kann mir das eigentlich sehr gut vorstellen, einfach weil es ist halt sehr viel lebensnaher und ich finde manchmal ist halt nur diese Wissensabfrage und halt auch nicht nur Grammatik abfragen, was ja schlussendlich das meiste ist im Lehrberuf irgendwo durch. Ähm, und ich glaube, dass würde auch einfach helfen für sich nochmal zu reflektieren, möchte ich diesen Beruf erlernen? Bin ich gut in dem und wenn ich nicht so gut in einem Bereich, wie kann ich mich

					verbessern, in dem Bereich? Viel mehr als, wenn ich bei einer Klausur Häkchen setzen muss, welche Aussage ist jetzt richtig zu irgendeinem Thema. (SCP15: 65 - 65)
				Washback	The assessment is mentioned as being a reason for acquiring the necessary competences. [0:06:37] TP: Hm die Idee an sich finde ich super, also dass man dann vorher auch mal diese Konfrontation hatte mit: Wie gebe ich Feedback? Äh worauf muss ich achten beim Feedbackgeben? Wie muss ich mit der Schülerin umgehen? Wie benimmt sich die Schülerin? Also, dass man diese ganzen ähm Punkt erstmal richtig einordnet und dann natürlich eine Person hat, die das schon im besten Fall sehr lange macht und dann auch nochmal Verbesserungsvorschläge zu kriegen, wie man in bestimmten Situationen richtig verhalten könnte. Also in Prüfungssituationen finde ich das eigentlich gar nicht mal so schlecht, weil ja dann doch direkt das gefördert wird, was wir später auch brauchen. (ADH08: 49 - 49)
				Relevance of feedback	The relevance of feedback as a key aspect of the profession is mentioned. 0:03:19] TP: Ich kann mir das sehr gut vorstellen und zwar deshalb, weil ich das Gefühl habe, dass sowas im Lehramtsstudium völlig untergeht, man es aber am Ende trotzdem irgendwie können muss. Man muss ähm mit Schülern reden können. Man muss auch mit Eltern reden können. Und keiner bereitet einen so wirklich darauf vor. Ähm deshalb finde ich, ist es sehr sinnvoll, dass mal als Prüfungsleistung anzubringen, weil sonst sind ja mündliche Prüfungen oft einfach, ne, das ist dann so ein Frage-Antwort-Spiel und ähm gefühlt wird halt einfach irgendwelches Wissen abgefragt, was man sich vorher über die Wochen irgendwie exzessiv angeeignet hat und danach ist dann irgendwie alles wieder weg. Und so ein Feedback- oder Beratungsgespräch ähm ist ja auf jeden Fall ja halt irgendwas, was man halt später auch braucht. Also finde ich das schon sinnvoll. (SNA28: 37 - 37)
				Practice-orientation	The assessment is describe as being practice-oriented. [0:04:37] TP: Hm das kann ich mir tatsächlich sehr gut vorstellen und ich würde das auch sehr begrüßen, weil ich finde, das ist eine sehr praxisnahe Prüfung. Und es hatte jetzt ja auch ein bisschen einen Prüfungscharakter. Ich konnte mich kurz vorbereiten und dann fand eben die Simulation statt. Und ähm doch das kann ich mir tatsächlich sehr gut vorstellen. Und man kann einerseits prüfen

		ähm, wie gut jemand eben im Englischen ist, wie schnell er die Fehler erfassen kann und gleichzeitig aber auch, wie er das dann gegenüber den Schülern rüberbringt. Und ich finde das kommt halt im Studium jetzt häufig zu kurz, bei einer Prüfung. Und das finde ich sehr sehr praxisnah und ich kann mir das sehr gut vorstellen. (HKV21: 31 - 31)
	Application of competences	The possibility of applying competences/knowledge is mentioned. [0:03:20] TP: Ähm ja ich könnte mir das schon ziemlich gut vorstellen, weil es ja auch ähm eher praktisch ähm anliegt und ähm dadurch kann man seine ähm Kompetenzen, die man in der Uni so ähm gelernt hat, besser anwenden finde ich. [0:03:38] I: Hm und welche Kompetenzen sind das, die man dann anwendet? [0:03:40] TP: Jetzt zum Beispiel die Sprachkompetenz oder ähm auch das, was wir jetzt hatten ähm, wie man den Schüler anregen kann ähm, ja Reflektionen zu machen und ähm mehr fällt mir jetzt auch nicht ein. (MMD31: 43 - 45)
	Comparison to other professions	Other professions that use simulation-based assessment formats are mentioned. [0:06:54] TP: Auf jeden Fall, habe ich auch schon öfters drüber nachgedacht. Ich meine in der Poli/ bei der Polizei wird es ja genauso gemacht, dass Situationen nachgestellt werden und warum nicht auch im Lehramtsstudium? Also ich denke mal was authentischeres bekommt man an der Uni so nicht hin und diese Prüfungsform ist auf jeden Fall deutlich näher an dem späteren Beruf als ähm als es eine Klausur jemals sein kann. (ALA02: 41 - 41)
	Test addresses gap/lack of practice in teacher education at university	The test is described as filling a gap in teacher education programs. [0:05:19] TP: Ich fände das super, wenn sowas passieren würde äh also ich kenne das auch schon so aus der Medizin, dass man das da auch macht und ich glaube, dass das auch nochmal den Fokus auf eine Kompetenz legt, die zurzeit in der Universität eigentlich gar nicht geprüft wird, also sowas zwischenmenschliches, kommunikatives was dann aber doch glaube ich super viele Schülerinnen und Schüler als extrem wichtig einschätzen. Also wenn man jetzt irgendwelche Schülerinnen und Schüler fragt, was ist denn ein guter Lehrer, dann sind das meistens die Lehrer, die ähm ja die Schüler als Menschen gut behandeln und die juckt es dann eigentlich gar nicht ja wie gut die jetzt in der Fachwissenschaft sind oder äh ja wenn die jetzt so

					total fachdidaktisch tollen Unterricht machen, aber wenn es dann an diesen zwischenmenschlichen Kompetenzen hapert, ist das ja für die Schülerinnen und Schüler am prägnantesten. (AJB12: 39 - 39)
				Replacement of oral exams	Using the assessment as a replacement of oral exams is mentioned. Hm ich kann mir das ähm, ich habe das ja vorhin schon erwähnt, sehr gut als Prüfung äh vorstellen. Da man vor allem auch nicht so in das kalte Wasser geschmissen wird, wenn man dann wirklich in die Situation kommt. Und äh deswegen würde ich mir auch gerne solche Prüfungen als mündliche Prüfung zum Beispiel ähm wünschen. Und äh dann auch nicht nur als Leistung so als qualifizierte Teilnahme, sondern als richtige Prüfung, damit man das auch richtig ernst nimmt. Und ja. (AAS13: 39 - 39)
				Rather negative	
				Generally rather negative	The assessment is seen rather negatively without providing specifications. n.a.
				Specifications	
				Difficult standardization	The difficulty of standardizing the assessment is mentioned. [0:05:09] I: Hm und das ist jetzt so ein bisschen die Richtung: Wie kann man das zum Üben nutzen? Oder können Sie sich das vorstellen, dass man das wirklich als Prüfung einsetzt? [0:05:17] TP: Als Prüfung stelle ich es mir schwierig vor, weil die Rolle, die gespielt wird von der Person, die die Rolle spielt, muss halt immer gleich gespielt werden. Und ich glaube, ich bin mir nicht sicher, ob diese Möglichkeit eins zu eins immer da ist. Ähm ja allerdings könnte das auch irgendwie den Reiz der Sache ausmachen. Ähm die Frage ist halt, okay, wie/ Was sind so die Bewertungskriterien? Wie können die unabhängig von dem Rollenspiel der anderen Person quasi übersetzt werden und bewertet werden? [0:05:53] I: Also Ihrer Meinung nach ist das sehr sehr schwierig, diese Rolle zu standardisieren? [0:05:58] TP: Ja genau. (EEH01: 38 - 41)
				Complexity of evaluation criteria	The complexity of evaluation criteria is mentioned. [0:04:26] TP: Hm (...) jetzt gerade kann ich es mir noch nicht so gut vorstellen, ähm einfach weil ich das so als Prüfungsformat nicht kenne und da dann wirklich sehr stark ähm/ Die Bewertungskriterien müssen ja wirklich

				sehr stark ausgefächert sein, dass das auch wirklich ja fair bewertet werden kann und dass es bei jedem die Situation gleich ist, dass die ähm Ansprechpartner, die Simulationspartner dann auch bei jeder Simulation gleich reagieren und dass es halt identisch bleibt. (SLG21_M, Pos. 43)
		Lack of knowledge about the learner	Lack of knowledge about the learner is mentioned.	[0:05:40] TP: Ach im Lehramtsstudium. Ähm ich finde es schwierig, einerseits fände ich das gut, wenn es zum Beispiel Teil einer AQT wäre, wenn man das mal machen würde, aber als Prüfungsleistung fände ich das schwierig. Dadurch, dass was ich vorhin auch gesagt habe, wenn man wirklich eine Schülerin hat oder einen Schüler hat, auf lange Sicht kann man sich da, glaube ich besser mithineinversetzen als wenn man das jetzt gerade einmal hat und man nicht genau weiß, was sind jetzt eigentlich die genauen Fähigkeiten, weil man ja so einen groben Überblick darüber hat, aber nicht genau und detailliert und ein Gefühl dafür hat. Und des Weiteren im Lehramtsstudium bis auf wo es die Praxisphasen gibt, aber ansonsten macht man ja mehr theoretisches ähm, deswegen würde ich dann sagen, wenn dann sollte es dann einen Kurs geben, wo man auch wirklich mehr auf Theorie achtet und das dann ein bisschen ausprobiert, wenn man das zu einer PL machen will. Aber ich würde es eher als QT. (APD10: 39 - 39)
EFL Teacher Education				
	Rather negative			
	Limitation: Lack of knowledge about learners' language competence	The lack of knowledge about the learner's language competence as a limitation for adaptation is addressed.		[0:06:00] I: Hm okay. Und kannst du dir vorstellen, dass man sowas dann auch wirklich im Englischlehramtsstudium einsetzt oder siehst du das ähnlich kritisch oder vielleicht anders? [0:06:11] TP: Ne, das ist eigentlich ähnlich dann. Jetzt so im Nachhinein betrachtet, ich habe mein Feedback auf Deutsch gemacht. Das würde ich dann auch in der Praxis dann an den Schüler anpassen äh je nach dem wie der wie das Level der Sprache bei dem ist. Ähm genau, aber sonst wäre das im Englischen auch relevant und ja das wäre halt so ein besonderer Punkt, wo man nochmal drauf achten muss in welcher Sprache man das dann macht.

			[0:06:39] I: Und das konnte man hier in der Simulation nicht machen? [0:06:42] TP: Hm doch (lachen). [0:06:43] I: (lachen) [0:06:44] TP: Also aber wenn man das im Unterricht macht oder in der Praxis macht, dann hat man ja schon eine größere Bindung zu den Schüler innen, dass man dann auch weiß, ob die dann auch auf Englisch das verstehen, was sie dann, was man grad vermitteln möchte oder ob die dann vielleicht doch eher ein bisschen Hilfe bräuchten mit dem Deutschen dann. [0:07:03] I: Und das kann man in der Simulation nicht so gut nachstellen, weil man das nicht so alles weiß? [0:07:07] TP: Genau, ja.(SLG21_M: 50 - 57)
		Rather positive	
		Comparable to other subjects	TP see no substantial difference in applying the concept to the subject of EFL compared to other subjects. [0:13:39] I: Und wie ist das speziell im Englischen? Also genauso wie alle anderen oder besonders gut, besonders schlecht? Oder? [0:13:46] TP: Ne, genauso gut eigentlich. Also vielleicht ähm aber auch besonders, weil das irgendwie andere Fehler sind, also es ist ja auch so eine andere Sprache, wo viele Schüler vielleicht so der Meinung sind: ‚Ich kann das einfach nicht.‘ Obwohl das hat man auch in jedem Fach wahrscheinlich. Schwierig irgendwie, also ich glaube, es ist in jedem Fach machbar auf jeden Fall. Ähm und ja genau. (BNW18: 88 - 89)
		Positive with limitations	
		Limitations	
		Age of the actor generally	The age of the actor is mentioned as a limiting factor generally [0:12:45] TP: Ja, das Ding ist, das müsste man ja auf jeden Fall mit einer jüngeren Person machen, so mit einem Dozenten wäre es jetzt glaube ich sehr unauthentisch gewesen. Hm, ja aus so einer Prüfungsperspektive das jetzt so zu betrachten (lacht), ja also es ist ja jetzt auch im Englischstudium, ja genau das ist ja jetzt nicht Biwi oder so, genau. Ähm, puh, ja also ich kann nur sagen, ich bin mega dafür, dass man mehr so Praxissituationen nachstellt, also [0:13:30] I: (unv.) [0:13:31] TP: Weil Lehrer ist ja auch einfach ein, man muss, also ich finde es sehr gut,

				dass man dann im Bachelor schon am Anfang so Situationen nachstellt und dann relativ gut auch nochmal neben den Praktika die man hat ähm und also allgemein Feedbackgespräche sehe ich auch als sehr relevant an so darum habe ich den Kurs auch genommen so, ich hatte da auch die Wahl, ich konnte mir das auch aussuchen ähm, Feedbackgespräche sind auch wichtig auf jeden Fall, sind auch eine der wenigen Situationen, wo man überhaupt mit den Schülern so interagiert ähm. (MMK21, Pos. 89-91)
	Only in combination with a theoretical part	The test should be amended with another theoretically-oriented part.		[0:14:32] TP: Ähm, ich fände es cool, wenn es so eine Mischung aus Theorie und Praxis geben würde, dass man vorher auch vielleicht eine Teilnote für die Theorie kriegt und eine Teilnote für die Praxis, weil es kann natürlich immer sein, dass man dann in dieser Situation irgendwie ein Black out hat oder dass es nicht richtig funktioniert oder keine Ahnung, es gibt ja viele Faktoren in diesem/ in dieser Situation. [0:14:54] I: Mhm, ja. [0:14:55] TP: Und ich denke dann wäre der Fokus auch noch ein bisschen mehr auf der Theorie und man müsste das auch vertiefen. (SJH28, Pos. 123-125)
	Unsuitable for beginners	The test requires some advancement in the teacher education program and should be implemented at a later stage, e.g., after internships.		[0:12:23] TP: Mhm. Also es müsste auf jeden Fall aus meiner Sicht jetzt nach dem Praxissemester sein, also ich denke es ist schon wichtig, dass man ähm bereits gewisse Praxiserfahrung da hat, um in so eine Prüfung auch gehen zu können ähm ich glaube nicht dass es viel Sinn machen würde das vorher zu machen vielleicht ja weil es dann eine Situation ist die für einen selber dann eben vielleicht nicht so authentisch ist, weil man es eben noch nicht so hatte. Also ähm ich denke nach dem Praxissemester könnte man das machen, wenn man eben entsprechend darauf vorbereitet wird und wie gesagt, wenn auch die entsprechenden Raster dafür äh so konzipiert werden äh da könnte ich mir das dann schon vorstellen das es eben ähnlich wie jetzt gerade auch in dem in der Simulation ähm im Fach Englisch ähm gemacht wird. (FCP15, Pos. 83)
	Structural barriers	The actual implementation is critically		[0:09:31] TP: Hm vorstellen jetzt gerade kann ich es mir nicht, aber es würde mich freuen, wenn es passieren würde. Also ähm ich weiß

				discussed from a systemic point of view.	nicht, ob keine Ahnung, ob die Uni bereit dafür ist sozusagen, aber ich glaube, da gibt es viele Leute, die dagegen lenken würden so vom Gefühl her, aber ich glaube, es wäre super interessant und auch sehr hilfreich, es zu machen. Also ich finde es toll. (EEH01, Pos. 57)
			Specifications despite limitations		
			Focus on language	The focus on language is highlighted.	[0:09:59] TP: Ich würde sagen besonders die Sprache und Sprachsituation, also ganz häufig wird man ja auch da quasi konfrontiert mit denen man vielleicht in anderen ja Unterrichtseinheiten und ähnlichen Dingen nicht wirklich konfrontiert wird. Also wenn jetzt jemand keine Ahnung plötzlich eine Frage hat zu irgendwelchen Sprachen und ähnlichen Sachen, also man muss halt so schnell reagieren auf die Person. (EEH01, Pos. 59)
			Importance of feedback	The importance of feedback is highlighted.	[0:13:31] TP: Weil Lehrer ist ja auch einfach ein, man muss, also ich finde es sehr gut, dass man dann im Bachelor schon am Anfang so Situationen nachstellt und dann relativ gut auch nochmal neben den Praktika die man hat ähm und also allgemein Feedbackgespräche sehe ich auch als sehr relevant an so darum habe ich den Kurs auch genommen so, ich hatte da auch die Wahl, ich konnte mir das auch aussuchen ähm, Feedbackgespräche sind auch wichtig auf jeden Fall, sind auch eine der wenigen Situationen, wo man überhaupt mit den Schülern so interagiert ähm. (MMK21, Pos. 91)
			Positive without limitations		
			Specifications		
			Focus on language		
			Language in general	The issue of language is highlighted.	[0:03:44] I: Mhm, okay. Und könnten Sie sich vorstellen, dass man so eine Prüfung tatsächlich dann auch im Englischlehramtsstudium einsetzt? [0:03:50] TP: Ja. [0:03:51] I: Mhm und warum für Englisch besonders oder warum nicht besonders? [0:03:55] TP: Weil man ja gerade in Englisch viel mit der Sprache arbeitet und auch Feedback geben muss. (GJS22: 54 - 57)

				TPs' appropriate use of language	The demand to use the English language in an interaction with the learner is highlighted positively.	[0:07:21] TP: Also (...) ähm gerade im Englischstudium fände ich das, glaube ich, gar nicht so schlecht, weil wir dann ähm gerade ähm, also was die Schauspielerin/ Ich will jetzt immer Schülerin sagen. Ähm wie die Schauspielerin das gemacht hat: ‚Darf ich lieber auf Deutsch reden?‘ Ich habe/ Ich bin natürlich direkt darauf eingegangen. Ich habe gesagt: ‚Klar, können wir machen.‘ Aber war das so richtig? Ne, dass man das dann vielleicht lernt und ich finde in Fremdsprachen muss man echt ein bisschen gucken, wie man dann mit den Schülerin interagiert. (ADH08: 51 - 51)
				Learners' accurate use of language	The difficulty to use the English language correctly is highlighted as feedback is especially necessary here.	[0:08:42] TP: Gute Frage. Im (...) Ich sage es mal so: Sprachen lassen vielmehr Spielraum zu ähm auch Spielraum für Fehler. Also was heißt, also ich sage jetzt mal ein naturwissenschaftliches Fach da gibt es meistens nur eine richtige Antwort zu den ganzen falschen Antworten. Aber ähm auch in Sprachen da kann man auch äh zwar richtig liegen, aber man hat immer noch vielleicht einen falschen Ausdruck gewählt, also das ja genau das da muss man dann ganz oft spezifisch werden konkret werden. Also ich finde ja und Englisch ja hm. [0:09:30] I: Also würden Sie sagen, es bietet sich an oder eher nicht oder? [0:09:33] TP: Es bietet sich auf jeden Fall an. (AOS13, Pos. 45-47)
				Focus on feedback	The focus on feedback is highlighted.	0:10:51] TP: Ähm, ja gerade/ also da kann ich es mir auf jeden Fall vorstellen, weil/ also mein Zweitfach ist Erziehungswissenschaften also Pädagogik und da wird eher so/ also so über Feedbackkulturen, wertschätzende Feedbackkulturen noch eher gesprochen als wirklich im Englischlehramt und deswegen oder im Fach Englisch und ich finde, dass da eigentlich wirklich noch viel wichtiger, weil es um eine Fremdsprache geht und man da, gerade weil es noch ein Hauptfach ist und da wirklich von der Fünften bis zur 13. Klasse dann halt oder in der Q2 äh oder in der Zehnten, je nach dem wann die Schüler dann abgehen und dass man da einfach wirklich ähm ich sage mal ganz vorsichtig viel kaputt machen kann, wenn man ein falsches Feedback gibt oder ein nicht wertschätzendes oder ja ein herabwürdigendes Feedback, genau. (GBP25: 49 - 49)

				Focus on English language didactics	The implementation in the didactic-focused portions of the EFL teacher education program is emphasized (in comparison to, e.g., literature courses)	[0:11:10] TP: Ähm, also ich denke einsetzbar ist es auf jeden Fall. Das hatte ich ja schon gesagt. Ähm, und ich denke vor allen Dingen könnte man das im Bereich der Didaktik noch miteinbeziehen ähm, weil im Bereich Didaktik wird in Englisch doch recht viel einfach Hausarbeiten geschrieben und ich glaube vor allen Dingen in dem Bereich wäre es gut die Sachen, die man halt auch dann teilweise dort gelernt hat, nochmal tatsächlich in die Praxis umzusetzen und zu testen, habe ich die Sachen, die wir da auch gemacht haben, tatsächlich aufgenommen und kann die auch ins reale Leben umsetzen? Also ich denke im Bereich Didaktik wäre das. (SCP15, Pos. 81)
				Different situations	TP mention different scenarios that could also be used as a performance-based assessment.	[0:08:08] TP: Ja, auf jeden Fall. Also besonders äh ich habe jetzt diesen Test gesehen, da wurde auch viel abgefragt. Wie kann man Aussprache verbessern, dass man halt für so was gezielt mal so eine Prüfung macht oder ein Modul irgendwie. Wie kann ich Schüler anleiten ähm ja, die Aussprache zu verbessern? Wie kann ich da Strategien anwenden? Also das fände ich besonders in Englisch sinnvoll. [0:08:34] I: Also auch noch für weitere Bereiche außerhalb jetzt dieser Text. [0:08:38] TP: Ja, auf jeden Fall. (KMP05, Pos. 63-65)
				Focus on error selection	TP positively evaluates the necessity of error selection.	[0:07:40] TP: Ziemlich gut, weil ich finde, das, also in den Kursen, die ich jetzt bisher besucht habe, hat man halt auch nie gelernt, so wirf den Schülern nicht alles um die Ohren. Das habe ich eher in der Nachhilfe gemerkt, wenn die mir Texte gegeben und ich alles rot anmale, dann fällt denen halt alles aus dem Gesicht, ne. Und deshalb das habe ich praktisch selber gelernt und das wäre schön, wenn es da eben so einen Kurs gebe, wo man zum Beispiel einfach einen Kurs ähm Schüler-feedback. So wie gebe ich das und dann prüft man das halt so. Das wäre genial. (FLA04, Pos. 61)
				Focus on interpersonal aspects	TP mentions the importance for interpersonal aspects in the EFL classroom that could be assessed in the test.	0:02:02] I: Mhm. Und wie sieht das für das Fach Englisch aus, also irgendwie für das Englischlehramtsstudium? [0:02:08] TP: Fast noch besser. [0:02:09] I: Warum?

						[0:02:10] TP: Ähm, weil man im Englischen glaube ich, ähm, in einer Prüfungssituation, die jetzt gerade, ähm, simuliert wurde, kann man besser auf die Fähigkeiten der Schüler eingehen, ähm, und vor allem Schülern vielleicht auch oder lernen den Schülern die Angst zu nehmen vor dem Englischen, was glaube ich auch ein Riesenproblem sein könnte. (AEM23, Pos. 32-35)

F.3. Expert Interviews

Part 1 – Evaluative comments

Code	Explanation	Example
Text deep-level features	Expert addresses text deep-level features.	[0:14:24] E: Also eigentlich/ Ich war irritiert, dass es jetzt auf Deutsch stattgefunden hat und ich nehme an, die an-gehenden Lehrpersonen die müssen sollten das eigentlich in der Zielsprache Englisch durchführen. Also davon bin ich jetzt ausgegangen, dass sie das tun sollten und das hat sie nicht gemacht. (E1)
Dialogue	Expert addresses the engagement of the learner in a dia-logue.	[00:11:07] E: [...] Also gerade in dem Gespräch wird ja auch der dialogische Charakter vom Feedback ziemlich deutlich, den man oft dann NICHT so hat vielleicht in anderen Situ-ationen oder diejenigen, die noch einem alten Feedback-verständnis folgen. Und ich finde, sie hat das Format auch in der Hinsicht gut genutzt, also, dass sie Fragen gestellt hat an die Schülerin und sie auch aktiv eingebunden hat und, dass sie, ja, auf ihre Rückfragen dann auch eingegan-gen ist und/ Also das jetzt zu der Gesprächsführung. (E3)
Text surface-level features	Expert addresses text surface-level features.	[0:14:24] E: [...] Einfach auch mit den Schluss so: ‚Ah, das muss man nicht überarbeiten‘ mit dem Recency-Effekt, da gehe ich dann all diese konstruktiven Anmerkungen ha-ben dann eigentlich wahrscheinlich nicht viel gebracht (E1)
Revision	Expert addresses the advocacy of revi-sion.	[00:25:04] E: [...] Also da hätte er zum Beispiel auch fragen können, so was wie, ja, anstatt einfach nur so zu fragen ja das ist irgendwie schon klar, dass er da nochmal fragt was jetzt, dass sie nochmal am Text zeigen soll, was jetzt so diese Kernelemente sind von einem news report. Also das auch hier dieses über die Fragen geht, die Arbeit an dem Material, mehr, ja, Dialog gibt und vielleicht auch eine konkretere Lernunterstützung. (E3)
Specificity	Expert addresses specificity.	[00:13:26] E: She virtually now, with few exceptions, she never listened to the student at all. It was a straight mono-logue with very few exceptions and given my knowledge of feedback, it is about, not even, the quality of what is given but the nature of what is heard and understood and there was very little checking to see whether the student understood. (E5)
Learning objec-tives	Expert addresses learning objectives (Feed Up).	[0:39:13] E: [...] Dann hat er paar mal, ich glaube ein zwei-mal hin und her geschwitzt und ist dann bei Deutsch ge-blieben und hat eigentlich den ganzen Text Satz für Satz besprochen, einfach erklärt, aber war defizitorientiert. Das hat einfach alle Fehler herausgepickt, aber nicht erklärt, weshalb es falsch ist, sondern einfach: ‚Das hätte ich jetzt so und so gemacht.‘ Und hat gleichzeitig auch ein eigent-lich das den Text, also es war auch (würde er es) sich selbst erklären (E1)

Code	Explanation	Example
Structure of the feedback conversation	Expert addresses the structure of the feedback conversation.	[0:54:37] E: [...] Also sie, ich habe das Gefühl, sie ist noch viel mehr an ihrer Lernprogression orientiert in ihrem Feedback als die anderen. Ja. (E1)
Mentioning positive aspects (first)	Expert addresses mentioning positive aspects (first).	[0:25:18] E: [...] Was mir aufgefallen ist, beim Wissen über die Sprache, dass unregelmäßige Verben keine Regeln haben. Und das ist so das sehe ich nicht, stimmt so eigentlich nicht unbedingt. Es gibt schon Regelmäßigkeiten bei den unregelmäßigen Verben. (E1)
Final message (Recency)	Expert addresses the finale message that learners have of the feedback conversation.	[00:40:17] E: Genau, ja, das ist halt so eine Frage der Gewichtung. Das finde ich schon wichtig, dass er halt sprachliche Dinge auch explicit adressiert hat. (E3)
Learner support	Expert addresses support the learner.	[0:11:57] E: [...]Have you got the guts of the things and letting the student know that she had important information there with something positive and talking about that for a while and then going perhaps to an important part of the genre for reporting, TENSE and returning to that later but then going back to the hook, the need to have a title to hook the reader in and you know there, she had the right facts for the reader (E4)
Learners ability to use the information / Checking understanding	Expert addresses learners' ability to use the feedback and participants making sure of it.	[0:39:13] E: [...] Also alles sehr aus dem Kontext genommen und sehr auf die grammatikalische Ebene geblieben, ohne also sehr viel zerdenkt, sehr viel Informationen auch, aber Oberflächeninformationen, mit denen die Schülerin wahrscheinlich nichts nicht wirklich was anfangen kann. (E1)
Deficit orientation	Expert addresses deficit orientation, i.e., a detailed focus on mistakes and errors.	[0:09:16] E: [...]Also wenn man jetzt sozusagen vom Feedbackmodell Hattie und Timperley ausgehen, dann könnte man sich so orientieren, dass man erstmal sagt, was ist eigentlich das Ziel dieses Artikels gewesen. Das ist sozusagen dieser Aspekt des Feed-Ups (E2)
Learner progression	Expert addresses the reference to learner progression, either to past or future.	[0:14:24] E: [...] und sie hat auch versucht Vorwissen zu aktivieren, (E1)
Accuracy (Participant)	Expert addresses participants accuracy, either with regards to language or other professional aspects.	[0:09:16] E: [...] Ganz gut gesagt, so ich gehe jetzt nicht auf alle Grammatikfehler ein, hat sie irgendwie gesagt, aber jetzt zum Beispiel eine Sache, die man sich angucken sollte, ist die Vergangenheit. (E2)
Emotional-motivational level	Expert addresses the emotional-motivational level and relationship.	[00:56:42] E: [...] Ja, ich finde es wirklich sehr schwierig, mit den beiden. (...) Also, ich fand das erste strukturierter, soweit ich mich erinnern kann. (E3)

Code	Explanation	Example
Use of (external) material	Expert addresses the use of (external) material.	[0:55:33] E: [...] Anyway, I thought / and also her student came away with the same: do I have to revise this? Although the student was obviously keen to ask do I have to write it again? She said yes you DO have to revise this. So, a couple of the others really did not make it clear, you know, revision was needed. (E4)
Feed forward	Expert addresses the provision of feed forward.	[00:40:27] E: It was straight to the details. The poor student, I am sure, was overwhelmed with everything was a detail, it was about low level content. Of course it is important but there was no concept, it was no depth of understanding, he focused all on tense. When she asked one question: how do I learn? He just said learn it by heart. And then contradicted the whole thing in the end by saying it was not too bad. Every single thing was a negative and about detail. So, if I was the student, I would go away thinking that all I have to do is learn everything by heart, road learn, memorise, no concept of tense, no concept of irregular verbs. I would be demoralised after that. (E5)
Use of target language	Expert addresses the use of target language.	[00:25:04] E: [...] Also das merkte man auch an der Körperhaltung und das er dann auf den Text mehr eingegangen, so gezeigt hat, also diese Interaktion über das Blatt mit der Schülerin gab und das gab es auf jeden Fall im ersten Video ziemlich oft und hier kam das erst später, wodurch ich es dann auch distanzierter fand in diesem Video zu BEGINN. (E3)
Contextualization of errors and mistakes	Expert addresses the contextualization of errors and mistakes in the text context.	[00:11:07] E: [...] Da war jetzt bei mir so die Frage: Ist vielleicht an manchen Stellen zu viel hedging gewesen? Also diese abschwächende Ausdrücke, da könnte man noch was anderes machen. An anderen Stellen hat sie aber dann das doch auch deutlicher gesagt, also, dass SOLL anders, als Journalist SOLL man neutral sein und kein wertenden Satz schreiben in diesem Schlusssatz. Zwischendurch hat sie auch so paar Zusammenfassungen nochmal gemacht. (E3)
Activating prior knowledge	Expert addresses the activation of prior knowledge.	[0:30:54] E: [...] The student is going away with absolutely no idea, in my view, of what exactly / at the end I think she says, and I should look and see exactly what to improve, yeah, she should find out exactly what to improve because he was not specific, and I find his explanations incomprehensible. (E4)
Selecting errors	Expert addresses the selection of errors.	[00:12:51] E: [...] When I was the student, I would have walked out feeling a little bit demeaned by the whole exercise because other than a few platitudes, she corrected me on everything but there was no, where-to-next. And given [the] whole theory of feedback is feedback is primarily a function of where am I going. I do not think that student walked out with a going-next. She kind of offered/ The student kind of offered do I have to do it again? No, you do not. So I think the student walked out

Code	Explanation	Example
		outside like: well, that is over. And there was very little learning (E5)
Balancing hedged communication	Expert addresses balancing hedged communication with more direct feedback.	[0:30:54] E: [...] Now, and I did not like the notion, I mean, it is all about us and students and it was not a rhetorical question, like what is your view, are you happy with the text? What do you think about this? If they are happy with it, how are they going to be open for feedback? (laughs). It is sort of an interesting TECH. (E4)
Self-evaluation	Expert addresses the initiation of self-evaluation.	[00:53:01] E: [...] She started off with some comments about what the student had done right, the catchy headlines, what she had done right on the tenses and she gave examples where the student got it right(E5)
Metalinguistic language	Expert addresses the use of metalinguistic language.	[0:30:54] E: I think I would fail that student. Because he does not have a meta-language to talk about text, and so, he just simply talks about THINGS or other students (laughs). (E4)

Part 2 – Evaluation of the scoring instrument

Code	Sub-code	Explanation	Example
General comments			
	Overall evaluation	The scoring instrument is evaluated overall.	[01:31:52] E: No, I like where you are going. I like what you are doing. I think you are going to have good set of (unv.). It is going to be interesting to see/ (E5)
	Comprehensiveness	The general scope of the scoring instrument is addressed, positively or negatively	[1:17:29] E: Also ich finde es erstens, also es aus persönlicher und auch aus fachlicher Sicht spannend, wie ihr das operationalisiert habt, dass ihr miteinbezogen habt. Und ich finde jetzt so als ersten Gesamteindruck sehr sehr umfassend, also sehr/ Ihr habt wirklich aus meiner Sicht sehr viele Aspekte oder ist es euch gelungen, aus meiner Sicht, sehr viele Aspekte abzubilden, die relevant sind in diesem Kontext. Auch mit Bezug auf die relevanten Theorien. Ich denke jetzt gerade an zum Beispiel an Feed-Up, Feedback, Feed-Forward. Das ist alles auch drin ist, also die/ Ja, es ist sehr umfassend und ich glaube, wenn man das wirklich alles so gut anwenden kann, würde man ein sehr, wie sagt man, detailliertes Bild von der Leistung und von der Performanz von den Testpersonen erhalten können. Sofern sie das halt alles auch zeigen. (laughs) Aber ich kann mir vorstellen, dass vielleicht viele von diesen Aspekten nur teilweise beurteilt werden können, wenn es halt nichts gezeigt wird in den Performanzen, aber so allgemein sehr umfassend und sehr, wie sagt man, so einfach sehr/ Das Wort kommt mir wahrscheinlich

Code	Sub-code	Explanation	Example
			nachher noch in den Sinn, also ich habe das Gefühl, es deckt alles ab, was aus meiner Sicht relevant ist. Ja. (E1)
	Based on previous research	The use of previous research for the scoring instrument is addressed.	[0:53:14] Aber ich glaube nicht, dass sich etwas doppelt, also mein Gesamteindruck ist gut, die Kategorien ergeben sich ja aus der Literatur, finde ich sind schön herausgearbeitet, wie gesagt, ich glaube in den meisten Fällen würde es schon helfen, wenn wir sozusagen hinterher gucken könnten, wie gut ist das Kriterium erfüllt. (E2)
	Positive highlights	Specific aspects are highlighted	[01:27:45] E: [...] I like/ We did not see it very often at all, but I like the notion of trying to understand how the student was receiving and understanding the feedback.
	Missing aspects	Aspects that are potentially missing are addressed	[01:40:17] E: [...] In den Kriterienraster weiß ich jetzt nicht, ob es schon irgendwo untergebracht war, das man diese 3D Perspektive hat, so zu sagen. Also das man die Materialien, die Arbeitsblätter einbindet, das man auf konkrete Textstellen verweist, weil das findet ja immer in einem bestimmten Kontext statt, in dem Materialien sind, sei es jetzt hier in dem Vorortkontext oder online, und diese Materialien, die haben ja auch eine Rolle in dem Feedbackprozess. (E3)
	Redundant aspects	Redundancy of aspects is addressed.	[1:33:39] E: Gut, das sind ja teilweise vielleicht schon, aber es sind dann Aspekte, die aus unterschiedlichen Perspektiven angegangen werden. Von dem her wahrscheinlich, ah eher nicht. Also nicht jetzt, dass mir jetzt wirklich aufgefallen ist, das was doppelt vorkommt (E1)
	Scoring	The actual scoring is addressed. This includes the distribution of points as well as issues of reliability, pragmatics and grading.	[2:12:02] E: I think you are being quite exhaustive in terms of coverage. My concern will be pragmatics. (E4)
	Context of the RobS	The context of the RobS is addressed.	[01:26:44] E: Ja, also an sich finde ich bisher alles nachvollziehbar. Ich hatte ja als ich diesen Text bekommen hatte, auch schon einige Sachen, die ich mir direkt markiert und da waren auch solch Unterscheidungen mit bei, von daher kann ich das bisher so nachvollziehen. Im Hinblick auf dieses Design, so einer Prüfungsaufgabe, fände ich es halt sehr sinnvoll, wenn jetzt nicht nur die Studierenden, diese Simulation machen, sondern es danach noch irgendwie ein Gespräch geben würde, wo sie dann sagen, warum sie etwas nicht

Code	Sub-code	Explanation	Example
			gemacht haben oder gemacht haben. Also, weil es kann für jedes dieser Sachen, die jetzt gerade auf der Folie sind, immer gewisse Gründe geben, weil sie vielleicht angeben, ok dafür brauch man jetzt keine Erläuterung nochmal, zu diesem Fehler oder das sie dann nochmal halt sagen, ok das ist ein wiederkehrender darum hab ich da nochmal Kommentare oder zusätzliche Ressourcen genannt. Also das fiel mir jetzt so darüber hinaus noch ein. (E3)
	General comments on "General aspects of feedback"	The set of items for "General aspects of feedback" is addressed generally.	[01:04:26] E: No, I think that is pretty much going on all different levels. [01:05:21] E: No, I am sorry. It is good. Obviously I like the criterion. (E5)
Self-evaluation		The criterion "self-evaluation" is addressed.	[1:23:12] E: Ja, also ich finde ganz zentral auch auf jeden Fall diese den Einbezug von den Lernenden durch die Self-Evaluation, also Selbstevaluation gleich am Anfang. Also wirklich, dass es ein die Lernenden auch kognitiv auch aktiviert werden und ihnen nicht einfach nur Feedback (laughs) ja gegeben wird ohne, dass die das nur aufnehmen und nicht selber mitdenken.(E1)
Positive feedback		The criterion "positive feedback" is addressed.	[1:23:12] E1: [...] Das positive Feedback finde ich gut und spannend und ich finde es vor allem spannend, weil da kommt mir die Frage: Wenn so ein explizites Schicht-Sandwich zu Tage gebracht wird, positiv, negativ, positiv, irgendwie wenn das eine Rolle spielt, also was die damit macht, vor allem wegen auch so Recency-Effekt und so, wenn am Schluss ja, genau. Also es gibt ja unterschiedliche Auffassungen so davon, wie wirksam diese Sandwiche sind. (laughs)(E1)
Learning objectives (Feed Up)		The criterion "learning objectives (Feed Up)" is addressed.	[1:17:29] E: Also ich finde es erstens, also es aus persönlicher und auch aus fachlicher Sicht spannend, wie ihr das operationalisiert habt, dass ihr miteinbezogen habt. Und ich finde jetzt so als ersten Gesamteindruck sehr sehr umfassend, also sehr/ Ihr habt wirklich aus meiner Sicht sehr viele Aspekte oder ist es euch gelungen, aus meiner Sicht, sehr viele Aspekte abzubilden, die relevant sind in diesem Kontext. Auch mit Bezug auf die relevanten Theorien. Ich denke jetzt gerade an zum Beispiel an Feed-Up, Feedback, Feed-Forward. Das ist alles auch drin ist, also die/ Ja, es ist sehr umfassend und ich glaube, wenn man das wirklich alles so gut anwenden kann, würde man ein sehr, wie sagt man, detailliertes Bild von der Leistung

Code	Sub-code	Explanation	Example
			und von der Performanz von den Testpersonen erhalten können (E1)
Feedback (Feed Back)		The criterion "feed-back (Feed Back)" is addressed.	[1:17:29] E: Also ich finde es erstens, also es aus persönlicher und auch aus fachlicher Sicht spannend, wie ihr das operationalisiert habt, dass ihr miteinbezogen habt. Und ich finde jetzt so als ersten Gesamteindruck sehr sehr umfassend, also sehr/ Ihr habt wirklich aus meiner Sicht sehr viele Aspekte oder ist es euch gelungen, aus meiner Sicht, sehr viele Aspekte abzubilden, die relevant sind in diesem Kontext. Auch mit Bezug auf die relevanten Theorien. Ich denke jetzt gerade an zum Beispiel an Feed-Up, Feedback, Feed-Forward. Das ist alles auch drin ist, also die/ Ja, es ist sehr umfassend und ich glaube, wenn man das wirklich alles so gut anwenden kann, würde man ein sehr, wie sagt man, detailliertes Bild von der Leistung und von der Performanz von den Testpersonen erhalten können (E1)
Next Steps (Feed Forward)		The criterion "next steps (Feed Forward)" is addressed.	[1:17:29] E: Also ich finde es erstens, also es aus persönlicher und auch aus fachlicher Sicht spannend, wie ihr das operationalisiert habt, dass ihr miteinbezogen habt. Und ich finde jetzt so als ersten Gesamteindruck sehr sehr umfassend, also sehr/ Ihr habt wirklich aus meiner Sicht sehr viele Aspekte oder ist es euch gelungen, aus meiner Sicht, sehr viele Aspekte abzubilden, die relevant sind in diesem Kontext. Auch mit Bezug auf die relevanten Theorien. Ich denke jetzt gerade an zum Beispiel an Feed-Up, Feedback, Feed-Forward. Das ist alles auch drin ist, also die/ Ja, es ist sehr umfassend und ich glaube, wenn man das wirklich alles so gut anwenden kann, würde man ein sehr, wie sagt man, detailliertes Bild von der Leistung und von der Performanz von den Testpersonen erhalten können (E1)
Task		The criterion "task" is addressed.	n.a.
Process		The criterion "process" is addressed.	n.a.
Self-regulation		The criterion "self-regulation" is addressed.	[0:53:14] Auch bei der Selbstregulation, wie hilfreich ist es jetzt im oxford dictionary nachzugucken, wenn wir im Zeitalter von Digitalisierung sind, ja machen wir einen tick Strategie, aber ist diese wirklich noch so der Brenner. Oder der Hinweis, guck mal wir haben hier am Ende des Buches nochmal die Liste der irregular verbs, das finde ich ist eigentlich eine ganz gute Sache, geh die

Code	Sub-code	Explanation	Example
			nochmal durch und vielleicht gibt es die dann auch auf irgendeiner App was auch immer, wo man die dann nochmal lernen könnte (E2)
Self		The criterion "self" is addressed.	[1:19:47] E: Yeah, where he asked the student about structure, she says she think so, she tried and what. It is / I think that was problematic and many things are more, not just self-regulation, but more affect, that is meta-cognitive, where it is as you said, you have got nothing, and this is one the self about affect, a criterion related to bolstering self-efficacy, you know a setup as a writer. I am not sure what self does not address the learner's personal level would actually relate to, making him feel at EASE Making the / the fourth one which was one that kept saying it takes courage to / takes courage to hand this in. Well, I cannot remember. [1:20:52] I: This is related to Hattie's four dimensions of feedback. And he says that like a simple praise, like you are a good student, this would fall under that category, so if someone said that in the simulation, he would not get that point. However, this did not occur in the videos and this is something that I believe / [1:21:13] E: Oh, this is something negative. You get a point if you do not say it (laughs). Okay, sorry, I misinterpreted that entirely. So it does not address the learner's / there is no generic feedback about, on a / [1:21:30] I: Yes, no generic praise. [1:21:36] E: Although, some say, in the videos that it was a good text. [1:21:40] I: Yeah, this is a grey area right. It is somehow related to the task because the task was the text but it is still generic, so. [1:21:51] E: It has to not related to the text. A good text is alright but all good effort which someone / effort like in video four, he made a good effort, that is a personal thing is it not? [1:22:05] I: That is true. [1:22:09] E: Or no, I do not know (laughs). To me that is relatively problematic (...). So, what is missing is, yes maybe, something that has to do with the effect of aspects of both self-evaluating, of getting feedback, of the self of the writer or even, I mean, some students may find this really engaging and others sort of was not, I do not know. (E4)

Code	Sub-code	Explanation	Example
	Dialogue Orientation	The criterion "dialogue orientation" is addressed.	[01:02:08] [...] I like the dialogue orientation because it should be/ it is not just a dialogue it is does both parties hear and understand the other party. And so, after the ? conversation where you reinforce you know, no that is not what I meant, this is what I meant, et cetera. So I can see how that comes out. So, I ?? with the dialogue as much as involving, that checking the learners understanding. (E5)
	Quality of Feed Back	The criterion "Quality of Feed Back" is addressed.	[01:09:44] E: Ja, das sind glaube ich so Standartsachen, sowie mit diesem möglichst genau und klar, beschreibend und hier auch mit diesem dialogischem Charakter. Ja. So spontan (...) ja noch nichts weiter. (E3)
	Quality of Feed Forward	The criterion "Quality of Feed Forward" is addressed.	[01:03:00 (...) Now quality of feed-forward, now I like that, it is the one that matters the most to me and it is also coming back to the dialogue. Do the teacher check that the student understood what the call of action was or just go back and do it again. Well, this is what I want you to find out, I would like a little bit more of that in that criteria. (E5)
	Securing understanding / content adaptation	The criterion "securing understanding / content adaptation" is addressed.	[01:27:45] E: Overall impression, like until I saw the elaboration I was getting worried that you were privileging content. And I probably would have put more emphasis on the elaboration, across the whole marks. But overall I think you got it, in terms of how are they going where are they going. I like/ We did not see it very often at all, but I like the notion of trying to understand how the student was receiving and understanding the feedback. Because as I do more and more work in this area, it becomes more and more critical. We can get teachers, and you know this, who can give brilliant feedback that is not understood. And my mentor is, was not heard or understood (unv.). When I look at one of your slides, it comes across that you at least (unv.) about was it heard, none of that is really checked whether the student understood it. And the only example of action, it was: do it again. But in the scoring, there are options for that. (E5)
	Linguistic adaptivity	The criterion "linguistic adaptivity" is addressed.	[1:03:34] E: Ja, finde ich auch interessant. Ich meine hier kann man ein bisschen, hier kommt man in so einen Graubereich, der ja nicht gut erforscht ist. Das ist sozusagen ein didaktisches Desiderat möglichst viel auf Englisch zu machen, ob das jetzt sozusagen im Feedback den Lernenden

Code	Sub-code	Explanation	Example
			hilft, ist ja so ein bisschen unklar und wenn wir jetzt von code switching ausgehen würden, dann würde das wahrscheinlich hilfreich sein, wenn man switched, aber so richtig hilfreich war das noch nie, deswegen finde ich, also das ist sozusagen auf einer anderen Ebene, das ist eher auf einer didaktisch normativen Ebene, aber einer empirisch gesicherten Ebene. Ich glaube das ist sozusagen ein Bereich, den man für sich selber so einmal klar-kriegen muss, denn hier geht man auf eine andere Ebene, das heißt nicht unbedingt, das also vielleicht wären die Lernenden auch fähig das auf Englisch zu verstehen, also das weiß man halt einfach nicht so. Deswegen habe ich jetzt hier keine direkte Anregung, ich finde es nachvollziehbar, gerade aus fachdidaktischer Richtung, würde jetzt aber keine Aussage darüber machen, was wirklich so wichtig ist jetzt für den Lernenden, außer das Zeitargument ist immer eins, also je mehr Kontakt die mit dem Englischen haben ist natürlich desto besser. (E2)
Adaptive error correction		The criterion "adaptive error correction" is addressed.	[1:28:01] E: Ich finde das sehr (es bringt seine (unv.) zusammen?) und gute Herangehensweise. Vor allem die Probleme, die wir hatten mit diesem Raster von Lukas Bleichenbacher, war halt schon, dass es war sehr vage, also es ist. Was bedeutet dann Adaptivität? Ist das eine sprachliche Adaptivität? Ist das eine adaptiere ich die Inhalte auf das Vorwissen von den Schülerinnen und Schülern? Und das finde ich hier schön ausdifferenziert auf das Wissen, auf die Sprache und auch auf die Error Correction.(E1)
Prioritization of content/structure		The criterion "prioritization of content/structure" is addressed.	[01:31:06] E: Ja vielleicht. Ja genau und mit diesem Fokus auf den Inhalt und, also Struktur, das war ja auch das, was ich eingangs fragte, wie stark soll das denn jetzt gewertet werden? Weil das ja auch in der Aufgabenstellung auch so rüberkam, also hätte es eigentlich den Kandidatinnen da auch klar sein sollen. Ich glaube aufgrund dieser Menge an Fehlern die da waren, haben sie sich oft daran aufgehalten an den sprachlichen Sachen. Und auch mit der Menge, genau, das war da auch noch so ein Punkt wo ich da dachte: ok, dieses Gespräch kann nicht alles abdecken, sollte es ja auch nicht, damit dann die Schülerin nicht überfordert wird und wie sie da den Fokus setzen.

Code	Sub-code	Explanation	Example
Appropriate amount of information		The criterion "appropriate amount of information" is addressed.	[1:07:36] E: Deswegen, also ich meine wir haben ja eh schon festgestellt, dass es nicht so leicht ist, dass man da drei gibt und dann da zwei, das ist halt immer ein bisschen schwierig das man wirklich Punkte gibt, aber man könnte es sicherlich abticken und da würde man glaub ich eine super interater reliability hinbekommen bei allen Punkten bis auf appropriate amount (unv.), vielleicht könnte man da noch eine (unv.) irgendwie geben, also wie viel soll es denn sein in einem zehn minütigen Gespräch müssten zumindest irgendwie drei Punkte genannt werden, aber es dürfen auch nicht zwanzig genannt werden, also das müsste man irgendwie doch, weil ich könnte mir sonst vorstellen, dass man nicht mit diesem Bogen auf eine gute interater reliability kommt. (E2)
Language Correctness		The criterion "language correctness" is addressed.	[2:01:34] E: The one and two seem to be almost something, you could subsume them, language accuracy and language appropriacy, inappropriate use of tense, forming of the simple past, well / at some point or some sort of grammatical, anyway. (E4)
Language Adequacy		The criterion "language appropriateness" is addressed.	[2:01:34] E: The one and two seem to be almost something, you could subsume them, language accuracy and language appropriacy, inappropriate use of tense, forming of the simple past, well / at some point or some sort of grammatical, anyway. (E4)
Content		The criterion "content" is addressed.	[1:08:45] E: Ne, also ich meine der Punkt mit den Punkten da haben wir ja schon drüber gesprochen, dass das schwierig ist und ansonsten die Idee zum content structure und zum Prozess was zu sagen, finde ich auch eine gute Idee auf jeden Fall.(E2)
Structure		The criterion "structure" is addressed.	[1:08:45] E: Ne, also ich meine der Punkt mit den Punkten da haben wir ja schon drüber gesprochen, dass das schwierig ist und ansonsten die Idee zum content structure und zum Prozess was zu sagen, finde ich auch eine gute Idee auf jeden Fall. (E2)
Process		The criterion "process" is addressed.	[1:08:45] E: Ne, also ich meine der Punkt mit den Punkten da haben wir ja schon drüber gesprochen, dass das schwierig ist und ansonsten die Idee zum content structure und zum Prozess was zu sagen, finde ich auch eine gute Idee auf jeden Fall. (E2)

Code	Sub-code	Explanation	Example
Process orientation		The criterion "process orientation" is addressed.	[1:15:32] E: [...]und bei der process orientation das finde ich gut, also das klar wird, dass man es auf jeden Fall überarbeiten muss. Jetzt müsste man sich natürlich fragen, ob man auch fragen soll, wie die Lernende den Text geplant hat, also wenn man jetzt prozessorientiert vorgehen würde, könnte man auch nochmal sagen in diesen Bereich Planung zu gehen, aber ich finde es auch legitim, sich nur auf die revision zu stürzen, weil wir ja sehen, dass eigentlich das Planen nicht das Problem ist, das machen ja auch Lehrkräfte häufig im Unterricht, aber sie fordern zu wenig Überarbeitung ein und insofern finde ich es auch durchaus ein valides Vorgehen hier zu sagen, ich beschränke mich jetzt hier auf revision, finde ich völlig okay. Es war ja auch die letzte glaube ich, die gesagt hat nein das wird jetzt überarbeitet, das war ja durchaus positiv zu sehen. (E2)
Error pattern		The criterion "error pattern" is addressed.	[2:04:45] E: Why would they get three points if they do that exclusively? Which is really poor-quality feedback if you do that exclusively (laughs). [2:05:07] E: No, no. You could say PREDOMINANTLY (...) addresses error patterns, which I do not think is good. I mean, I would not give them three points for that (laughs). This is the problem with your frequency scale. You are implying that it is good to do this a lot. (E4)
Target-oriented error selection		The criterion "target-oriented error selection" is addressed.	[1:11:16] E: Und wenn wir sagen TP always in exclusively addresses error patterns in the feedback conversation, also dann ist nicht ganz klar, dass ist ja eigentlich nicht besonders gut, wenn er das nur tun würde, weil dann geht es ja eigentlich um surface-level areas, weil hier jetzt z.b die conclusion fehlt ist ja kein error pattern, sondern das ist ja eher ein strukturelles Problem. Brauchen Sie eigentlich diese Kategorie, die zweite, also eigentlich ist für mich die wichtigere Kategorie ist Ihre dritte Kategorie, nämlich die Target oriented error selection TP always selects error and/or patterns and relates them to the learning objectives, das ist super, weil darum geht es ja, dass man jetzt irgendwie drei auswählt und das wären jetzt vier gewesen, die tenses meinetwegen noch die false friends, wo man sagt, okay da gehe ich jetzt ran, aber, also wir wissen ja, dass diese Überdosierung an error correction eher negativ ist und ich glaube damit kommen Sie ja dann, gerade wenn wir es jetzt hoch skalieren dann hätten Sie ja im Grunde, würden Sie ja sozusagen was positiv bewerten was

Code	Sub-code	Explanation	Example
			ja ihr schwächster Kandidat ja gemacht hat, der hat ja eigentlich nur, der ist ja nur die Fehler durchgegangen. Also der hat nicht gesagt, dass es pattern ist, aber selbst wenn er jetzt ganz viele pattern genannt hätte, wäre es ja auch nicht besser gewesen, weil was soll man damit anfangen, wenn man jetzt zwhn pattern gefunden hätte. (E2)
Detailedness		The criterion "detailedness" is addressed.	[01:19:32] E: Well, I think your elaboration is very critical because it is probably getting to what I talked before about that balance and maybe that's the place to put it, alongside that. (E5)
Linguistic accuracy		The criterion "linguistic accuracy" is addressed.	[1:32:05] E: Fachlichkeit genau. Ja. (...) Gut und das finde ich auch das finde ich toll auch hier diese Ausdifferenzierung zwischen diesen beiden Kategorien, was auf jeden Fall Sinn macht und/ Das finde ich auch spannend, weil ich habe manchmal so das Gefühl, es ist fast wie ein Spektrum oder ein Kontinuum von Linguistic accuracy und pedagogical accuracy. Wo bewegt man sich dazwischen? Schwierigkeit ist wirklich das so klar auseinander zu halten, weil man das ja irgendwie/ Man muss ja irgendwie. Und ich finde das ist/ (E1)
Pedagogical accuracy		The criterion "pedagogical accuracy" is addressed.	[2:30:34] E: So, you have a lot of focus on learning strategies, yet I do not actually / in the four videos I saw very little comment on. (E4)

G. Coding Samples

G.1. Stimulated Recall

1

Transkript Stimulated Recall Interview (SIM)

Fachzuordnung

☐ Bildungswissenschaften

☒ Englisch

Pseudonymisierungscode

MJZ08

Länge

00:21:15 Minuten

Aufnahmeort

Universität Paderborn

Transkriptionskonvention

Inhaltlich-semantische Transkription

Interviewkürzel

TP: Testperson

I: Interviewer

TPV: Testperson Video

SV: Schülerin Video

MV: Mutter Video

„Test situation“

1 [0:00:00] I: So. Der Code ist M J Z 0 8. Meine erste Frage lautet
 2 aber erstmal gar nicht auf dem Video, sondern das ist erstmal so
 3 eine generelle Einschätzfrage. Wie war die ganze Situation für Sie
 4 Erzählen Sie einfach mal.
 5
 6 [0:00:17] TP: Hm. Ja ich musste mich erst in die Materialien
 7 einarbeiten, um das überhaupt gut verstehen zu können, weil nur da
 8 kann ich ja auch so richtiges oder angemessenes Feedback geben. So
 9 dass halt sich alles zu merken das war so ein bisschen die
 10 Herausforderung. Dann natürlich auch die spezielle Situation. Das
 11 mit der Kamera und so. Das war dann auch ein bisschen, ja, was Neu
 12 dann.
 13
 14 [0:00:40] I: Hm. Okay. Ich mache einmal nur die Fenster zu, weil d
 15 ist sonst zu laut. Hatten letztens ist hier auch einmal ein
 16 Bollerwagen mit Schlagermusik vorbeifahren. Das war auch /
 17
 18 [0:01:01] TP: Oh ((lacht)). Hat man dann gehört in der Aufnahme?
 19
 20 [0:01:06] I: Ja das war als wir da drüben waren /
 21
 22 [0:01:09] TP: Ah, okay.
 23
 24 [0:01:10] I: Und das aufgenommen haben. Das war zum Glück noch nie
 25 so wichtig, aber das war / deswegen muss man sehr vorsichtig sein
 26 (unverständlich, ca. 2 Sekunden). Okay, gut. Dann würde ich sagen,
 27 fangen wir auch schon /
 28
 29 [0:01:19] TP: Ja.
 30
 31 [0:01:20] I: an mit dem Video.
 32
 33 [0:01:28] [0:00:00] Vorbereitung [0:00:27] S. ((klopft)) [0:00:31]
 34 SV: [good morning] [0:00:32] TPV: [ah hallo] (-) morgen (.) komm
 35 rein mia (.) setz dich hin
 36
 37 [0:01:40] I: Was haben Sie in dem Moment gedacht als die Schülerin
 38 reinkam mit den Worten: „Good morning“?

..Test situation

39
 40 [0:01:48] TP: Eigentlich nicht so viel, weil ich es gar nicht
 41 erwartet habe, dass sie überhaupt reinkommt. Ich habe halt gedacht
 42 das findet irgendwie zwischen einer Tür oder auf dem Flur irgendwo
 43 statt. Aber jetzt im Nachhinein ist es ja schön, dass sie gleich
 44 Englisch gesprochen hat.
 45
 46 [0:02:01] I: Mhm.
 47
 48 [0:02:01] TP: Das ist ja gut.
 49
 50 [0:02:03] I: Okay.
 51
 52 [0:02:05] [0:00:37] TPV: ja erstmal finde ich das super von dir das
 53 du die chance nutzt für ein feedback gespräch zu holen [0:00:41] SV:
 54 dankesch[ön] [0:00:42] TPV: [ja] (.) das zeigt mir halt auch das du
 55 wirklich deine noten (.) daran interessiert bist deine noten wieder
 56 zu verbessern [0:00:47] SV: [hm_hm] [0:00:47] TPV: [das] ist super
 57 das merke ich mir (.) also [0:00:50] SV: ja [0:00:51] TPV: das freut
 58 mich (.) äm (.) genau du hast mir jetzt diesen bericht abgegeben
 59 [0:00:54] SV: hm_hm [0:00:55] TPV: ich habe den jetzt auf die
 60 schnell jetzt mal angeschaut ja (.) als lehrer korrigiert man ja
 61 immer zwei sachen man schaut sich erstmal die sprache an (.) hast du
 62 da grammatikfehler oder äm (.) welche vokabeln hast du benutzt ja
 63 das wäre die eine seite (.) und dann wir hatten jetzt ganz viel über
 64 die zeitungsberichte gesprochen (.) schaut man sich auch immer an
 65 (.) hast du dich an die merkmale gehal[ten] [0:01:14] SV: [hm_hm]
 66 [0:01:15] TPV: [die] wir besprochen haben ja (.) äm (-) genau das
 67 ist jetzt so ein großes wirrwarr ich sag dir erstmal was du alles
 68 super gemacht hast (.) äm (.) du hast schon mal viel richtiges
 69 vokabular verwendet nE [0:01:24] SV: [hm_hm] [0:01:25] TPV: [passt]
 70 ja auch zu dem thema umwelt und (.) äm (.) genau für dieses event
 71 gepasst hat (.) [das] [0:01:31] SV: [ja] [0:01:31] TPV: hat mich
 72 sehr gut gefreut ja (-) manchmal musst du noch ein bisschen auf die
 73 rechtschreibung achten [0:01:36] SV: [okay] [0:01:36] TPV: [ja] da
 74 [würde ich] [0:01:37] SV: [wo denN] [0:01:38] TPV: zum beispiel bei
 75 environment (-) [ja] [0:01:41] SV: [wie] wird das denn geschriebeN
 76

77 [0:03:12] I: Was haben Sie in diesem Moment gedacht?
78
79 [0:03:14] TP: „Boah, weiß ich gerade selber wie es geschrieben
80 wird?“ ((lacht)) Nein, aber ähm, ist ja auch gut, dass sie
81 nachfragt, ne? Dass sie da die richtige Lösung dann direkt wissen
82 möchte. Das zeigt ja auch, dass das Interesse vorhanden ist bei der
83 Schülerin.
84
85 [0:03:28] I: Mhm.
86
87 [0:03:30] [0:01:42] TPV: da musst du noch ein o und ein en zum
88 beispiel einfügen [0:01:46] SV: okay (.) können sie mir das
89 vielleicht einmal aufschreiben [0:01:48] TPV: ja (--) (...)
90
91 [0:03:35] I: Was haben Sie da gedacht?
92
93 [0:03:37] TP: Ja das war das mit dem „Wie schreib ich es jetzt
94 richtig?“ ((lacht)) Nein, aber auch, genau, das ist ja, wenn die
95 Schülerin das sehen kann, können die sich da meistens noch besser
96 merken, als wenn die das noch hören. Also das war auch sehr gut von
97 der Schülerin. Also, dass sie das wissen wollte.
98
99 [0:03:49] I: Mhm.
100
101 [0:03:52] [0:01:48] TPV: (...) äm das kriegst du ja wieder das blatt
102 nE [0:01:52] SV: [hm_hm] [0:01:52] TPV: [dann] schreibe ich dir das
103 hier drüber (6.0) [0:01:58] SV: ach so (.) ja [0:01:59] TPV: ja
104 genau (.) auch immer klein (.) im englischen werden wirklich nur die
105 wenigsten sachen großgeschrieben [0:02:03] SV: [ach so] [0:02:04]
106 TPV: [im] zweifel schreib es lieber klein [jA] [0:02:06] SV: [hm_hm]
107 okay (.) [0:02:07] TPV: wenn du dir unsicher [bist] [0:02:08] SV:
108 [also] muss es hier dann auch klein [0:02:10] TPV: genau (...)
109
110 [0:04:12] I: Was haben Sie da gedacht?
111
112 [0:04:15] TP: Als sie auf den zweiten Fehler so /
113

...Learner's motivation

114 [0:04:17] I: Als sie fragte: „Muss das hier denn auch klein
115 geschrieben werden?“
116
117 [0:04:20] TP: Ja. Ja das ist gut, weil die Schülerin denkt mit.
118 Also, ja. Ist gut.
119
120 [0:04:25] I: Mhm.
121
122 [0:04:27] {0:02:10} TPV: (...) da habe ich das drunter geschrieben
123 [0:02:12] SV: hm_hm (.) okay [0:02:12] TPV: und prize (~) das ist
124 auch mit zet aber das sind jetzt so kleinig[keiten] (.)
125 rechtschreibung nE [0:02:17] SV: [hm_hm] [0:02:18] TPV: in zukunft
126 auch die vokabeln gut lernen [0:02:21] SV: [okay] [ja] [0:02:21]
127 TPV: nE [das du] auch die rechtschreibung dann gleich mit anschaut
128 (.) dann passieren weniger solche fehler [0:02:26] SV: hm_hm
129 [0:02:27] TPV: genau (.) dann sind da auch ein paar grammatikfehler
130 drin äm (.) da würde ich dir raten ganz wichtig (.) äm (.) die
131 simple past formen nochmal zu wiederholen [0:02:34] SV: simple past
132 [okay] [0:02:36] TPV: [genau] die vergangenheit (.) die zweite und
133 die [dritte form] [0:02:37] SV: [hm_hm] (.) [0:02:39] TPV: [äh]
134 [0:02:39] SV: [wo] habe ich das denn da falsch gemacht in dem text
135 [0:02:41] TPV: bei manchen hast du es(...)
136
137 [0:04:58] I: Was haben Sie da gedacht als diese Frage kam?
138
139 [0:05:02] TP: Was sie falsch gemacht hat in dem Text. Weil ich sie
140 jetzt nicht gleich komplett demotivieren möchte oder so. Weil da
141 waren schon viele Fehler drin, aber das wollte ich nicht so deutlich
142 sagen.
143
144 [0:05:11] I: Mhm.
145
146 [0:05:13] {0:02:41} TPV: (...) super richtig gemacht (.) zum beispiel
147 bei get hast richtig zu [got] umgewechselt [0:02:45] SV: [hm_hm] ja
148 [0:02:47] TPV: und manchmal ist die falsche form draus geworden (.)
149 von make ist ja nicht maked, sondern [fällt dir] noch ein was das
150 sein könntE [0:02:52] SV: [(unverständlich)] ne fällt mir jetzt so
151 nicht ein (...)

..Positive
..Options for support



152
153 [0:05:27] I: Was haben Sie da gedacht? In dem Moment?
154
155 [0:05:32] TP: Ja ich wollte ihr nochmal die Chance geben. Vielleicht
156 ist es ihr ja hinterher nochmal eingefallen. Ne? Ähm. Weiß ich
157 nicht, dann sag ich es ihr lieber nochmal, dass sie das ab jetzt
158 dann vielleicht weiß. Ja.
159
160 [0:05:44] I: Mhm.
161
162 [0:05:46] [0:02:56] TPV: das ist made (.) [0:02:57] SV: made (.)
163 hm_hm [0:02:58] TPV: make made made (.) [0:03:00] SV: ja (.) wie
164 kann ich das denn übeN [0:03:02] TPV: genau (...)
165
166 [0:05:52] I: Was haben Sie in diesem Moment gedacht bei dieser
167 Frage, wie sie das üben kann?
168
169 [0:05:56] TP: Ja das ja auch wieder, das zeigt Interesse daran, dass
170 die Schülerin sich verbessern möchte. Und gleichzeitig überlegt man
171 halt so man im Kopf: „Was gibt’s denn für Strategien, ne? Wie kann
172 sie es denn machen?“ Das habe ich mir gedacht.
173
174 [0:06:07] I: Hm. Und können Sie sich vorstellen, was das für ein /
175 worauf diese Frage gezielt hat? Auf einer didaktisch-fachlichen
176 Ebene bei Ihnen?
177
178 [0:06:18] TP: Auf mein Wissen. Also, damit ich weiß, wie kann man
179 Vokabeln, äh, vermitteln.
180
181 [0:06:23] I: Mhm.
182
183 [0:06:24] TP: Also: „Was gibt’s für Strategien?“
184
185 [0:06:28] [0:03:05] TPV: hattest du da schon immer schwierigkeiten
186 ist das schon seit der sechs [so ein thema für dich] [0:03:08] SV:
187 [ja ich hab die] nie so [viel] gelernt [0:03:10] TPV: [hm] (-) ja
188 (...)
189

190 [0:06:37] I: Was haben Sie da gedacht?
 191
 192 [0:06:42] TP: Dass sie ehrlich ist. Das ist ja gut, dass sie / vor
 193 allem nur so kann man die Sachen ja dann auch verbessern oder
 194 angehen. Und das ist halt auch typisches Schülerverhalten. Also ist
 195 nichts Ungewöhnliches.
 196
 197 [0:06:52] I: Was genau ist typisches Schülerverhalten?
 198
 199 [0:06:55] TP: Hm. Ja, es gibt halt manche Schüler, die halt sehr
 200 gewissenhaft die Vokabeln lernen und viele aber halt auch nicht.
 201
 202 [0:07:00] I: Mhm.
 203
 204 [0:07:01] TP: (unverständlich, ca. 1 Sekunde)
 205
 206 [0:07:02] I: Okay.
 207
 208 [0:07:04] [0:03:10] TPV: (...) das ist ein ganz wichtiges thema am
 209 besten (.) äm ich kann dir nachher noch ein computerprogramm
 210 empfehlen (.) das ist so ein digitaler karteikarten [0:03:19] SV:
 211 hm_hm [0:03:20] TPV: und der ist mit einem algorithmus und damit
 212 lernst du ganz schnell dann die neuen vokbalen (.) [0:03:25] SV:
 213 [okay] [0:03:25] TPV: [äh] die neuen unregelmäßigen verben [0:03:28]
 214 SV: [ja] [0:03:28] [aber] das ist auch ein guter hinweis für mich
 215 weil ich glaube das du nicht die einzige bist die da noch
 216 schwierigkeiten [mit] hat in der klasse [0:03:32] SV: [hm_hm]
 217 [0:03:33] TPV: vielleicht können wir da mal eine extra stunde machen
 218 (...)
 219
 220 [0:07:26] I: Was waren Ihre Gedanken in dem Teil als Sie sich
 221 entschieden haben, dass so zu formulieren?
 222
 223 [0:07:34] TP: Ja, ich lerne als Lehrkraft ja auch von den Schülern.
 224 Und das war einfach für mich ein Hinweis, vor allem wenn man eine
 225 neue Lehrkraft ist, ich bin ja erst vier Monate dabei.
 226
 227 [0:07:42] I: Mhm.

228
 229 [0:07:43] TP: Und das ist halt auch so ein typischer Fehler.
 230
 231 [0:07:45] I: Mhm.
 232
 233 [0:07:45] TP: Ja.
 234
 235 [0:07:46] I: Und warum haben Sie sich dazu entschieden in diesem
 236 Moment noch zu sagen, dass das auch im ganzen Klassenverband
 237 aufgetaucht ist? Weil das ist ja aus dem Material jetzt nicht so
 238 ersichtlich geworden.
 239
 240 [0:07:55] TP: Ja, das stimmt. ((atmet aus)) Vielleicht um
 241 (unverständlich, ca. 3 Sekunden) Weiß ich nicht. Ich wollte sie halt
 242 nicht so alleine dastehen lassen, dass sie da die Einzige ist, die
 243 den Fehler gemacht hat. Um das ein bisschen zu relativieren.
 244
 245 [0:08:14] I: Mhm. Okay.
 246
 247 [0:08:18] [0:03:36] TPV: (...) genau (-) also das sind erstmal so und
 248 noch ein satz ist mir aufgefallen [0:03:40] SV: [hm_hm] [0:03:40]
 249 TPV: [äm] 8.) immer an die satzstellung denken [nE] [0:03:44] SV:
 250 [okay] [0:03:44] TPV: wir haben ja im deutschen (.) können wir das
 251 subjekt ein bisschen verschieben (.) im [englischen] muss das aber
 252 immer vor dem verb kommen [0:03:48] SV: [hm_hm] ja [okay] [0:03:50]
 253 TPV: [ja] das klingt vielleicht komisch aber [0:03:52] SV: also
 254 einfach hier vertau[scheN] [0:03:53] TPV: [genau] ja (...)
 255
 256 [0:08:37] I: Was haben Sie in diesem Moment gedacht?
 257
 258 [0:08:42] TP: Ja, das zeigt ja auch wieder, dass sie mitgedacht hat,
 259 als sie gesagt hat: „Einfach vertauschen“. Ne?
 260
 261 [0:08:47] I: Mhm.
 262
 263 [0:08:48] TP: Sie sitzt da nicht nur passiv da, sondern hört aktiv
 264 zu.
 265

Participant's lack of knowledge

266 [0:08:50] I: Und wussten Sie, nachdem sie das gesagt hat, wo das
 267 Problem so bei ihr war?
 268
 269 [0:08:59] TP: Hinsichtlich der Satzstellung?
 270
 271 [0:09:01] I: Mhm.
 272
 273 [0:09:03] TP: Ja, das ist ja noch sehr dem Deutschen entsprechend.
 274 (unverständlich, ca. 1 Sekunde)
 275
 276 [0:09:07] I: Mhm. Okay.
 277
 278 [0:09:10] [0:03:54] SV: ja (.) [0:03:55] TP: richtig (.) aber schau
 279 hier hast du es auch (.) es gibt sätze wo du das richtig gemacht
 280 hast [0:03:57] SV: hm_hm (.) ja (.) und ist denn da inhaltlich alles
 281 drin [0:04:01] TPV: hm (...)
 282
 283 [0:09:19] I: Was haben Sie in diesem Moment gedacht als diese Frage
 284 kam?
 285
 286 [0:09:23] TP: Ich habe erstmal an mich selber gedacht, ne? Weil ich
 287 ja auch auf die Schnelle jetzt nicht alle Merkmale eines
 288 Zeitungsartikels weiß. Ähm, aber grundsätzlich zeigt das ja auch
 289 wieder, dass sie / sie hat mir am Anfang zugehört, dass es halt
 290 verschiedene Merkmale gibt oder auch Ebenen, auf der man korrigieren
 291 kann. Und sie möchte auch sich verbessern und das ist
 292 (unverständlich, ca. 2 Sekunden)
 293
 294 [0:09:44] I: Mhm.
 295
 296 [0:09:48] [0:04:01] TPV: hm genau (.) das ist jetzt der zweite punkt
 297 nE [0:04:04] SV: [hm_hm] [0:04:05] TPV: [mit] dem inhalt (.) äm (-)
 298 richtig (.) wir hatten ja jetzt wie gesagt die kriterien besprochen
 299 (.) ich habe dir nochmal ein beispiel (.) artikel mitgebracht
 300 [0:04:14] SV: ja [0:04:15] TPV: wir schauen mal was du alles richtig
 301 gemacht hast (.) hier gibt_s eine überschrift und das ist deine
 302 über[schrift] [0:04:19] SV: [hm_hm] [0:04:20] TPV: nE [0:04:21] SV:
 303 [genau] [0:04:21] TPV: [da] hast du alles richtig gemacht (.) da

304 hast du dir eine überschrift ausgesucht [0:04:24] hm_hm [0:04:25]
 305 TPV: sehr gut ja (.) dann äm (.) fällt dir was auf was hier anders
 306 ist nur vom [ersten] [0:04:31] SV: [hm] [0:04:32] TPV: drauf also
 307 vom ersten draufschauen [0:04:34] SV: h° ne ich habe da jetzt
 308 nicht_s unterstrichen zum beispiel (...)
 309
 310 [0:10:24] I: Was haben Sie in diesem Moment gedacht?
 311
 312 [0:10:28] TP: Hm, das war halt gerade nicht die Antwort, die ich
 313 hören wollte ((lacht)). Aber ja sie hat ja trotzdem überlegt und mir
 314 dann was gesagt. Also, sie macht mit, ne? Sie ist da, ähm,
 315 interessiert wie gesagt. Sich zu verbessern.
 316
 317 [0:10:40] I: Mhm.
 318
 319 [0:10:45] [0:04:37] TPV: ja gut (.) das war für euch nur so eine
 320 hilfe wegen den vokab[eln] [0:04:40] SV: [ja] [0:04:41] TPV: aber
 321 schau mal hier hast du eins zwei drei paragrpaph[en (.) ab]schnitte
 322 [0:04:44] SV: [ach so] ach so (.) ja [0:04:45] TPV: und bei dir ist
 323 es halt ein text [0:04:47] SV: hm_hm [0:04:48] TPV: und das nE (.)
 324 das wäre besser wenn du dann in der klassenarbeit wenn das dann
 325 drankommt das du auch dann in [abschnit]ten arbeitest [0:04:52] SV:
 326 [okay] [0:04:53] TPV: hier ist die einleitung den hauptteil und den
 327 schluss [0:04:57] SV: [ja] [0:04:57] TPV: [das] ist wirklich eine
 328 ganz einfach sache die du umsetzen kannst [0:04:59] SV: hm_[hm]
 329 [0:05:00] TPV: [und] da kriegst du dann schon punkte dafür das du
 330 dich an die merkmale gehalten hast [0:05:03] SV: wo müsste ich die
 331 jetzt hier machen in dem text (...)
 332
 333 [0:11:13] I: Was haben Sie in diesem Moment gedacht?
 334
 335 [0:11:17] TP: Ja, wusste ich gerade selber nicht, wo sie das hätte
 336 machen können ((lacht)). Aber, ähm, das ist das gleiche wie vorhin.
 337 Sie ist interessiert und stellt Nachfragen, ja? Sie nickt nicht nur
 338 einfach so und sagt: „Ja, ja, ja“.
 339
 340 [0:11:28] I: Mhm.
 341

342 [0:11:29] TP: Wie man das von anderen Schülern in dem Alter erwarten
 343 könnte. Also, ja.
 344
 345 [0:11:35] I: Mhm.
 346
 347 [0:11:35] TP: Das ist gut.
 348
 349 [0:11:36] I: Und wussten Sie, was Sie hier machen sollten? Nach
 350 dieser Frage? Also, was Sie als nächstes sagen wollen? Oder tun?
 351
 352 [0:11:46] TP: Ja, das war ein bisschen schwierig, weil der Text das
 353 auch jetzt nicht so hergegeben hat, die zwei Teile einzuteilen. Aber
 354 man muss ja trotzdem, ähm, was Positives der Schülerin rückmelden.
 355 Aber ich fand's eine schwierige Situation dann.
 356
 357 [0:12:02] I: Mhm.
 358
 359 [0:12:05] [0:05:06] TPV: ja (4.0) genau du könntest zum beispiel (.)
 360 es geht ja erstmal darum um was ging es überhaupt [0:05:13] SV:
 361 [hm_hm] [0:05:14] TPV: [nE] (-) zum beispiel hier könntest du jetzt
 362 die einleitung (-) [ja] [0:05:18] SV: [hm_hm] [0:05:18] TPV: hier
 363 ist die einleitung nE das ist der hauptteil [0:05:22] SV: ja
 364 [0:05:23] TPV: und dann zum schluss was der minister gesagt hat
 365 [0:05:23] SV: ach so (.) okay ja [0:05:27] TPV: genau (.) [0:05:29]
 366 SV: hm_hm [0:05:29] TPV: und äm (.) dann was hatte ich mir noch
 367 aufgeschriebeN (.) genau die überschrift hatte ich (.) die
 368 paragraphen (.) ja und du kannst im englischen ((lacht)) äm den
 369 englischen artikel nicht einfach auf deutsch schreiben nE das
 370 hättest du [übersetzen] müssen [0:05:41] SV: [ja] [0:05:42] TPV:
 371 aber natürlich schon [(unverständlich, ca. 2 Sekunden)] [0:05:43]
 372 SV: [ach so] ich dachte nur weil das in der aufgabenstellung auch so
 373 steht [0:05:47] TPV: in der aufgabenstellung (...)
 374
 375 [0:12:46] I: Was haben Sie in diesem Moment gedacht?
 376
 377 [0:12:47] TP: Ich habe mich gewundert, welche Aufgabenstellung sie
 378 meint, weil ich hatte jetzt gar keine gesehen. Kann auch sein, dass
 379 ich das überlesen hatte. Und ähm, habe mir vielleicht überlegt, dass

380 wenn's dann doch eine gab, dass man die als Lehrkraft hätte
 381 irgendwie eindeutiger formulieren können, damit das klarer wird. Ja.
 382 Oder wie man das besprechen kann.
 383
 384 [0:13:07] I: Mhm. Also Sie haben das (unverständlich, ca. 1 Sekunde)
 385
 386 [0:13:10] TP: Ja. Hatte ich da nicht im Kopf.
 387
 388 [0:13:11] I: Mhm. Okay. Und wussten Sie danach, was Sie jetzt machen
 389 sollten, nachdem diese Aussage kam?
 390
 391 [0:13:19] TP: Hm. Nein, nicht so genau, weil ich eben verwirrt war.
 392 Weil ich hatte nicht das Gefühl es gab eine Aufgabenstellung auf dem
 393 Material.
 394
 395 [0:13:27] I: Mhm. Okay.
 396
 397 [0:13:31] [0:05:52] TPV: hm_hm (.) äm (-) die aufgaben (.) also
 398 einfach nur merken du hättest das erwähnen müssen bei deinem text
 399 [aber] [0:05:59] SV: [hm_hm] [0:06:00] TPV: vielleicht war das ein
 400 bisschen missverständlich dann die [aufgaben]stellung [0:06:02] SV:
 401 [okay] ja also einfach auf deutsch danN äh [auf englisch] [0:06:05]
 402 TPV: [auf englisch] [0:06:06] SV: ok[ay]
 403
 404 [0:13:52] I: Was haben Sie in dem Moment gedacht, wo sie sagt: „Also
 405 einfach auf Englisch“?
 406
 407 [0:13:59] TP: Hm. Ja, das ist glaube ich zwar eine deutsche Aussage,
 408 also das Zitat des Politikers /, dass das gar nicht so einfach geht.
 409 Also, dass halt / das hat die Schülerin nicht ganz so hinterblickt,
 410 was sie da machen muss.
 411
 412 [0:14:13] I: Mhm.
 413
 414 [0:14:13] TP: Das das einfach gar nicht so einfach geht, sondern
 415 dass das ein kompliziertes Zitat ist, was dann irgendwie leichter
 416 verständlich umgeschrieben werden kann.
 417

Learner's motivation/cognition
Learner's motivation/cognition

418 [0:14:21] I: Mhm.
419
420 [0:14:24] [0:06:06] TPV: [genau] ja (.) und da hätte man sich auch
421 nochmal äm (.) die (.) die wörtliche rede das stimmt nicht so ganz
422 überein wie man das macht (.) mit kommasetzung und auführungsstriche
423 nE (.) aber das können wir uns auch nochmal angucken [0:06:17] SV:
424 okay (.) also das üben wir nochmaL [0:06:19] TPV: ja ich denke mal
425 (...)
426
427 [0:14:39] I: Sie sagte: „Das üben wir nochmal.“ Was haben Sie da
428 gedacht?
429
430 [0:14:44] TP: Ja, dass sie das ja können wird, wenn man das nochmal
431 übt. Weil wörtliche Rede ist ja ein Thema, das macht man irgendwie
432 einmal und dann nie wieder. Dann ist das klar, dass man das schnell
433 vergisst. Aber wenn das halt für diesen Zeitungsartikel wichtiges
434 Thema ist, dann wird sie das können, wenn man das nochmal geübt hat.
435
436 [0:14:59] I: Okay.
437
438 [0:15:00] [0:06:19] TPV: (...) wenn noch ein zwei andere schüler mir
439 das geben und da sehe ich den gleichen fehler dann ist das auf alle
440 [fälle] [0:06:24] SV: [ja] [0:06:26] TPV: was wir wiederholen können
441 (-) genau aber ansonsten (.) also die grundstruktur hast du auf alle
442 [fälle] erfüllt [0:06:30] SV: [hm_hm] okay [0:06:31] TPV: ja (.)
443 hier hatte ich nochmal angeschrieben (.) schau mal das sind alles
444 sachen die du erwähnt hast [0:06:37] SV: hm_hm [0:06:38] TPV: und
445 nur zwei sachen haben gefehlt (.) das die eltern donations gemacht
446 haben (.) weißt du was donations bedeuteT [0:06:43] SV: ähm (.) das
447 war doch das die uns geld gegeben habeN [0:06:46] TPV: genau das hat
448 gefehlt (-) und äm (.) das das in einem gymnasium [stattgefunden]
449 hat [0:06:52] SV: [ja] [0:06:53] TPV: das ist ganz wichtig bei einem
450 zeitungsbericht das du schreibst was passiert und wo passiert ist
451 [0:06:56] SV: hm_hm [0:06:57] TPV: diese we: fragen (.) die wir dann
452 [besprochen] haben [0:07:00] SV: [ja okay] [0:07:01] TPV: das du an
453 die nochmal denkst (.) aber das meiste hast du ja alles sehr schön
454 erwähnt [0:07:04] SV: okay also vom inhalt und der struktur war das
455 schon richtiG (...)

Learner's motivation/cognition

Learner's motivation/cognition

„Confirming the necessity

„Positive

456
 457 [0:15:50] I: Was haben Sie in diesem Moment gedacht, wo sie das
 458 nochmal gefragt hatte?
 459
 460 [0:15:53] TP: Hm. Ja, da habe ich mir überlegt: „Wie kann ich ihr
 461 eine hilfreiche Antwort geben, die sie halt auch nicht entmutigt?“
 462 Weil es war ja eigentlich sehr viel / war relativ viel verkehrt.
 463 Aber, ähm, ich wollte sie halt nicht so entmutigen, weil sie halt
 464 wirklich auf mich zugekommen ist. Hätte sie ja nicht machen müssen.
 465
 466 [0:16:16] I: Mhm. Und wussten Sie nach der Aussage, was sie jetzt
 467 tun wollen?
 468
 469 [0:16:23] TP: Also bei der / schwierig.
 470
 471 [0:16:24] I: Mhm.
 472
 473 [0:16:27] TP: Was ich tun will?
 474
 475 [0:16:29] I: Ja, also als sie nochmal gefragt hat. Wussten Sie da
 476 ganz genau, wie Sie / was Sie jetzt darauf antworten wollen?
 477
 478 [0:16:35] TP: Ich denke im Großen und Ganzen schon, ja. Vielleicht,
 479 ähm, mehr auf die Sprache bezogen, also auf die, ähm, Kriterien des
 480 Zeitungsartikels. Aber im Großen und Ganzen schon.
 481
 482 [0:16:47] I: Mhm. Okay.
 483
 484 [0:16:48] [0:07:09] TPV: ja (-) äm (.) du hast vieles richtig
 485 gemacht schon [0:07:12] SV: hm_hm [0:07:13] TPV: aber luft nach oben
 486 gibt_s [natürlich] immer noch nE [0:07:14] SV: [okay] ja [0:07:15]
 487 TPV: deshalb an ein zwei punkte noch dran halten nächstes mal
 488 [0:07:18] SV: ja (.) muss ich den text denn jetzt äh noch neu
 489 schreiben (...)
 490
 491 [0:17:02] I: Was haben Sie in dem Moment gedacht als diese Frage
 492 kam? Dass sie den Text neu schreiben muss.
 493

„Confirming the necessity
„Positive

494 [0:17:10] TP: Hm. Ja muss. Ähm. Also ich habe mir halt den
495 Verlaufsplan der Unterrichtsreihe angeguckt und da stand ja drin,
496 dass die halt als nächstes / in der nächsten Stunde in der Klasse
497 zusammen gemacht wird. Deshalb muss sie das nicht jetzt machen, zu
498 dem Zeitpunkt. Ähm, genau. Und halt auch wieder, sie zeigt
499 Arbeitsbereitschaft. Also sie ist willig da dran zu arbeiten.
500
501 [0:17:30] I: Mhm.
502
503 [0:17:31] TP: Was ja auch sehr löblich ist.
504
505 [0:17:33] I: Und können Sie sich vorstellen, was / worauf auf einer
506 didaktisch-fachlichen Ebene so eine Frage gezielt worden ist?
507
508 [0:17:44] TP: Vielleicht auch einfach das Neuschreiben dann zur
509 Verbesserung führt.
510
511 [0:17:48] I: Mhm. Okay.
512
513 [0:17:53] [0:07:22] TPV: äm (.) wir machen das ja noch (.) du hast
514 jetzt von mir feedback bekommen und in der nächsten stunde kriegst
515 du ja auch von deinen mitschülerinnen feedback [0:07:29] SV: hm_hm
516 [0:07:30] TPV: und das kannst du dann sammeln am besten wenn du
517 jetzt hier rausgehst mit dem gespräch (.) mach dir ein paar
518 stichpunkte zu dem was ich gesagt [habe] [0:07:36] SV: [ja]
519 [0:07:37] TPV: (.) zu (.) zu den tipps das du die nicht vergisst (.)
520 morgen gibst du noch tipps (.) kriegst du noch tipps von deinen
521 mitschülerinnen (.) und äm (-) genau und dann kannst du das
522 anschließend nochmal überarbeiten [0:07:46] SV: okay [0:07:47] TPV:
523 wenn du dann alle tipps hast [0:07:48] SV: hm_hm (.) äm können sie
524 nochmal vielleicht ganz kurz zusammenfassen was denn jetzt das
525 wichtigste nochmal waR (.) damit das ein guter zeitungsartikel ist
526 [0:07:56] TPV: genau (...)
527
528 [0:18:28] I: Was haben Sie in diesem Moment gedacht?
529
530 [0:18:29] TP: Das ist eine gute Frage von ihr. Äh, ja hilfreich für
531 sie.

532
 533 [0:18:37] I: Und wussten Sie in dem Moment schon, was Sie dann
 534 darauf antworten sollen?
 535
 536 [0:18:45] TP: Hm. Ich glaube, ich musste nochmal kurz überlegen.
 537 Ähm, ich weiß nicht. Also ich hatte mir ja schon Stichpunkte
 538 gemacht. Ich hätte einfach nochmal die Stichpunkte wiederholt. Das
 539 ist ja auch für die Schülerin ganz gut dann, wenn sie das zwei,
 540 dreimal hört.
 541
 542 [0:19:00] I: Mhm.
 543
 544 [0:19:01] TP: Dann kann sie sich das besser merken.
 545
 546 [0:19:08] I: Okay.
 547
 548 [0:19:10] [0:07:56] TPV: (...) du schaust dir nochmal die
 549 unregelmäßigen verben an [0:07:59] SV: hm_hm [0:07:59] nE habe ich
 550 auch nochmal drunter geschrieben unregelmäßige verben (.) machst
 551 bitte die paragraphen rein [0:08:05] SV: ja [0:08:06] TPV: ja und
 552 (.) ah ich hatte noch vergessen (.) die eigene meinung (.) die kommt
 553 nicht in einem zeitungsbericht vor [0:08:13] [ach so] [0:08:13] TPV:
 554 [das wäre] dann eher ein (unverständlich, ca. 1 Sekunde) zum
 555 beispiel wenn du ein tagebucheintrag schreibst nE [0:08:15] SV:
 556 hm_hm [0:08:15] TPV: aber ein zeitungsbericht soll ja objektiv
 557 bleiben [0:08:18] SV: ach so ist das dieser satz hier dann [0:08:19]
 558 TPV: hm_hm genau (.) it was an amazing event (.) das glaube ich dir
 559 das war ja auch wirklich toll [0:08:23] SV: hm_hm [0:08:24] TPV:
 560 aber das schreibt man nicht in [einem] zeitungsbericht [0:08:25] SV:
 561 [ach so] okay (.) aber ist ist das mit amazing (.) ist das da in
 562 ordnung
 563
 564 [0:19:44] I: Was haben Sie in dem Moment gedacht?
 565
 566 [0:19:49] TP: Wo sie gefragt hat: „Hier ist das in Ordnung?“
 567
 568 [0:19:51] I: Mhm.
 569

570 [0:19:54] TP: Ja (unverständlich, ca. 1 Sekunde) so positiv von ihr
 571 war, weil das war ja das gleiche Wort. Nur einmal ist es eben
 572 angebracht und einmal nicht.
 573
 574 [0:20:00] I: Mhm.
 575
 576 [0:20:01] TP: Und dass ihr das auffällt. Also (unverständlich, ca. 1
 577 Sekunde - in dem Text?)
 578
 579 [0:20:03] I: Mhm.
 580
 581 [0:20:06] [0:08:29] TPV: ja das ist ja so eine art zitat [0:08:31]
 582 SV: hm_hm [0:08:32] TPV: des äm (.) direktors (-) und du hast es
 583 halt mit amazing übersetzt (.) hätte man auch anders machen können
 584 aber das ist schon hier in Ordnung auch [0:08:39] SV: okay (.) ja
 585 [0:08:41] TPV: okay hast du noch fragen [0:08:42] SV: äm (.) ne ich
 586 glaube nicht [0:08:45] TPV: ja soweit verständlich [0:08:46] SV: hm
 587 [0:08:47] TPV: also wie gesagt schau dir nochmal die unregelmäßigen
 588 verben an (.) die paragraphen und an diesem satz nE da frag auch
 589 gerne nochmal deine mitschülerinnen [morgen] [0:08:53] SV: [hm_hm]
 590 [0:08:54] TPV: ob die dir den übersetzen können (.) vielleicht haben
 591 die ja schon eine idee gehabt [0:08:57] SV: ach so (.) ja okay [das
 592 könnten wir machen] [0:08:58] TPV: [das ihr da] zusammen dran
 593 arbeitet [0:09:00] SV: hm_hm [0:09:01] TPV: [okay] [0:09:01] SV:
 594 [okay] gut (.) dankeschön [0:09:03] TPV: ja super dann sehen wir uns
 595 morgen [0:09:04] SV: danke (.) tschüss [0:09:06] Ende Dialog.
 596
 597 [0:20:44] I: Was haben Sie gedacht als es zu Ende war?
 598
 599 [0:20:51] TP: Ja. Ja also ich hoffe, dass, hm, dass sie verstanden
 600 hat oder dass es verständlich genug war mein Feedback, dass sie
 601 versteht, woran sie arbeiten muss. Was das / was die wichtigsten
 602 Punkte sind.
 603
 604 [0:21:04] I: Mhm. Okay. Gut, dann war's das auch schon für den
 605 ersten Teil.
 606
 607 [0:21:11] TP: Mhm.

608
609 [0:21:12] I: Ich stoppe diese Aufnahme einmal.
610
611 Ende.
612
613
614
615
616
617
618

G.2. Simulation

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Transkript der Gesprächssimulation Englisch

Pseudonymisierungscode

MMM07

Länge

00:07:27 Minuten

Aufnahmeort

Universität Paderborn

Transkriptionskonvention

Erweiterte Inhaltlich-semantiche Transkription

Englische Sprachliche Norm der Versuchsperson

☒ Received Pronunciation ☐ General American English

Ankerbeispiel und Timecode

/rɪpt/ [0:01:21]

Interviewkürzel

TP: Testperson | S: Schülerin

Videoevent	Audioevent
1. [0:00:00] Vorbereitung	1. [0:00:00] Vorbereitung
2. [0:00:12] TP: legt beide Arme auf den Armlehnen ab und richtet den Blick in die Richtung der Tür	1. [0:00:13] (klopfen) 2. [0:00:14] TP: JA heREIN
3. [0:00:14] TP: bewegt den Oberkörper nach vorne, richtet den Blick dabei nach unten, legt die Hände auf dem Tisch neben dem Material ab	1. [0:00:16] S: morning
4. [0:00:15] TP: richtet den Blick in die Richtung der S	1. [0:00:17] TP: morning (...) [mia] 2. [0:00:18] S: [räuspern] JA

1/2

1/20

	5. [0:00:16] TP: lehnt sich zurück und legt die Arme auf den Armlehnen ab 6. [0:00:17] S: geht zum Stuhl 7. [0:00:18] TP: richtet den Blick auf das Material und berührt mit der linken Hand den unteren Teil von M2, legt dann auch den zweiten Arm auf das Material (?) 8. [0:00:19] S: zieht den Stuhl zurück und setzt sich darauf 9. [0:00:24] TP: richtet den Blick abwechselnd in die Richtung der S und auf das Material 10. [0:00:24] S: richtet den Blick abwechselnd in die Richtung der TP und auf das Material	1. [0:00:21] TP: first of all thank you for handing in a first written text of the class
	1. [0:00:32] TP: verschränkt die Hände miteinander und legt diese auf den Beinen ab 2. [0:00:33] S: hebt leicht die Mundwinkel	1. [0:00:25] S: mhm 2. [0:00:26] TP: that takes a lot of courage ähm what do you think about your TEXT (.) itself 3. [0:00:32] S: mh können sie die frage nochmal auf deutsch stellen 1. [0:00:34] TP: natürlich ähm (..) wie ähm findest du denn selber den TEXT (räuspern) 2. [0:00:41] S: mh also eigentlich ganz okay ich habe halt versucht alles mit reinzubringen aber an manchen stellen bin ich mir nicht so sicher 3. [0:00:48] TP: mhm 4. [0:00:49] S: (räuspern) 5. [0:00:49] TP: von mir aus können wir tatsächlich das ganze gespräch auch auf deutsch machen das ist immer ein bisschen

Language Adaptation: 2		persönlicher 6. [0:00:54] S: ja gerne (lachen) 7. [0:00:54] TP: englisch kreiert ja ein bisschen so mehr distanz [ähm]
	1. [0:00:42] TP: berührt mit beiden Händen kurz das Material und verschränkt diese dann wieder vor sich	8. [0:00:57] S: [mhm] 9. [0:00:58] TP: du bist ähm also ganz zufrieden mit deinem text hast du gesagt [ähm]
Dialogue orientation	1. [0:00:53] S: nickt mit dem Kopf 2. [0:00:54] S: hebt die Mundwinkel	10. [0:01:01] S: [mh ja] 11. [0:01:02] TP: wenn das für dich okay ist wenn du jetzt vorher keine nachfragen hast würde ich jetzt ein bisschen feedback geben 12. [0:01:07] S: mh [ja]
	1. [0:00:56] TP: bewegt das Material etwas	1. [0:01:08] TP: [ähm] und zwar ist mir erstmal aufgefallen ihr hattet ja hier dieses ähm arbeitsblatt dazu [mit den]
Positive feedback	1. [0:01:00] TP: legt die Arme auf der Armlehne ab	1. [0:01:12] S: [mhm]
Prioritization of deep	1. [0:01:07] S: nickt mit dem Kopf und richtet den Blick auf das Material 2. [0:01:07] TP:	2. [0:01:13] TP: informationen drauf und ich finde es gut dass du wirklich viele informationen dazu ähm genutzt hast um in den text zu integrieren

		richtet den Blick auf das Material	
..(Feed Up) Learnin		1. [0:01:10] TP: hebt leicht die untere linke Ecke von M2 mit der linken Hand	1. [0:01:19] S: hmh 2. [0:01:19] TP: das ist ja ein news report also ist es auch wichtig dass viele informationen drinne sind
..Condition: Quality		2. [0:01:13] TP: legt die Hand neben M2	3. [0:01:24] S: mhm
		3. [0:01:15] TP: legt den Arm auf der Armlehne ab	4. [0:01:25] TP: das hast du sehr gut gemacht
		4. [0:01:16] TP: richtet den Blick in die Richtung der S	5. [0:01:26] S: okay 6. [0:01:26] TP: ähm (..) genau ähm (..) hier ist ja auch ähm steht ja auch dass ihr die die struktur benutzen solITET [ähm würdest du]
..Prioritization o		5. [0:01:16] S: richtet den Blick in die Richtung der TP und nickt mit dem Kopf	1. [0:01:36] S: [JA]
..Use of materia		6. [0:01:17] TP: richtet den Blick auf das Material	2. [0:01:36] TP: sagen dass das hier ähm wie es hier aufgebaut ist die struktur ist die in einem news report gebraucht werden solLTE (...)
..Deep: Structur		7. [0:01:18] S: richtet den Blick auf das Material und nickt mit dem Kopf	1. [0:01:45] S: mh JA
..Quality of feed		1. [0:01:24] TP: richtet den Blick kurz in die Richtung der	1. [0:01:46] TP: mhm 2. [0:01:46] S: eigentlich schon

			S und dann auf das Material	(lachen)
..Dialogue o				3. [0:01:47] TP: ja ähm kannst du dich noch an deinen text erinnern und würdest sagen dass du das ähm so auch genutzt HAST (.)
..Use of mat			1. [0:01:27] TP: berührt mit der linken Hand die untere Kante von M2 und hebt und senkt diese mehrmals	
			2. [0:01:32] TP: zeigt auf M2	1. [0:01:54] S: mh ja ich habe es auf jeden fall versucht
			3. [0:01:34] TP: legt den Arm auf der Armlehne ab	2. [0:01:55] TP: mhm (.) ich ich finde auch dass äh die struktur dir eigentlich ganz ähm ganz in ordnung gelungen ist
..Prioritizatio			1. [0:01:37] TP: legt die rechte Hand flach auf M2	1. [0:02:01] S: okay
..Deep: Struc			2. [0:01:38] TP: tippt mit dem Zeigefinger der rechten Hand mehrmals auf M2	2. [0:02:01] TP: ähm was am ende oder was vielleicht ein bisschen zu kurz gekommen ist wenn ich dir das mal so als tipp mitgeben darf
..Quality of fi			3. [0:01:40] TP: legt den Arm auf der Armlehne ab	3. [0:02:07] S: hmh
			4. [0:01:41] TP: richtet den Blick abwechselnd in die Richtung der S und auf das Material	4. [0:02:07] TP: ist dass ähm dieser aim dieses ziel [ähm]
..Deep: Cont			5. [0:01:45] S: hebt die Mundwinkel und richtet den Blick in die Richtung der TP	1. [0:02:11] S: [mhm] 2. [0:02:12] TP: hättest du vielleicht noch ein bisschen mehr informationen mit in den text schreiben können 3. [0:02:16] S: okay also inhaltlich

..Prioritization o Deep: Content ..Deep: Structure ..Quality of feed for ..Error patterns	und dann auf das Material 6. [0:01:46] TP: richtet den Blick auf das Material 1. [0:01:48] TP: berührt mit der rechten Hand M2 2. [0:01:51] TP: legt den Arm auf der Armlehne ab 3. [0:01:53] TP: richtet den Blick in die Richtung der S 1. [0:01:56] TP: richtet den Blick auf das Material 2. [0:01:57] TP: berührt mit der rechten Hand M2 und legt den Arm wieder auf der Armlehne ab 3. [0:02:00] TP: richtet den Blick abwechselnd in die Richtung der S und auf das Material 1. [0:02:07] S: richtet den	[hier] 4. [0:02:17] TP: [genau] 5. [0:02:18] S: okay mhm 6. [0:02:18] TP: vielleicht noch ähm weil ähm hier wird ja so richtig klar ähm (.) warum eure schule diesen preis äh gewonnen hat NE ihr habt ja fünfhundert bäume gepflanzt [du hast] 1. [0:02:28] S: hmh 2. [0:02:29] TP: nur als kleiner tipp vielleicht ähm nächstes mal da noch ein bisschen mehr informationen hier von miteinbringen 1. [0:02:33] S: okay 2. [0:02:33] TP: ähm und vielleicht mehr betonen weil das ist ja echt was besonderes dass unsere schule diesen preis gewonnen hat ähm warum die schule jetzt gewonnen hat (.) und ähm oder warum vielleicht warum diese aktion auch überhaupt gemacht wird 1. [0:02:47] S: mh

6/20

...Error patterns ...Dialogue orientation		Blick kurz in die Richtung der TP, nickt mit dem Kopf und richtet den Blick auf das Material 2. [0:02:07] TP: richtet den Blick auf das Material 3. [0:02:09] TP: streckt die Finger von der linken Hand ab und berührt M2 4. [0:02:12] TP: legt den Arm auf der Armlehne ab 5. [0:02:14] TP: richtet den Blick in die Richtung der 6. [0:02:16] TP: richtet den Blick auf das Material 7. [0:02:17] S: zeigt auf das M2 1. [0:02:21] TP: bewegt die rechte Hand über M2 2. [0:02:24] TP: legt den Arm auf der Armlehne ab 3. [0:02:26] TP: richtet den Blick in die Richtung der S und dann auf das Material	okay 2. [0:02:48] TP: ähm genau ähm ansonsten ist mir mh sind mir nur also grammatikmäßig nur so ein paar sachen aufgefallen 3. [0:02:47] S: was DENN 1. [0:02:58] TP: das ist eigentlich ganz gut grammatik ist ähm vor allem das simple past was mir aufgefallen ist 2. [0:03:02] S: mh ja da war ich mir unsicher [an manchen Stellen 3. [0:03:04] TP: [JA] (.) ähm falls wir da im unterricht nochmal zeit für haben ähm oder ich weiß nicht nimmst du nachhilFE selbst oDER [zur zeit NICHT] 4. [0:03:13] S: [eigentlich nicht] 1. [0:03:14] TP: vielleicht ähm kann man das nochmal kurz anreißen auf jeden fall ähm war das jetzt das auffälligste da würde ich dir vielleicht als tipp mitgeben nochmal die ähm unREGelmäßigen verben zu lernen

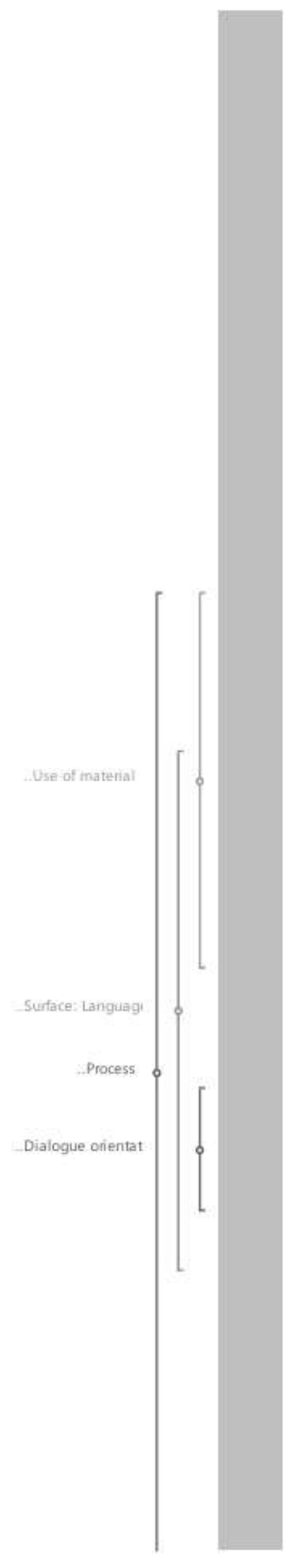
4. [0:02:26] S: nickt mit dem Kopf 1. [0:02:31] TP: berührt mit dem Daumen und Mittelfinger ausgestreckt M2 2. [0:02:32] TP: legt den Arm auf der Armlehne ab 1. [0:02:37] TP: hebt beide Hände etwas über Tischhöhe und öffnet die Handflächen zu einander 2. [0:02:38] TP: richtet den Blick in die Richtung der S und legt die Arme auf der Armlehne ab 3. [0:02:39] S: nickt mit dem Kopf 4. [0:02:39] TP: richtet den Blick auf das Material 1. [0:02:48] TP: berührt mit beiden Händen die Tischkante	1. [0:03:24] S: mhm 2. [0:03:25] TP: ähm und dass du da weißt wie man von von der infinitiv die vergangenheit formt 1. [0:03:30] S: aber wie kann ich das den lERNEN 2. [0:03:31] TP: das äh steht zum beispiel hinten in den büchern immer 3. [0:03:34] S: hmh 4. [0:03:34] TP: ähm die vokabeln werden wahrscheinlich jetzt nicht mehr in den units der vokabeln sein aber ganz hinten im buch [ist] 5. [0:03:40] S: [hmh] 6. [0:03:40] TP: äh so eine seite da ist so eine ganz große tabelle 1. [0:03:43] S: okay (räuspern) 2. [0:03:44] TP: und da sind immer die unregelmäßigen verben aufgelistet ähm hinter dem infinitiv 3. [0:03:48] S: mhm okay 4. [0:03:48] TP: das heißt da kann man das nochmal ganz gut lernen weil bei denen ist es ja leider so dass man die auswendig lernen muss

8/20

..Surface: Language Appr ..Condition: Dialog ..AEC: 1 Point (ope	1. [0:02:50] TP: legt die Arme auf der Armlehne ab 2. [0:02:56] TP: richtet den Blick abwechselnd in die Richtung der S und auf das Material 1. [0:03:13] S: bewegt den Kopf nach rechts und links hin und her 2. [0:03:13] TP: richtet den Blick auf das Material 1. [0:03:16] TP: berührt M2 mit der rechten Hand 2. [0:03:20] S: richtet den Blick abwechselnd in die Richtung der TP und auf das Material 3. [0:03:23] TP: richtet den Blick abwechselnd in die Richtung der	5. [0:03:53] S: hm 6. [0:03:53] TP: da gibt es ja keine regeln für (räuspern) [genau] 1. [0:03:56] S: [okay] wo habe ich das denn falsch geMACHT 1. [0:03:58] TP: ähm zum beispiel als äh beispiel einmal hier oben ähm du erzählst hier ja über die zeremonie 1. [0:04:04] S: mhm 2. [0:04:05] TP: und die war ja am diensttag das heißt das liegt in der vergangenheit [deswegen] 3. [0:04:08] S: [ah okay] 4. [0:04:08] TP: musst du hier nicht win nutzen sondern das unregelmäßige verb vielleicht weißt du es grad auch selber aus dem KOPF 1. [0:04:13] S: nein weiß ich nicht 2. [0:04:14] TP: okay ähm das ist dann won 3. [0:04:16] S: mhm

...Surface: Language Core	S und auf das Material 4. [0:03:24] S: nickt mit dem Kopf 1. [0:03:28] TP: berührt mit der linken Hand zweimal (M0?) 1. [0:03:38] TP: legt die linke Hand hochkant auf (M0?) ab 1. [0:03:40] TP: dreht die Hand mit der Handfläche zum Material 2. [0:03:42] TP: bewegt die Hand die Tischmitte berührend nach links und rechts 1. [0:03:45] TP: hebt und senkt die Hand und bewegt die Hand mehrmals nach links und rechts 2. [0:03:48] S: nickt einmal mit dem Kopf 3. [0:03:49] TP: legt die Arme auf der Armlehne ab	4. [0:04:16] TP: also w o n (.) wie man sieht gibt es da keine regeln für man kann nicht einfach ein ed hinten dranhängen wie das normalerweise ist 1. [0:04:23] S: mhm 2. [0:04:23] TP: das ist dann so ein beispiel [ähm] 1. [0:04:25] S: [okay] 2. [0:04:26] TP: genau oder zum beispiel auch ähm hast du geschrieben we become even one hunderd thousand euros prize money ähm das ist auch ein unregelmäßiges wort ähm (..) da kann ich dir jetzt auch sagen das bedeutet we äh das muss heißen we became 3. [0:04:42] S: mhm 4. [0:04:43] TP: aber das kannst du alles hinten im buch nachschauen und das kann man auch echt gut lernen mit so

10/20

	1. [0:03:54] S: hebt die Mundwinkel 2. [0:03:55] TP: richtet den Blick auf das Material 1. [0:03:57] S: richtet den Blick auf das Material und lehnt sich näher dahin 2. [0:03:59] TP: legt M1 zwischen sich und S 3. [0:04:01] TP: die linke Hand liegt neben M1 und zeigt mit der rechten Hand auf M1 4. [0:04:03] TP: winkelt den rechten Arm an, sodass die rechte Hand neben M1 liegt und zeigt mit der linken Hand auf M1 5. [0:04:07] TP: richtet den Blick in die Richtung der TP und dann auf das Material 6. [0:04:08] TP:	einer tabelle 5. [0:04:46] S: ja 6. [0:04:47] TP: die seiten zu halten und sich selber das abfragen 1. [0:04:49] S: ja okay mh 2. [0:04:50] TP: als kleiner tipp das ähm (.) wird in dem nächsten text auf jeden fall schon bereichern 3. [0:04:56] S: mhm 4. [0:04:56] TP: ähm genau ähm und (.) noch als (.) als letzter tipp den ich dir mit auf den weg geben will ähm (.) habe ich hier einen text mitgebracht ähm den wir in der (..) stunde oder in der unterrichtsreihe schon einmal durchgenommen habt 1. [0:05:12] S: mhm 2. [0:05:12] TP: ich weiß nicht ob du dich erinnerst 3. [0:05:13] S: ja (räuspern) 4. [0:05:14] TP: ähm hier sind ja auch so ein paar neue vokabeln (räuspern) wahrscheinlich bei 5. [0:05:19] S: mh 6. [0:05:20] TP: auch so ein bisschen was was ähm bäume angeht ähm oder oder generell die natur

..Process	zeigt auf M1 7. [0:04:10] TP: hebt und senkt die Hand 8. [0:04:12] TP: richtet den Blick in die Richtung der S 9. [0:04:13] S: hebt die Mundwinkel 10. [0:04:14] TP: richtet den Blick auf das Material 11. [0:04:16] TP: streckt die Finger von der Hand und knickt sie wieder 12. [0:04:18] TP: streckt die Finger Richtung M1 13. [0:04:19] TP: lehnt sich zurück 14. [0:04:21] TP: zeigt mit der Hand in die Richtung von M1 und legt sie dann vor sich auf dem Tisch ab und richtet den Blick in die Richtung der S und dann auf das Material 15. [0:04:23] S: nickt einmal mit dem Kopf 16. [0:04:24] TP: zeigt mit den Zeige- und	wo es ja hier rum geht [um das Klima] 1. [0:05:27] S: [mhm] 2. [0:05:27] TP: dass du vielleicht ähm ihn nochmal anschaut weil dann kannst du vielleicht ein paar vokabeln hier von den neuen vielleicht sogar in deinen text integrieren 1. [0:05:35] S: ah okay 2. [0:05:36] TP: dass würde das ganze so ein bisschen aufschmücken sage [ich mal] 1. [0:05:38] S: [mhm] 2. [0:05:39] TP: das ist immer ganz gut wenn man dann auch sachen die man neu lernt mit integriert weil man dann auch die vokabeln auch automatisch besser lernt 1. [0:05:45] S: ja

12/20

..Process orientation: 1 Pc ..(Feed Up) Learning o ..Condition: Quality of ..Use of material	Mittelfinger der linken Hand auf M1 1. [0:04:28] TP: zeigt auf M1 2. [0:04:39] TP: zeigt mit allen Fingern der Hand auf M1 1. [0:04:43] TP: legt die Arme auf der Armlehne ab 1. [0:04:47] TP: richtet den Blick in die Richtung der S 2. [0:04:48] TP: legt die Hand flach halb auf den Tisch und halb auf (M0?) 3. [0:04:49] S: nickt mit dem Kopf 4. [0:04:50] TP: richtet den Blick in die Richtung der S und legt M1 vor sich 5. [0:04:56] TP: richtet den Blick in die Richtung der S und dann auf das	okay 2. [0:05:45] TP: ähm genau ähm 3. [0:05:48] S: heißt ich soll den text jetzt einmal neu schreiben und dann versuchen die sachen noch miteinzubauen (.) 4. [0:05:53] TP: wenn du die motivation hast ist das natürlich ein optimaler nächster schritt 1. [0:05:56] S: (lachen) okay 2. [0:05:56] TP: wenn du ähm sowas versuchst aber das soll jetzt nicht bedeuteten dass ich nichts von dem text halte 3. [0:06:02] S: okay 4. [0:06:03] TP: wie gesagt das lernziel war es ja ähm dass ihr erstmal einen ersten entwurf äh ähm erfassen ne entwerfen konntet 1. [0:06:11] S: mhm 2. [0:06:12] TP: ähm und das lernziel hast du auf jeden fall erfüllt 3. [0:06:15] S: okay 4. [0:06:15] TP: ähm es sind nur diese paar tipps die ich dir eben schon gegeben habe [um das]

Use of material

Material	5. [0:06:18] S: [ja]
6. [0:04:56] S: richtet den Blick in die Richtung der TP und dann auf das Material	6. [0:06:18] TP: ganze ein bisschen ähm (.) bisschen ähm zu bereichern
7. [0:04:59] TP: bewegt das Material etwas	1. [0:06:23] S: mhm
8. [0:05:03] TP: legt die Arme auf der Armlehne ab	2. [0:06:24] TP: dann kannst du vielleicht auch nochmal ein paar mehr informationen geben
9. [0:05:05] TP: zieht M4 aus einem Materialstapel heraus und legt es auf M2	3. [0:06:26] S: okay
1. [0:05:16] TP: zeigt mit dem Daumen und Mittelfinger auf die untere Hälfte von M4	4. [0:06:27] TP: ähm genau und ähm einfach noch ein paar informationen ergänzen wie ich das ja eben schonmal meinte von diesem aim part (.) die vokabeln und vielleicht ein bisschen auf die grammatik achten [dann kann man]
1. [0:05:22] TP: zeigt auf M4	5. [0:06:40] S: [ja das versuche ich]
2. [0:05:23] S: nickt mit dem Kopf	6. [0:06:41] TP: äh auf jeden fall was draus machen wie gesagt es ist jetzt der erste entwurf
3. [0:05:24] TP: macht kreisende Handbewegungen mit beiden Händen	7. [0:06:45] S: mh
4. [0:05:25] TP: richtet den Blick in die Richtung der S und deutet auf	8. [0:06:45] TP: und ähm es erfordert ja auch erstmal mut als erstes ähm aus der klasse überhaupt einen text (.) mir zu geben um den zu [korrigieren]
	9. [0:06:53] S: [mhm]

14/20

...Self-regulation		das Material 5. [0:05:26] TP: richtet den Blick auf das Material und legt die Arme auf der Stuhllehne ab	1. [0:06:54] TP: ähm das entwickelt sich auf jeden fall in die richtige richtung ich merke ja bei dir auch dass du jetzt feedback möchtest zu dem [text] 2. [0:06:59] S: [ja] (räuspern)
...Checking understanding		1. [0:05:29] TP: zeigt mit dem Daumen und Mittelfinger der rechten Hand auf die untere Hälfte von M4 2. [0:05:32] TP: legt die Hand an die Tischkante, sodass die Finger M4 berühren 3. [0:05:34] TP: deutet mit der linken Hand auf M1 1. [0:05:37] TP: richtet den Blick in die Richtung der TP 2. [0:05:38] S: richtet den Blick in die Richtung der TP und nickt mit dem Kopf 3. [0:05:39] TP: richtet den Blick auf das Material 4. [0:05:43] TP: bewegt beide Hände in die	1. [0:07:00] TP: um dich selber zu verbessern ähm deswegen will ich dir da auf jeden fall schon dankbar und ich glaube dass das in eine gute richtung gehen kann [wenn du die sachen] 2. [0:07:06] S: [okay] 3. [0:07:08] TP: so beachtest 4. [0:07:09] S: okay 5. [0:07:10] TP: genau (.) dann ähm ich weiß nicht ob du vielleicht noch weitere fragen HAST 6. [0:07:15] S: mh ne eigentlich nicht also ich würde mir den text dann nochmal angucken und dann genau gucken was ich dann verbessern kann 1. [0:07:21] TP: sehr gut 2. [0:07:22] S: ja 3. [0:07:23] TP: gut dann ähm wenn du magst kannst du den gerne auch wieder dann einreichen

Richtung von M4 und zurück auf die Armlehnen und richtet den Blick in die Richtung der S	[dann]
5. [0:05:44] S: nickt mit dem Kopf	4. [0:07:26] S: [mhm]
6. [0:05:44] TP: richtet den Blick auf das Material	5. [0:07:26] TP: gebe ich dir nochmal ein feedback da kann man vielleicht auch so einen prozess erkennen [ähm]
1. [0:05:50] S: zeigt auf M4 und richtet den Blick in die Richtung der TP	1. [0:07:29] S: [ja]
1. [0:05:55] TP: richtet den Blick in die Richtung der S	2. [0:07:30] TP: also wie du dich weiterentwickelt hast
2. [0:05:56] S: hebt die Mundwinkel und richtet den Blick auf das Material	3. [0:07:32] S: okay
3. [0:05:56] TP: richtet den Blick auf das Material	4. [0:07:32] TP: und genau
4. [0:06:01] TP: hält die Hand flach über M1	5. [0:07:34] S: okay
1. [0:06:06] TP: macht mit der linken Hand eine	6. [0:07:35] TP: dann würde ich sagen schöne restpause noch
	7. [0:07:37] S: ja danke (lachen) vielen lieben [dank]
	1. [0:07:39] TP: [und] bis zur nächsten stunde gerne
	2. [0:07:41] S: tschüss
	3. [0:07:41] TP: tschüss

16/20

	kreisende Handbewegung von sich weg 2. [0:06:08] TP: richtet den Blick in die Richtung der TP 3. [0:06:08] S: nickt mit dem Kopf 4. [0:06:09] TP: richtet den Blick in die Richtung der TP 1. [0:06:18] S: nickt mit dem Kopf 2. [0:06:18] TP: richtet den Blick in die Richtung der S und bewegt die Hände umeinander kreisend vor sich 3. [0:06:19] TP: richtet den Blick auf das Material 4. [0:06:22] TP: hebt die Hände und macht eine kreisende Bewegung von sich weg 5. [0:06:24] TP: tippt mit der linken Hand auf M1	

	1. [0:06:34] TP: hebt M4 hoch und legt es auf die linke Seite, sodass M2 sichtbar wird 1. [0:06:51] TP: richtet den Blick in die Richtung der S 2. [0:06:53] S: nickt mit dem Kopf 3. [0:06:53] TP: richtet den Blick auf das Material 1. [0:06:55] S: hebt die Mundwinkel 2. [0:06:57] TP: macht eine kreisende Handbewegung 3. [0:06:59] TP: richtet den Blick in die Richtung der S 4. [0:06:59] S: nickt mit dem Kopf 5. [0:07:00] TP:

	richtet den Blick auf das Material	
	1. [0:07:08] S: nickt leicht mit dem Kopf	
	1. [0:07:15] TP: richtet den Blick in die Richtung der S	
	2. [0:07:17] TP: richtet den Blick auf das Material	
	3. [0:07:19] TP: hebt und senkt den Kopf	
	1. [0:07:25] S: richtet den Blick in die Richtung der TP	
	2. [0:07:26] TP: richtet den Blick abwechselnd in die Richtung der S und auf das Material	
	3. [0:07:27] S: nickt mit dem Kopf	
	4. [0:07:28] TP:	

bewegt die linke Hand von sich weg und hebt und senkt diese dabei mehrmals 5. [0:07:29] S: nickt mit dem Kopf 6. [0:07:30] TP: richtet den Blick auf das Material	
1. [0:07:36] S: hebt die Mundwinkel und nickt mit dem Kopf 2. [0:07:38] TP: hebt die Mundwinkel und richtet den Blick in die Richtung der S 3. [0:07:38] S: erhebt sich vom Stuhl und verlässt den Raum 4. [0:07:39] TP: richtet den Blick auf das Material	

G.3. Guideline Interview– Main Survey

1	Transkript Leitfadeninterview (LF) - Nachbefragung
2	Fachzuordnung
3	☐ Bildungswissenschaften
4	☒ Englisch
5	Pseudonymisierungscode
6	NRS18
7	Länge
8	00:11:19 Minuten
9	Aufnahmeort
10	Universität Paderborn
11	Transkriptionskonvention
12	Inhaltlich-semantische Transkription
13	Interviewkürzel
14	TP: Testperson
15	I: Interviewer
16	[0:00:00] I: Der Code ist N R S 1 8. Wie war die Situation für Sie? Erzählen Sie.
17	[0:00:07] TP: Also es war was ganz Neues für mich. Ähm gerade auch, weil ich in so einem Bereich sehr wenig Erfahrungen habe, gerade was auch ein Interview angeht. Ähm das ganze dann aber nochmal so auf Englisch und auf Deutsch quasi gemixt zu machen war nochmal eine Stufe, wo ich gesagt habe ‚Okay, das ist etwas Neues für mich, was mir auch im Leben sehr weiterhelfen wird.‘ Ähm ich werde davon auf jeden Fall eine Menge mitnehmen auch an jetzt an Erfahrungen. Ähm ja im Großen und Ganzen war es sehr gut für mich. Äh natürlich könnte ich einige Sachen besser machen, aber dafür ist es halt da, dass ich halt im Endeffekt danach einmal richtig reflektieren kann. Wo sind meine Fehler? Wo kann ich mich bessern? Wo sind auch meine Stärken gewesen? Und ja.
18	[0:00:47] I: Hm inwiefern würden Sie sagen war das eben eine authentische Situation, also eine Situation wie man sie in der Schule

1/8

			vorfinden würde?
...Age appropriateness	19	[0:00:54] TP: Also ich finde, das war sehr authentisch meiner Meinung nach, weil ich halt solche Situationen damals auch erlebt habe. Ähm nicht halt direkt schon in dem frühen Alter, sage ich mal, so dass ich äh bisschen älter halt war mit Englisch. Äh und erst dann quasi meine Lehrerin mit mir Englisch gesprochen hat. Ich habe zum Beispiel oft den Fehler gemacht, dass ich direkt angefangen habe, Englisch zu sprechen, obwohl es ja eigentlich äh um eine Person ging, die Englisch nicht so gut versteht und nicht so gut sprechen äh und verstehen kann. Ähm bei mir jedoch war es halt so authentisch gewesen, dass ich halt damals auch immer auf Englisch versucht habe, zu antworten dem Lehrer, da der Lehrer auch mit mir versucht hat, Englisch zu sprechen bis zu einem gewissen Punkt, wo wir dann auch quasi gewechselt haben von Englisch auf Deutsch. So war das der Fall dann auch. Und auch natürlich, dass äh sie quasi diesen Switch haben wollte, also die andere Person ähm, hat das Ganze nochmal viel authentischer gemacht. Also es war jetzt nicht so irgendwie so als würde ich mich da in einer falschen Situation befinden. Also da wo ich mich wirklich wiedererkennen würde.	
...Actors	20	[0:01:56] I: Inwiefern hat sich die Schauspielerin, die die Schülerin gespielt hat, Ihrer Meinung nach auch authentisch verhalten?	
	21	[0:02:02] TP: Ähm wie schon gerade gesagt, also sie hat halt versucht, mir klar zu machen, dass äh sie nicht so gut Englisch sprechen kann. Darauf bin ich nicht halt so direkt eingegangen. Ähm jedoch erst im Nachhinein gemerkt, also dass sie gut ihre Rolle miteinbezieht. Also sie hat mir äh immer wieder klar gemacht, dass sie das und das nicht auf Englisch versteht. Und sprich ich das dann auf Deutsch auch nochmal erklären sollte, was für mich auch den Eindruck gemacht hat, dass sie äh die selben Erfahrungen auch schon gemacht hatte und daher auch dieses Authentische einfach da war und man sich auch/ Ich meine auch sie sich auch in so einer Situation wiedererkennen würde.	
...Communicative behavior (verbal)			
...Rather positive	22	[0:02:34] I: Hm ist sie Ihrer Meinung nach kontinuierlich in der Rolle	

2/8

..Rather positiv	23	der Schülerin geblieben?
		[0:02:39] TP: Äh ja definitiv.
..Rather positiv	24	[0:02:41] I: Und wirkte sie wie eine natürliche Gesprächspartnerin, also wie eine Neuntklässlerin?
		[0:02:45] TP: Ähm ja, im Großen und Ganzen schon. Ähm war halt ein bisschen schwer sich erstmal so in diese Lage reinzusetzen, dass jetzt nicht neben mir eine Person sitzt, die jetzt auf Neuntklässlerin tun muss, ähm aber ich meine so, es war ganz gut. Also ich war jetzt nicht so, dass ich gedacht habe ‚Okay, eigentlich müsste die das wissen.‘ Aber ähm sie hat das schon gut gemacht als Neuntklässlerin so zu schauspielern. Man hat das gut gemerkt.
..Feed Forward ..General: Dealing with em ..Feed Back	26	[0:03:12] I: Inwiefern wurden mit dem Rollenspiel Ihrer Meinung nach Kompetenzen erfragt, die für den Beruf als Lehrkraft notwendig sind?
		[0:03:19] TP: Ähm Kompetenzen wurden demnach erfragt, dass äh ich halt schauen musste, erstmal wie ich mit dem Schüler umgehen soll. Wie ich ihm etwas vermitteln soll, was er falsch gemacht hat, wie er etwas besser machen kann, ähm ob ich ihm dabei helfen kann oder ähm inwieweit auch. Es gab halt eine Situation, wo sie mich gefragt hatte, wie sie etwas lernen kann. Und dann habe ich ihr quasi ähm wie in einer Scaffolding-Methode gezeigt, dass ich ihr quasi eine Stütze würde in einer Aufgabe, sie aber nach und nach diese Stütze verliert und das dann quasi selber macht, dass sie weiß, wie es geht natürlich mit Hilfe, aber dann irgendwann mal selber auch anwenden kann. Und dabei auch bleibt.
..Use of language as medium ..Specific Issues of German L1	28	[0:04:00] I: Und welche Kompetenzen wurden in dem Rollenspiel erfragt, die besonders wichtig sind für Englischlehrkräfte?
		[0:04:06] TP: Für Englischlehrkräfte war es halt besonders wichtig, ähm so darauf einzugehen, dass ja halt jemand neben einem sitzt, der Englisch nicht wirklich gut besitzt, dass es halt eine

...Use of language as medium
...Specific issues of German L1

Fremdsprache ist, dass wir darauf auch als Lehrer dann quasi achten müssen, dass wir nicht zu sehr Englisch mit denen sprechen so wie ich es am Anfang getan hatte. Ähm und halt auch darauf achten müssen, dass halt Vieles, was im Deutschen geschrieben wird, äh nicht so eins zu eins ins Englische übersetzt werden kann, weil sonst würde Vieles keinen Sinn ergeben, äh da auch die Satzreihenfolge einfach anders sein soll. Auch Wörter ähm, die Vokabeln waren zum Beispiel halt so immer gesetzt, dass sie eben manchmal ähm quasi das so eins zu eins übersetzt hätte. Sprich äh bekommen hat sie als become übersetzt, was halt äh in dem Satz dann gar keinen Sinn mehr ergeben hat. Also es gibt halt ähm die Vokabeln, auf die man eingehen müsste, das Schreiben an sich äh, auch das Verstehen und das Sprechen.

30 [0:05:03] I: Hm unsere Idee ist es ja so ein bisschen so eine Simulation auch als Prüfung tatsächlich irgendwann mal einzusetzen. Und ähm meine Frage ist jetzt: Inwieweit können Sie sich das vorstellen, dass man das wirklich als Prüfung im Lehramtsstudium einsetzt?

...Personal development/ Oppor
...Practice-orientation

31 [0:05:16] TP: Ich könnte es mir sehr gut vorstellen eigentlich sogar. Also ähm vielleicht sogar als äh Vorher- und Nachhereffekt, dass man das einmal quasi vorher einmal macht bevor man ähm richtig, also sagen wir mal am Beginn des Studiums oder im dritten Semester jetzt einfach so beispielsweise und dann später nochmal im Bachelor oder im Master irgendwann mal wieder macht. Ähm einfach nur, um zu sehen, wo die Entwicklung stattgefunden hat, weil solche Situationen werden halt täglich auf einen zu kommen. Nicht wahrscheinlich so eins zu eins so wie ähm heute jetzt, sondern, dass einer halt direkt auf einen zukommt und fragen wird äh 'Wie war denn meine Note? Wie war meine Klausur?' Also sich schon darauf vorzubereiten in verschiedenen Situationen finde ich sehr wichtig. Und von daher auch sehr praktisch, wenn das in

4/8

Personal development/ Topic		einer Prüfung vorkommen sollte.
Practice-orientation		
	32	[0:06:01] I: Und speziell als Prüfung im Englischstudium?
	33	[0:06:05] TP: Ähm.
	34	[0:06:06] I: Können Sie sich das da vorstellen?
	35	[0:06:07] TP: Ja, auch halt genau die Gründe wie ich gerade eben auch gesagt habe. Es ist halt wichtig, dass ähm wir halt lernen zu sprechen, aber so zu sprechen, dass äh wir quasi auch Werte vermitteln den Schülern, auch im Englischen, dass das halt alles so auch richtig verstanden wird. Ähm und ja dann weiß man es auch.
Focus auf Sprache		
TP's appropriate use of language		
	36	[0:06:29] I: Hatten Sie den oder haben Sie das Gefühl, dass man in so einer Prüfung wirklich fair bewerten kann? Mit fair meinen wir hier eher, dass man wirklich das, was man kann, auch wirklich zeigen kann.
	37	[0:06:38] TP: Ähm ich sage mal so fair wird halt schwierig, weil halt jeder eine eigene Persönlichkeit hat für sich, wie man mit so einer Situation umgeht. Ich nehme mich mal jetzt als Beispiel. Ich war halt da relativ nervös gewesen. Bin auch nicht so der richtige Sprecher in einer Gruppe. Bin halt quasi auf dem Weg erstmal dahin. Deswegen äh hätte ich auch halt diese Idee halt als Vorschlag, dass man halt so ein Vorher- und Nachherprüfung macht, um halt zu schauen, wie die Entwicklung von einem ist. Und je nach dem kann man dann halt auch etwas fairer, sage ich mal, das Ganze beantworten. Und ja.
Dependent on the individual		
	38	[0:07:10] I: Und würden Sie sagen, dass man das durchaus objektiv bewerten kann? Also, dass man das auch vergleichen kann, die Leistung von verschiedenen Studierenden?
	39	[0:07:18] TP: Ähm ja schwierig sage ich mal. Halt auch, wie gesagt, gerade weil es halt so verschiedene Persönlichkeiten gibt, muss man halt schauen ähm. Ich weiß nicht, wie oft man sowas vielleicht durchführen kann, ob es wirklich nur einmal notwendig oder zwei- oder dreimal, um halt wirklich eine Objektivität herzustellen.
	40	[0:07:37] I: Hm also würden Sie sowas mehrfach machen, um dann über mehrere Prüfungen hinweg das gut beurteilen zu können.

- 41 [0:07:43] TP: Gerade halt auch wegen der Entwicklung, die jemand jetzt quasi von dem Beginn des Studiums bis zur Mitte, bis zum Ende halt macht. Also so wäre es in meinen Augen halt auch etwas sinnvoller gerade mit Blick auf die Objektivität.
- 42 [0:07:54] I: Und hat so eine Prüfung für Sie Praxisrelevanz?
- 43 [0:07:56] TP: Ja. Ahm die Begründung ist halt äh in der Praxis werden wir auch auf solche Situationen stoßen. Und wenn wir das schon vorher auch gemacht haben, können wir halt auch sowas viel besser einschätzen und auch ein besseres Feedback rausgeben.
- 44 [0:08:14] I: Hm Sie haben ja im Studium auch schon andere Prüfungsformate kennengelernt, Klausuren und mündliche Prüfungen und Hausarbeiten. Wie authentisch war denn jetzt das Rollenspiel im Vergleich zu diesen Prüfungsformaten?
- 45 [0:08:27] TP: Ähm ich sage mal zu den anderen Prüfungsformaten gehört es ja immer so man muss halt lernen. Man muss lernen und dann weiß man etwas. Man merkt sich das und schreibt es auf einen Zettel. Das jetzt gerade war halt wirklich aus einer Situation heraus äh, die man nicht richtig lernen kann. Entweder man es halt drauf oder man hat es nicht drauf und bewegt sich auf den Weg dorthin. Ähm und ich meine das macht halt schon so einen großen Unterschied aus, von der normalen Prüfung oder nur zu einer Klausur, wo man halt lernt, weil ich kann jetzt für sowas gar nicht lernen, weil im Grunde habe ich ja nicht Vieles, was ich lernen kann. Ich kann halt nur an mir selber lernen. So etwas äh authentischer rüber zu bringen statt äh irgendwie etwas auswendig zu lernen wie in einer ganz normalen Prüfung oder einem Test.
- 46 [0:09:12] I: Hm und wie authentisch war den so ein Rollenspiel im Vergleich zu einem Unterrichtsbesuch, den man vielleicht schonmal aus Sicht eines Schülers kennengelernt hat?

- 47 [0:09:19] TP: Ähm ja es war halt anders. Ich/
Meinen Sie jetzt den Unterrichtsbesuch, wenn ich
jetzt jemanden oder/
- 48 [0:09:28] I: Ja, angenommen/ Also wie authentisch ist so ein
Rollenspiel, wenn man das als Prüfung hat, im Vergleich zu einem
Unterrichtsbesuch als Prüfungsformat? Ich habe das jetzt so gesagt,
weil Sie haben ja wahrscheinlich noch keinen Unterrichtsbesuch
gehabt als/
- 49 [0:09:39] TP: Ich habe halt vieles gehört von sowas. Und äh es gab
halt sehr viel verschiedene Meinungen. Viele haben halt gesagt, sie
saßen äh einfach nur hinten in der Ecke, haben sich ein paar Notizen
für das Portfolio gemacht und das war es auch. Äh wurden nicht
quasi von der Lehrkraft richtig eingebunden. Sprich man hat keine
richtige Situation wie es /
- 50 [0:09:57] I: Also es ist jetzt im Praktikum? Also ich meinte mit
Unterrichtsbesuch so, wenn man dann wirklich als Referendar später
geprüft wird im Unterricht. Das ist da ja eine Prüfung. Wie
authentisch ist diese Prüfung im Vergleich zum Rollenspiel?
- 51 [0:10:09] TP: Das ist auch wieder ein ganz anderes
Thema, weil die das immer ganz verschieden erlebt
haben. Es gab äh Situationen, wo man in einem
Unterrichtsbesuch, wo wir vorher quasi äh
manipuliert worden sind, sage ich mal so in
Anführungszeichen, wo es dann halt immer äh hieß,
dass wir etwas bekommen würden. Viele haben halt
auch schon quasi den Unterrichtsablauf im
Vorhinein geplant, so wie es eigentlich sein soll,
aber nicht halt äh mit unserem Wissen. Wir haben
das dann quasi in dem Unterrichtsbesuch erst
gemerkt, dass wir das schonmal gemacht haben.
Dementsprechend auch ähm einen besseren
Unterrichtsablauf äh leisten konnten und das Ganze
für die Lehrkraft halt einfacher gemacht haben.
Also authentisch würde ich dann eher, glaube ich,
solche äh Rollenspiele finden als so einen
Unterrichtsbesuch.

- 52 [0:10:55] I: Hm okay. Haben Sie noch Fragen, Kommentare oder Anmerkungen?
- 53 [0:10:59] TP: Ähm im Großen und Ganzen eigentlich nicht mehr. Ähm es war halt so die Situation, ob wirklich immer Englisch oder Deutsch in Frage kommt. Und ähm was halt in so einer Situation vielleicht wichtiger wäre. Also was man beachten sollte.
- 54 [0:11:16] I: Okay. Dann beende ich jetzt die Aufnahme.

G.4. Expert Interview – Part 1

1		Übersicht Zitate Experteninterview – Teil 1: Videos
2		Qualitative Kommentare zu den Videos
3		Exp 5
4		Exp 5
Deficit-orientation Learners ability to use the i Dialogue Final message (Recency) Feed forward Text deep-level features		Video 43 [00:12:51] E: Well, you know (to assign something without any anchors?) is very, very hard without a reference point but given what the dialogue was and given what she talked about, the balance, I would assign a C plus. [00:13:26] E: She virtually now, with few exceptions, she never listened to the student at all. It was a straight monologue with very few exceptions and given my knowledge of feedback, it is about, not even, the quality of what is given but the nature of what is heard and understood and there was very little checking to see whether the student understood. And when she asked for instance about false friends and the student asked, how can I practice? The answer was right learning which I do not think helped the student, (when I heard her?), at all. When I was the student, I would have walked out feeling a little bit (demand?) by the whole exercise because other than a few platitudes, she corrected me on everything but there was no, where-to-next. And given my whole theory of feedback is feedback is primarily a function of where am I going. I do not think that student walked out with a going-next. She kind of offered/ The student kind of offered do I have to do it again? No, you do not. So I think the student walked out outside like: well, that is over. And there was very little learning. And of course what I want to do now, Thomas, is interview the student about what she understood and what she will be doing next. But it was a very traditional, old-fashioned teachers (ran out?) it was kind of here is a little bit of praise, overall it was pretty good. BUT, it did not play about the most important thing you have got the tenses role, the title was not short and concise, you had too much of your own opinion in there, the conclusion was not clear, the translation was all over the place. So I think that, in terms of understanding about where the student is going next I do not think

Text deep-level features		she really knew, she certainly understood about how she was going and that is why I gave her the C or C plus because not very well but okay and how she go; it was very hard to know, there was no feedback to the student about where she is such relative to the criteria the instructor wanted, like you are well on the way to mastering this, you are a long way to mastering this, there was no sense of that at all. So it was okay, I would give the student a C plus, I give the teacher a C minus.
Specificity Text surface-level features Text deep-level features		Video 24 [00:27:38] E: C plus. [00:27:40] I: C plus? [00:27:42] E: He was much more at a higher level and reached more information and grammar. There was a touch more dialogue and checking in with the student. He had far less specific content than the first one but I think most of that specific content just went in and out of the students ears of the first one. And he gave also, so where to next, in terms of looking at the back and how you use it particularly about the irregular verbs (unv.) much more specific about what she should do to improve where she went to next. He was clear about the ending was too short but the beginning was pretty good in the way she had good arguments moving into the ending. So, more specific, a much better balance between a higher levels and lower level of information.
		Video 23 [00:40:17] I: Okay, so, how about that one? How would you evaluate this teachers/ [00:40:25] E: A D. [00:40:26] I: A D? Okay.

2/4

Deficit-orientation
Emotional-motivational level

	[00:40:27] E: It was straight to the details. The poor student, I am sure, was overwhelmed with everything was a detail, it was about low level content. Of course it is important but there was no concept, it was no depth of understanding, he focused all on tense. When she asked one question: how do I learn? He just said learn it by heart. And then contradicted the whole thing in the end by saying it was not too bad. Every single thing was a negative and about detail. So, if I was the student, I would go away thinking that all I have to do is learn everything by heart, read learn, memorise, no concept of tense, no concept of irregular verbs. I would be demoralised after that. - [00:56:08] E: You are right. The problem I am having is that the relative spread and obviously I would like to ?????. But why did not I fail him? He did give a lot of information, very detailed, far to detailed, he lost the student very quickly. If we were sitting around moderating, I would say I am more likely go from a D to an F, I go from an F to a D.
Video 21	[00:52:48] I: Okay, so this was the last one. For the last time, how would you evaluate/ [00:52:56] E: B minus. [00:52:57] I: B minus. Okay, and why? [00:53:01] E: She started off by building or trying to build some credibility with the student, she was the only one who used the first name.

3/4

Mentioning positive aspects (first)	She started off with some comments about what the student had done right, the catchy headlines, what she had done right on the tenses and she gave examples where the student got it right, where the student got it wrong. She did not talk mainly/ She had a good balance between giving information about principles, subject, verb, action, and details. There was a part there I thought she went a bit off target, where I could not understand, or as a student did not understand what she was saying but then she came back about getting more principled. I did not understand why she insisted on the dictionary of google but I am not here to comment on that. I also liked the way she handled, that she gave the student an action that she had to follow-up, that she had to do something about it. The second one that we watched, he was much more specific than this one but this one, I think, got a slightly better balance. I as a student understand what the feedback was about and could actually use it.
Feed forward	
Learners ability to use the information	

G.5. Expert Interview – Part 2

	Code	Exp 4
General comments		
	Overall evaluation	[2:33:28] E: Okay, I think you are tapping into competency in a novel way for ITE students. I have never seen anything like this. So, I think that is novel. I think it is ambitious and I am concerned about the practicality of it
	Comprehensiveness	[2:12:02] E: I think you are being quite exhaustive in terms of coverage. My concern will be pragmatics. (...) [2:12:13] I: Absolutely. – 2:12:15] E: If you are thinking ahead to potential use by professional (unv.) teachers, you know, pragmatism is a really important thing. When, you know, John Hattie, John led the assessment tools for teaching and learning on project (unv.), I led the team that led the writing rubrics, that was scoring writing. And we had quite / we revised rubrics and wanted teachers to learn from and in utilizing them they were going to learn a lot about content knowledge about writing and about which works, but they proved problematic in terms of acquiring that knowledge and the time to employ them so that they have moved gradually from rubrics for different purposes for writing because what is in the various dimensions is different according to purpose to a much more generic rubric that is simplified. I do not / Well, I think if you train expert raters to utilize this it is possible, I think this is going to be quite complex (...) and I do not know that it would then translate into something rather professional even an experienced professional could really use. [2:13:46] I: Yeah. [2:13:48] E: This (unv.), but for an instrument as you are thinking, trying to validate something that could be used in ITE to assist students by lecturers who actually may or may not know a great deal about it (lachen), it is going to be quite hard. So, I suspect when you see if, you know, your six trained raters can do it with some degree of reliability then / because you and I are not reliable, we did not rely on this, we did not use this, we just used a / [2:14:28] I: A global scale, right? [2:14:29] E: The patterns that I have in my head that I know to be efficacious for improving writing, this sort of feedback that is known to be efficacious is what I have in my head. And so the interaction sort of known to remote this they might be different to what you have in your head too. Once you have them here, we probably would come to your

		agreement to make it this way. Hopefully (lachen). That is a good thing to test out (lachen). [2:15:06] I: Yeah, that is right. I am also very concerned about pragmatics but I find it really difficult to erase more than one or two criteria here, so / [2:15:23] E: Yeah, I think it is hard. [2:15:25] I: Yeah, so when it comes to the pragmatic use. So, it is not necessarily meant for teachers when they are at school, it is really meant for teacher-students, so to assess them, not / sadly the instrument, the purpose is not necessarily a formative one regarding on how can I improve, this is not the main focus, so, and the people who would grade the performance would be fellow researchers, so not necessarily teachers, so that would be further professors that teach at university-level, they would be the ones that have to assess that, so yeah. But pragmatics is a big issue because I totally see that for the raters it would be very difficult to do that in just one run through and not stopping it, rewinding and watching it a couple of times. This will be interesting. — [2:16:39] E: You might have to think of organizational ways that you could get around that.
	Scoring	[2:33:28] E: Okay, I think you are tapping into competency in a novel way for ITE students. I have never seen anything like this. So, I think that is novel. I think it is ambitious and I am concerned about the practicality of it, even for ITE professional lecturers, people teaching at the university applying this. So, it may be that for research purposes you establish that this is like, you know, like they do with many tests establish that this a valid and somewhat reliable instrument and then you get a shortened version that you also validate later that might actually be used in practice, I do not know, I just think it would be something novel. I mean our lecturers who visit in schools will have practical, observational schedules that they will complete but nothing as complex as this, so I think that few would have the experience, even researching something like this / So yeah, maybe there are organizational ways that you can facilitate even for you research in applying a scoring system. — [2:35:10] E: Maybe one group of people will look at the language competences and other groups look at this on the same videos or a whole group, so you get reliability, but no one looks at the WHOLE thing in one go.
Self-evaluation		[1:10:02] E: I am also looking at the left-hand column, here you have the points allocated and those vary from one to three. Three for self-evaluation and only one for feedback. (lachen) That seems very

		odd to me. Is that based on what is on the next slide or / - [1:12:45] E: You probably read the article I published in 2010 where we would make a task for teachers. They were just asked to give a general comment, then three specific, or two specifics, I cannot remember, and then we scored them in terms of these sort of things but with the emphasis on do they tell, how they are going against the learning objectives or against the expected performance or whatever it be. For me it is a lot on the self-evaluation in a time when you are actually focusing on the teacher's ability to give / the training teacher's ability to give feedback compared to only one for feedback and two / I do not care about the positive, I think (unv., undeutlich und lachend) for always trying to / people that always tell kids they are doing well, but mentioning specifically what you have done well is good. - [1:15:19] E: Interesting, because you see, speaking about your scoring for self-evaluation, only video two actually asked their students how they felt about it, but I have / but going on and on with that, that did not actually build on, or probe why the student thought that. Which, yeah, anyway, so this might be why he has scored more on something like that. [1:15:49] I: Absolutely. (...) I am just curious, you do not believe that asking the learner for their own evaluation is something that is favorable? [1:16:01] E: No, no. It is fine as long as you (...) do not just say are you happy with the text? Yeah, I am happy with the text. And from that there is nowhere to go (lachen). [1:16:17] I: Okay. [1:16:19] E: What are you happy with or where do you think you could improve is different. I read about a wonderful article about Charlie who was five and said he came up to us teachers (unv. undeutlich) and he could identify which ones he has done correctly and which ones he could not get, you know that is what you are trying to (unv.) here. Yeah, I probably would not / one of my (doc?) students did work on this, worked on a sense of what did you think about your text? What did you think about my feedback to your text? This was with students, 16 to 17 year olds, and so was looking at both a sub-evaluation and their response to her feedback. But these students are not seeing the feedback, they are just receiving it like in a conference, oral confrontation.
Positive feedback		[1:12:45] E: You probably read the article I published in 2010 where we would make a task for teachers. They were just asked to give a general comment, then three specific, or two specifics, I cannot remember, and then we scored them in terms of these sort of things but with the emphasis on do they

		tell, how they are going against the learning objectives or against the expected performance or whatever it be. For me it is a lot on the self-evaluation in a time when you are actually focusing on the teacher's ability to give / the training teacher's ability to give feedback compared to only one for feedback and two / I do not care about the positive, I think (unv., undeutlich und lachend) for always trying to / people that always tell kids they are doing well, but mentioning specifically what you have done well is good. — [1:14:37] E: I think (...) our teachers operate on / teachers of younger students operate on the notion that you give a star and then two things that you can actually improve type thing that is the sort of ratio, something positive you have done really well and maybe some wishes. A star and two wishes, I think you call it something like that (lachen).
Learning objectives (Feed Up)		[1:11:10] E: I would have given, for the state the learning objectives one, i.e. knowing what they are, because they should have known them already and then they give feedback against them as being the BATTLE thing you know rather than / you know, that just seems very odd. — [1:11:41] E: If you are going to give them different points, I think from my perspective these are not weighted they I would weight them. I would not put so much emphasis on just stating what the learning objectives are, maybe revisiting briefly but the IMPORTANT thing is, the learner receives feedback about performance AGAINST the learning objectives, because you are stating the learning objective and then you only give them hint for learning objective. There are several objectives, we need to know how we are going against them. For me, the important thing, it is often the feedback is not related to what the aims were, what the learning objectives were, but you got to give three for feed forward. (...) Anyway.
Feedback (Feed Back)		[1:10:02] E: I am also looking at the left-hand column, here you have the points allocated and those vary from one to three. Three for self-evaluation and only one for feedback. (lachen) That seems very odd to me. Is that based on what is on the next slide or / — [1:11:10] E: I would have given, for the state the learning objectives one, i.e. knowing what they are, because they should have known them already and then they give feedback against them as being the BATTLE thing you know rather than / you know, that just seems very odd. — [1:11:41] E: If you are going to give them different points, I think from my perspective these are not weighted they I would weight them. I would not put

		so much emphasis on just stating what the learning objectives are, maybe revisiting briefly but the IMPORTANT thing is, the learner receives feedback about performance AGAINST the learning objectives, because you are stating the learning objective and then you only give them hint for learning objective. There are several objectives, we need to know how we are going against them. For me, the important thing, it is often the feedback is not related to what the aims were, what the learning objectives were, but you got to give three for feed forward. (...) Anyway.
Next Steps (Feed Forward)		[1:11:41] E: If you are going to give them different points, I think from my perspective these are not weighted they I would weight them. I would not put so much emphasis on just stating what the learning objectives are, maybe revisiting briefly but the IMPORTANT thing is, the learner receives feedback about performance AGAINST the learning objectives, because you are stating the learning objective and then you only give them hint for learning objective. There are several objectives, we need to know how we are going against them. For me, the important thing, it is often the feedback is not related to what the aims were, what the learning objectives were, but you got to give three for feed forward. (...) Anyway.
Task		
Process		
Self-regulation		[1:17:46] E: I wonder why self-regulation is on the lowly inferential because I think you have to make a lot of inference to actually (...) be able to say something about whether the student is employing self-evaluation strategies. And I presume that is / does the initial teacher education candidate asked the sort of questions that tap into the meta-cognitive use of strategies. I do not know that I have / well I will have to watch them again closely after having seen these criteria. I am not sure that I saw anyone do the sort of thing I would expect around self-regulation. And I think it does require quite high levels of inference to work out whether they are doing that or not. Anyway, other than telling them, you know, other than referring to the textbook and telling them to go and look at that, in other words putting the (unv.) on them or look at the Oxford Dictionary, I do not know, but I will have to look at them more closely again. I just think that is going to be a hard one. Did you find it hard when you scored?
Self		[1:19:47] E: Yeah, where he asked the student about structure, she says she think so, she tried and what. It is / I think that was problematic and many things are more, not just self-regulation, but more affect, that is meta-cognitive, where it is as you said, you have got nothing, and this is one the self about affect, a criterion related to bolstering self-efficacy, you know a setup as a writer. I am not sure what self does not address the learner's personal level would actually relate to, making him

		feel at EASE Making the / the fourth one which was one that kept saying it takes courage to / takes courage to hand this in. Well, I cannot remember. [1:20:52] I: This is related to Hattie's four dimensions of feedback. And he says that like a simple praise, like you are a good student, this would fall under that category, so if someone said that in the simulation, he would not get that point. However, this did not occur in the videos and this is something that I believe / [1:21:13] E: Oh, this is something negative. You get a point if you do not say it (lachen). Okay, sorry, I misinterpreted that entirely. So it does not address the learner's / there is no generic feedback about, on a / [1:21:30] I: Yes, no generic praise. [1:21:36] E: Although, some say, in the videos that it was a good text. [1:21:40] I: Yeah, this is a grey area right. It is somehow related to the task because the task was the text but it is still generic, so. [1:21:51] E: It has to not related to the text. A good text is alright but all good effort which someone / effort like in video four, he made a good effort, that is a personal thing is it not? [1:22:05] I: That is true. [1:22:09] E: Or no, I do not know (lachen). To me that is relatively problematic (...). So, what is missing is, yes maybe, something that has to do with the effect of aspects of both self-evaluating, of getting feedback, of the self of the writer or even, I mean, some students may find this really engaging and others sort of was not, I do not know.
Dialog Orientation		[1:24:59] E: I mean it is pointless involving the learner with a question if it is not a high-level cognitive question or if the prompt is just sort of you know, eliciting something that is mundane. (...) We found that our teachers who score highly on promoting self-regulation ask many more high-level cognitive demanding questions as opposed to questions. So, I mean I think I would not give anyone three that just asked questions or prompted that was not demanding the student to engage at a high level but as you say, it is very difficult because your students, your actors were possibly / I do not know what instructions, whether they were / well, even if the teachers had asked questions at a high-level (reality?) initial teacher education candidates then they deserve to score well. But if they just asked questions or prompt, they do not deserve necessarily to score. (...) The next one, this is sort of almost the same thing. You involve the learners in your questions or prompts / [1:26:35] I: Yeah, they are based on their frequency. [1:26:41] E: (unv., vermischt mit Aussage vom I) Yeah, so it is not (...) I cannot understand how the

		quality relates to the notions of questions and prompts. [1:27:01] I: They are vertically designed. So, they are three different categories. [1:27:12] E: Oh yeah, got that. So, you have just got a frequency thing there it has nothing to do with the actual nature of the question or prompt. My critique is that frequency is not a good indicator of how effective a question or prompt is at involving a student's cognitive work. [1:27:32] I: So, it would be preferable to say involve the learner through highly cognitively demanding questions, three points? [1:27:40] E: Yeah, yeah. [1:27:43] I: Then two points rather cognitively demanding, one point at least asks them some sort of question and not just talks to them. [1:27:50] E: At least involves them, yeah. I mean, it is that OR the converse of that is / one of my early doctoral students found that teachers who allowed the students to INITIATE actually were promoting for example self-regulation and also promoting a much higher-quality conference, in this case they were analyzing conferences. So, it is also opening the opportunity for the student to initiate and some of them did (not?) invite questions. So, that does not cover it here at all, you are just saying, involving the learner through questions or prompts. The prompts and the questions should be valuable, but it is also the sign that they can be just opening up the possibility of an initiation by the student. (...) That would also be highly inferential (lachen). And so, I was trying to think that there was some opportunity, several of these where they actually asked if they had questions, so I am not sure whether that qualifies as something valuable, but it did give an opportunity for the student to initiate a direction that the interaction would take.
Quality of Feed Back		1:29:32] E: Okay, good. Okay, feedback is to be accurate, clear / I do not know how in video two you ever thought his feedback was accurate or clear (lachen) because I did not (lachen). But that could be just the way that your teachers might express things. Accurate or clear, often, so again it is a frequency measure? [1:30:00] I: Yes. [1:30:03] E: Right, okay. So, all of the / (...) Why is the frequency even identifying this requires a high level of inference? So, if I go through the transcript and highlight all of the things that I think relate to quality of feedback that that is highly inferential as opposed to being inferential / all the time / sometimes. [1:30:43] I: So, the idea is that when watching a simulation, you do not have the chance to rewatch it. So, you only get the chance to see it once and

		you do not necessarily have a video tape if it was an actual assessment at university. So, then of course you have to have / So, if it is not an either-or-category, so I can see someone gives feed forward, check, three points. If I have to infer some sort of evaluation, so if it is frequency, then it is necessarily that you have to say, okay, I think this was more in the category of often rather than sometimes, right, and therefore there is more inference in that regard.
Quality of Feed Forward		-
Securing understanding / content adaptation		[1:32:50] E: So, by giving alternative answers means, is the sort of thing that I saw in video three doing, which I thought was actually bad practice. Constantly giving what I would / what is it based around so / [1:33:08] I: No, that is not what I tried to / No, it is giving an answer to the learner and then seeing that the learner did not really understand it, so for example, give an explanation of things that were wrong, asking did you understand what I mean? And the learner saying no I did not understand and then giving another / [1:33:23] E: And then giving alternative explanations or something like that. So, not answer. Answer sounds like giving you the right answer. Okay, now I understand, that is, again it is often, sometimes and never ensures himself of understanding and does adapt to their needs by rephrasing or giving alternative possibilities or whatever. Okay.
Linguistic adaptivity		[1:38:50] E: Okay, so across your total corpus of videos, do you have / I thought each one of these students asked the learner / each one of these ITE students, initial teacher education, asked these students, school students whether they wanted to continue in German or English, so they get a range across your transcripts. - [1:51:32] E: Adaptivity, yeah, we have talked about that one for the alternative (unv.) and this one I think would not be SO difficult to do because it is mainly at the start of the conversation, and then, well if they start in one language like in English you will have to see was there a request of change and did they respond to that? - [2:21:50] E: But they do emphasize accuracy and appropriacy rather than to the detriment of rhetorical situation and the actual melding of the structural context to (unv., schnell und undeutlich) as well as the language. Anyway, that is just my sense, I am not sure what, going through this, what you could DUMP. Maybe simplify some of the ones like the

		language in which the interaction is conducted. Maybe those categories could be simplified, I do not know, and the one that we talked about that could be a dichotomy rather than the error pattern.
Adaptive error correction		[1:42:29] E: I would not like to be doing this on the fly as it were, in other words in real-time. I would not like to be (in account?) to make all of these decisions at the time I am viewing the video. Do your raters have the chance to stop the video, to rewind, relook? — [1:44:43] E: In each of the three, the first three, you have to identify an opportunity for self-correction or the first two, anyway, and in the third one you have to find the implicit and indirect feedback before the explicit and (unv.) and only explicit. So. — [1:45:10] I: I am thinking for example at the third video that person usually / [1:45:15] E: Yeah, that was almost the last category preferably but the others, you know, I suppose if you train your raters and you train them on examples of this from the videos, then it might be possible. We had, when teachers were giving feedback to one another we trained them using videos and identifying the instances and where it would fall on the scoring sheet. I would have to have the (unv., lachen). This is where I suddenly, I mean the previous one I think is relatively straightforward compared to this because there are / you would have to think of, you know, more than one instance, I find, because if I see an instance of providing opportunity for self-correction through implicit, followed by that do I tick that, because I have seen one instance of it, or the others are frequency type measures, you know, how often do I see it, and this one is not, this one is phrased like, did I see it or not? [1:46:35] I: Yes. [1:46:36] E: Yeah, so if I saw except the last thing has got only in it, so it becomes a frequency measure, okay, so I am not sure did I see, or am I going to think overall that I see some opportunities for the first one. Opportunities to self-correct, to let them have a go and then I tell them what the answer is. And the last one is I hinge around a bit and then I say what it is (lachen). It is quite hard. — [1:48:40] E: Mhm, these last two have no self-correction in them, and they are only / they are relatively straightforward in a sense there is only this / I can recognize there is only explicit and direct AND some indirect and implicit. But yeah, for some people rating will be hard to differentiate implicit and indirect from waffly, sort of generalized feedback, you know so that would be a hard one, but those two (unv., schnell und undeutlich), it is this

		one where you have to look at the / first of all they give opportunity for self-correction, then they / and they do it through implicit and indirect or they just give opportunity through some other way, but opportunity for self-correction. — [1:50:20] E: The second category there is / gives them an opportunity, and when they do not get it right, I just go straight to the direct but this one you are saying the opportunity is THROUGH the implicit and indirect and not just by invitation. So, the second one could be, you know, you see any problem with this particular reader's perspective, with this particular construction? And then I say NO, you see this (lachen)? On the first one I have got to do it much more indirectly and I am not sure what that would look like and if I see it. So, maybe those, the important aspect is, is there an opportunity for self-correction?
Prioritization of content/ structure		[1:53:47] E: Yeah, something about the features of the writing purpose or the genre.
Appropriate amount of information		[1:55:25] E: Most teachers would know the students well, so they would know how much would be appropriate to get some enhanced quality but, in this case, they should be making this decision based on quality of the text that they are seeing and, you know, their students (weighting?) for every single mistake in that thing is totally over the top. If the school student had to remember all of that, they would (unv.) in hell. Not in that situation. [1:56:03] I: Absolutely. [1:56:06] E: It is quite hard and again, your right, depending on the length is, yeah / if you think of sitting a / this is like a conference in a classroom and you only have five minutes to do this? — [1:57:28] E: Yeah, I mean/ I know when one of my doctor students and her doctor, when she was setting, when the students were setting goals for the next piece, given her feedback was a dialogic journal, and she had / I think she had three things, they were going to focus on. So, you know the problem with some of these feedback conversations is they go over and over the spelling mistakes and all of the irregular verbs and they identify all of them and so on, whereas that is just an aspect. So, an appropriate amount of information might be three different aspects of the, whatever. All of the spelling ones, even though they have gone each one, they become an aspect / or all of the conventions could become an aspect and something to do with the headline or structure becomes an aspect or topic.

Language Correctness		[2:01:34] E: The one and two seem to be almost something, you could subsume them, language accuracy and language appropriacy, inappropriate use of tense, forming of the simple past, well / at some point or some sort of grammatical, anyway.
Language Adequacy		[2:01:34] E: The one and two seem to be almost something, you could subsume them, language accuracy and language appropriacy, inappropriate use of tense, forming of the simple past, well / at some point or some sort of grammatical, anyway. — [2:02:17] E: I have not heard of genre conventions. I would put we in the genre conventions. Yeah, anyway. [2:02:26] I: Yeah. This is appropriate to the genre maybe I could make it more clear, yeah. I will just talk about the next because I have seen we have been talking longer than I have agreed, longer than I have asked you to, so /
Content		
Structure		[2:00:00] E: Is there any evidence of that one, when you are giving that three points? I might think that the notion of / the features of text normally associate with a specific purpose or certain genre. This could be perhaps THE most important. So those will be things like the headline and the structure and the useful things like WE are appropriate which some people might put into / oh, I do not know I would put it there. I do not know, I just think you are giving a lot of emphasis to process when in these videos I do not think I saw any discussion of the process. — [2:17:03] E: No, you could leave out / There is actually (...) / There is no category that deals with the reader and the rhetorical purpose of the text which is one that we / If you are thinking about hierarchy, we start with the rhetorical purpose and then we might move to the content and structure, and then we might move to language resource use and then we might move to the mechanics of punctuation, spelling, grammar. But there is nothing there that is dealing with the / in fact I do not even think in the task the audience, it is a newspaper report that I was not quite sure who it was for. I do not know if it is stated, anyway. [2:18:07] I: I think it says for the local newspaper. [2:18:10] E: Sorry? [2:18:11] I: It is for the local newspaper. [2:18:13] E: Local newspaper, okay. So, but there is no assessment of the extent to / well one or two people talked about / [2:18:24] I: They talked about the reader but basically only in the notion of their English-speaking

		readers, so you should have the text in the English language. [2:18:34] E: That was about all that I saw. So, there is no / it is a really important / because of-ten students who have not got all of the mechanics and the structure, right, sometimes are actually making the point for the reader in doing it really well, you know, the reader takes the message. [2:18:57] I: So, maybe one could add that as another aspect of the text. So, besides accuracy, appropriacy, content, structure, one could add (...) the purpose or the reader, focusing on the reader as another criterion. [2:19:17] E: I think I would address the audience for this and then extend to if all of this, is that actually appropriate for the audience, not just accurate and appropriate use of tense but what it / the way it is structured, so the headline / those who talked, I mean our teachers and classes will immediately talk about hooking the reader into your piece and you would do that with a headline. And some of them talk about something that was interesting and grabbing attention and so on. And that would complete the (unv.). [2:19:52] I: So maybe the aspect of structure should be split into more overt structure and then into more genre-related aspects. [2:20:02] E: Yeah, you have got content structure there and I suppose the content is that they are talking about the important information that the reader needed, in structure as accessibility to the reader or highlighting the important things for the reader or whatever. But yeah, it is actually quite difficult. And these are not so hard to actually identify in a conversation. [2:20:31] I: Yeah. But it is a good thing to keep in mind. However, that would add another criterion and we just talked about pragmatics. [2:20:40] E: But it is more important to me than some of these other things because, so you have got accuracy and appropriacy there but it is accurate, appropriate in terms of the audience, because we use different language, we go on different language resources for different audiences and so on, and within different genre. It is hard to know how to, on the criteria, how to categorize them.
Process		[2:00:00] E: Is there any evidence of that one, when you are giving that three points? I might think that the notion of / the features of text normally associate with a specific purpose or certain genre. This could be perhaps THE most important. So those will be things like the headline and the structure and the useful things like WE are appropriate which some people might put into / oh, I do not know I would put it there. I do not know, I just think you are giving a lot of emphasis to process when in these videos I do not think I saw any discussion of the process.

Process orientation		[2:04:08] E: Did you find it relatively easy to apply? I think that would be okay.
Error pattern		[2:04:45] E: Why would they get three points if they do that exclusively? Which is really poor-quality feedback if you do that exclusively (lachen). [2:05:07] E: No, no. You could say PREDOMINANTLY (...) addresses error patterns, which I do not think is good. I mean, I would not give them three points for that (lachen). This is the problem with your frequency scale. You are implying that it is good to do this a lot. [2:06:20] E: Not all of the videos / In not all of the videos where they identified an error pattern did they / were they able to articulate what it was or show how you might address it accurately. You know, it is not just simply identifying, maybe the frequency is to do with showing how to / the frequencies around addressing accurately or explaining how to address the error pattern accurately. Always does this, sometimes does this, so you are not asking for the number of times they address error patterns but given they do, do they do it accurately, do they identify it accurately and address it accurately because yeah. I know that is not entirely what / it could be problematic because they could miss some very important error patterns and he would not capture that under that sort of phrasing, but I do think it is better than this. [2:07:50] E: Yeah, and it is not really highly inferential either because there are, obviously, accurate ways of mediating and ways that are not so accurate, but they are also ways of explaining that are accessible and ways of explaining that are confusing. — [2:21:50] E: But they do emphasize accuracy and appropriacy rather than to the detriment of rhetorical situation and the actual melding of the structural context to (unv., schnell und undeutlich) as well as the language. Anyway, that is just my sense, I am not sure what, going through this, what you could DUMP. Maybe simplify some of the ones like the language in which the interaction is conducted. Maybe those categories could be simplified, I do not know, and the one that we talked about that could be a dichotomy rather than the error pattern. [2:22:41] E: The target and the pattern, the target-oriented (pattern?) that goes into one.
Target-oriented error selection		[2:08:52] E: Relates rather to the learning objectives and genre because as you used genre, the writing function purpose because, yeah. [2:09:02] I: Yes, that would be the learning objectives. Writing a news report and then being able to follow the demands that come with that, so yeah. Okay, so then frequency would be okay with this one, right? So, then sometimes.

		[2:09:22] E: Yeah. [2:09:25] I: So, if I understand correctly get rid of the error pattern category and just have the target-oriented error selection. [2:09:32] E: Yeah, except I would put select error patterns and always relate them to the learning objectives, so it is not often selects but that they often relate them to the learning objectives. Just change where your descriptor is and then relates how often they relate to the learning objectives rather than how often they select them. [2:09:56] I: Ah, I see. Yeah (...). [2:10:03] E: Because in the bottom they never select error patterns and never relate them to the learning objectives, in other words they never relate error patterns and learning objectives. [2:10:13] I: Yeah, okay. [2:10:14] E: They might select them but they never relate them which is the important thing. — [2:22:41] E: The target and the pattern, the target-oriented (pattern?) that goes into one.
De-tailedness		
Linguistic accuracy		[2:24:32] E: I think that that is the first time I got used to that, that is why I was so slow on the first one, I was busy trying to DECODE the sort of almost haunting English where they, I mean hardly oral English, because we do not necessarily use full sentences and so on, but it was quite difficult. So, looking as opposed to listening to them, looking at the translated text makes you particularly / makes them vulnerable to receiving only one and zero. — [2:25:56] E: Yeah, that is a hard one. I think where I see and listening here is where I noticed in video two when I said to you, he does not have the meta-language to talk about this. So, there is nothing in there. You know, we work hard at trying to get our teachers to have a (unv.) mind, I do not know if you know a work from (exodus?) published quite a lot on this, so that you have a meta-language to talk about how language works and our teachers, so that they can actually use the terminology correctly. [2:28:00] E: You are combining two scales here. You have got a pass scale and a frequency scale. Always and often are frequency these ones are (unv.), so you cannot do anything with that scale. [2:28:16] I: Okay. [2:28:19] E: Digging back to / I trained the / my PHD students of psychology are trained psychologists but going back to my (unv.) and stats training, you have to stick to one measure, so it is either a

Appendix G - Coding Samples

		(unv.) measure, well not often, but this one is (...) noticeable and grave which are not frequency measures. So, what I am interested in. [2:29:30] E: Or maybe language facilitates learning, in other words it does not hinder learning in any way, language has some minor impact on learning or here it is some minor impact / you have got three that are not necessarily hindering anyone that hinders.
Pedagogical accuracy		[2:30:34] E: So, you have a lot of focus on learning strategies, yet I do not actually / in the four videos I saw very little comment on.

H. Statistical Information

H.1. Intercoding Sub-Codes PSETs Guideline Interview

Table 81: *Intercoding sub-codes for the PSETs guideline interview.*

Code	Agreement	Disagreement	Total	%
Actors_Behavior_Rather authentic	11	0	11	100.00
Actors_Continuity_Missing	1	0	1	100.00
Actors_Continuity_Rather positiv	10	0	10	100.00
Actors_Natural Conversational Partner_Missing	3	0	3	100.00
Actors_Natural Conversational Partner_Rather positive	8	0	8	100.00
Fairness_Partially	1	0	1	100.00
Fairness_Rather negative	1	0	1	100.00
Fairness_Rather positive	5	1	6	83.33
Fairness_Unclear	1	0	1	100.00
Situation_Authenticity_Rather authentic	8	1	9	88.89
Situation_Authenticity_Rather inauthentic	1	0	1	100.00
Situation_Authenticity_Unclear	1	1	2	50.00
Situation_From School_Rather negative	1	0	1	100.00
Situation_From School_Rather positive	10	0	10	100.00
Summative assessment_General Teacher Education_Positive	9	0	9	100.00
Summative assessment_General Teacher Education_Negative	2	0	2	100.00
Summative assessment_EFL Teacher Education_Rather positive	10	0	10	100.00
Summative assessment_EFL Teacher Education_Missing	1	0	1	100.00
Tested Competences_EFL Competences_Assessment	3	1	4	75.00
Tested Competences_EFL Competences_Feedback	2	1	3	66.67
Tested Competences_EFL Competences_Other	1	0	1	100.00
Tested Competences_EFL Competences_Subject-specific Competences	6	1	7	85.71
Tested Competences_EFL Competences_T-S Interaction	5	0	5	100.00
Tested Competences_General Competences_Assessment	2	1	3	66.67
Tested Competences_General Competences_Feedback	6	2	8	75.00
Tested Competences_General Competences_Other	0	1	1	0.00

Code	Agreement	Disagreement	Total	%
Tested Competences_General Competences_Professional Knowledge	1	0	1	100.00
Tested Competences_General Competences_Situation-related	2	1	3	66.67
Tested Competences_General Competences_T-S Interaction	7	2	9	77.78
Total	119	13	132	90.15

H.2. Discrepancies in Inter-Rating I

Table 82: Discrepancies between rater 1 and rater 2 in rating I.

	AEM23	CCF27	KMP05	MMM07	MMK21	CHS10	GJS22	EME03	EEH01	MJO28
Self-evaluation										
Positive feedback	R1=0; R2=2									
Learning objectives (Feed Up)							R1=2; R2=0			
Process			R1=0; R2=2		R1=0; R2=2			R1=2; R2=0		R1=0; R2=2
Self-regulation		R1=0; R2=2				R1=0; R2=2		R1=2; R2=0	R1=2; R2=0	
Use of material*							R1=0; R2=2			
Prioritization of deep-level features						R1=2; R2=0				
Error patterns			R1=0; R2=2		R1=0; R2=2	R1=0; R2=2				
Surface: Language correctness						R1=2; R2=0				
Surface: Language appropriateness	R1=2; R2=0					R1=2; R2=0				
Deep: Structure	R1=2; R2=0					R1=2; R2=0				
Deep: Content	R1=2; R2=0					R1=2; R2=0				
Deep: Process*									R1=0; R2=2	R1=0; R2=2
Self	R1=2; R2=0									
Body language			R1=0; R2=2		R1=2; R2=0					
Adaptation of language*	/	/	/	/	/	/	/	/	/	/
Revision	/	/	R1=0; R2=2	/	/	R1=1; R2=2	/	R1=2; R2=1	R1=1; R2=2	R1=1; R2=2
Adaptive error correction	/	/	/	/	/	/	/	/	/	/

	AEM23	CCF27	KMP05	MMM07	MMK21	CHS10	GJS22	EME03	EEH01	MJO28
Condition: Dialogue orientation		X			X	X	X			
Dialogue orientation	/	R1=1; R2=0 (0)	/	R1=2; R2=1	R1=1; R2=0 (0)	R1=1; R2=0 (0)	R1=2; R2=0 (0)	/	/	/
Condition: Quality of feedback						X	X			
Quality of Feed Back	/	/	/	R1=1; R2=2	/	R1=1; R2=0 (0)	R1=2; R2=0 (2)	/	R1=1; R2=2	R1=2; R2=1
Quality of Feed Forward	/	/	/	/	R1=0; R2=1	R1=1; R2=0	/	R1=2; R2=1	/	/
Checking understanding	R1=1; R2=0	/	/	/	/	/	R1=2; R2=1	/	R1=1; R2=2	/
Amount and balance of information	R1=1; R2=0	R1=1; R2=0	R1=0; R2=1	/	R1=2; R2=1	/	/	/	R1=2; R2=1	/
Target-oriented error selection	R1=1; R2=0	R1=1; R2=0	R1=1; R2=2	/	R1=2; R2=1	/	/	/	/	R1=2; R2=1
Detailedness	/	/	R1=2; R2=1	/	/	R1=1; R2=0	/	R1=2; R2=1	/	R1=2; R2=1
Language accuracy	/	/	R1=1; R2=2	R1=1; R2=2	/	R1=0; R2=1	/	/	/	/
Professional accuracy	/	/	R1=2; R2=1	/	R1=2; R2=1	R1=0; R2=1	/	/	/	/

Table 83: Discrepancies between rater 1 and rater 3 in rating 1.

	AEM23	CCF27	KMP05	MMM07	MMK21	CHS10	GJS22	EME03	EEH01	MJO28
Self-evaluation								R1=2; R3=0		
Positive feedback	R1=0; R3=2									
Learning objectives (Feed Up)						R1=2; R3=0	R1=2; R3=0			
Process	R1=2; R3=0						R1=0; R3=2			
Self-regulation							R1=2; R3=0	R1=2; R3=0	R1=2; R3=0	
Use of material*							R1=0; R3=2			
Prioritization of deep-level features						R1=2; R3=0				
Error patterns										
Surface: Language correctness										
Surface: Language appropriateness										
Deep: Structure						R1=2; R3=0				

Appendix H - Statistical Information

	AEM23	CCF27	KMP05	MMM07	MMK21	CHS10	GJS22	EME03	EEH01	MJO28
Deep: Content						R1=2; R3=0				
Deep: Process*										R1=0; R3=2
Self	R1=2; R3=0									
Body language										
Adaptation of language*										
Revision			R1=0; R3=1					R1=2; R3=1		
Adaptive error correction										
Condition: Dialogue orientation		X			X	X				
Dialogue orientation		R1=1; R3=0 (1)			R1=1; R3=0 (1)	R1=1; R3=0 (1)	R1=2; R3=1			
Condition: Quality of Feed Back						X	X			X
Quality of Feed Back			R1=1; R3=2	R1=1; R3=2		R1=1; R3=0 (1)	R1=2; R3=0 (2)		R1=1; R3=2	R1=2; R3=0 (2)
Quality of Feed Forward		R1=0; R3=1			R1=0; R3=1			R1=2; R3=1	R1=2; R3=1	R1=2; R3=1
Checking understanding	R1=1; R3=0						R1=2; R3=1			
Amount and balance of information	R1=1; R3=0									R1=2; R3=0
Target-oriented error selection	R1=1; R3=0	R1=1; R3=2								R1=2; R3=1
Detailedness			R1=2; R3=1						R1=2; R3=1	R1=2; R3=1
Language accuracy		R1=1; R3=2	R1=1; R3=2	R1=1; R3=2						
Professional accuracy			R1=2; R2=1	R1=1; R3=2	R1=2; R3=1					

H.3. Inter-Item Correlation Matrix for 27 Items

Table 84: Inter-Item Correlation matrix of the 27 items.

	Language_adaptivity	Self_Evaluation	Positive_Feedback	Learning_objectives	Process	Self_Regulation	Use_of_material	Prioritization_of_deep_level_features	Error_patterns
Language_adaptivity		-0.19	-0.29	-0.08	0.01	0.00	-0.16	0.21	-0.05
Self_Evaluation	-0.19		0.12	0.12	0.33	0.31	0.10	-0.07	-0.01
Positive_Feedback	-0.29	0.12		0.38	-0.03	0.09	-0.10	0.07	0.22
Learning_objectives	-0.08	0.12	0.38		0.07	0.09	-0.10	0.47	0.33
Process	0.01	0.33	-0.03	0.07		0.29	0.16	0.17	-0.04
Self_Regulation	0.00	0.31	0.09	0.09	0.29		-0.25	-0.03	0.25
Use_of_material	-0.16	0.10	-0.10	-0.10	0.16	-0.25		-0.01	-0.13
Prioritization_of_deep_level_features	0.21	-0.07	0.07	0.47	0.17	-0.03	-0.01		0.01
Error_patterns	-0.05	-0.01	0.22	0.33	-0.04	0.25	-0.13	0.01	
Surface_Correctness	0.06	-0.15	0.15	-0.10	-0.05	0.16	-0.04	-0.01	0.09
Surface_Appropriateness	0.16	0.12	0.09	0.09	-0.15	0.20	-0.05	-0.10	0.20
Deep_Structure	0.11	-0.05	0.28	0.50	0.08	-0.08	-0.12	0.37	0.12
Deep_Content	0.00	0.08	-0.08	0.33	0.22	0.13	0.19	0.19	0.19
Deep_Process	-0.09	0.00	0.00	0.00	0.43	0.28	0.07	-0.18	0.07
Self	0.08	-0.04	0.04	0.04	0.22	0.23	-0.06	0.28	-0.03
Body_language	-0.16	0.18	0.13	0.13	0.16	0.17	0.24	-0.02	0.16
Process_orientation	-0.14	-0.14	0.07	0.07	0.22	-0.01	-0.18	0.23	-0.13
Adaptive_error_correction	-0.08	0.36	0.12	0.00	0.52	0.38	0.17	0.15	-0.02
Dialogue_orientation	-0.28	0.39	0.22	-0.06	0.42	0.18	0.04	0.01	-0.10
Quality_feed_back	-0.22	0.12	0.34	0.80	-0.06	0.04	-0.08	0.25	0.42
Quality_feed_forward	0.07	0.25	0.21	0.36	0.18	0.40	-0.04	0.24	0.17
Checking_understanding	-0.03	0.16	0.21	0.21	0.30	0.08	-0.09	0.15	0.08
Balanced_information	-0.11	0.28	0.26	0.26	0.19	0.15	-0.11	0.15	0.17
Target_oriented_error_selection	-0.18	0.27	0.37	0.50	-0.01	0.03	-0.06	0.23	0.29
Detailedness	0.07	0.11	0.34	0.34	0.07	0.26	-0.09	0.06	0.39
Language_accuracy	-0.36	0.15	0.21	0.14	0.14	0.01	-0.06	-0.06	0.40
Professional_accuracy	-0.08	0.21	0.27	0.34	0.36	0.28	-0.06	-0.11	0.26

	Surface_Correctness	Surface_Appropriateness	Deep_Structure	Deep_Content	Deep_Process	Self	Body_language	Process_orientation	Adaptive_error_correction
Language_adaptivity	0.06	0.16	0.11	0.00	-0.09	0.08	-0.16	-0.14	-0.08
Self_Evaluation	-0.15	0.12	-0.05	0.08	0.00	-0.04	0.18	-0.14	0.36
Positive_Feedback	0.15	0.09	0.28	-0.08	0.00	0.04	0.13	0.07	0.12
Learning_objectives	-0.10	0.09	0.50	0.33	0.00	0.04	0.13	0.07	0.00
Process	-0.05	-0.15	0.08	0.22	0.43	0.22	0.16	0.22	0.52
Self_Regulation	0.16	0.20	-0.08	0.13	0.28	0.23	0.17	-0.01	0.38
Use_of_material	-0.04	-0.05	-0.12	0.19	0.07	-0.06	0.24	-0.18	0.17
Prioritization_of_deep_level_features	-0.01	-0.10	0.37	0.19	-0.18	0.28	-0.02	0.23	0.15
Error_patterns	0.09	0.20	0.12	0.19	0.07	-0.03	0.16	-0.13	-0.02
Surface_Correctness		0.38	-0.12	0.19	0.07	0.32	0.24	-0.05	0.17
Surface_Appropriateness	0.38		0.04	-0.11	-0.20	-0.07	0.17	-0.12	0.01
Deep_Structure	-0.12	0.04		0.00	0.04	0.00	0.07	0.12	-0.10
Deep_Content	0.19	-0.11	0.00		0.14	0.07	0.34	-0.11	0.17
Deep_Process	0.07	-0.20	0.04	0.14		0.10	0.12	0.13	0.44
Self	0.32	-0.07	0.00	0.07	0.10		-0.11	0.02	0.25
Body_language	0.24	0.17	0.07	0.34	0.12	-0.11		0.15	0.24
Process_orientation	-0.05	-0.12	0.12	-0.11	0.13	0.02	0.15		0.05
Adaptive_error_correction	0.17	0.01	-0.10	0.17	0.44	0.25	0.24	0.05	
Dialogue_orientation	0.16	0.06	-0.12	0.03	0.30	0.15	0.22	0.23	0.72
Quality_feed_back	0.06	0.18	0.35	0.27	-0.09	0.08	0.19	0.02	0.01
Quality_feed_forward	0.27	0.27	0.20	0.08	0.11	0.28	0.11	0.16	0.31
Checking_understanding	0.21	0.13	-0.21	0.04	-0.04	0.08	0.01	0.08	0.37
Balanced_information	0.00	-0.01	0.23	0.17	0.15	-0.24	0.26	0.26	0.12
Target_oriented_error_selection	-0.06	-0.02	0.21	0.22	-0.03	-0.18	0.21	0.09	-0.03
Detailedness	0.37	0.64	0.25	-0.07	-0.15	0.20	0.11	-0.17	0.06
Language_accuracy	0.09	0.11	0.05	0.11	-0.04	-0.19	0.15	0.20	-0.05
Professional_accuracy	0.08	0.10	0.11	0.18	0.23	0.02	0.23	0.19	0.15

	Dialogue_ orientation	Quality_ feed_back	Quality_ feed_forward	Checking_ understanding	Balanced_ information	Target_oriented_ error_selection	Detailedness	Language_ accuracy	Professional_ accuracy
Language_adaptivity	-0.28	-0.22	0.07	-0.03	-0.11	-0.18	0.07	-0.36	-0.08
Self_Evaluation	0.39	0.12	0.25	0.16	0.28	0.27	0.11	0.15	0.21
Positive_Feedback	0.22	0.34	0.21	0.21	0.26	0.37	0.34	0.21	0.27
Learning_objectives	-0.06	0.80	0.36	0.21	0.26	0.50	0.34	0.14	0.34
Process	0.42	-0.06	0.18	0.30	0.19	-0.01	0.07	0.14	0.36
Self_Regulation	0.18	0.04	0.40	0.08	0.15	0.03	0.26	0.01	0.28
Use_of_material	0.04	-0.08	-0.04	-0.09	-0.11	-0.06	-0.09	-0.06	-0.06
Prioritization_of_deep_level_features	0.01	0.25	0.24	0.15	0.15	0.23	0.06	-0.06	-0.11
Error_patterns	-0.10	0.42	0.17	0.08	0.17	0.29	0.39	0.40	0.26
Surface_Correctness	0.16	0.06	0.27	0.21	0.00	-0.06	0.37	0.09	0.08
Surface_Appropriateness	0.06	0.18	0.27	0.13	-0.01	-0.02	0.64	0.11	0.10
Deep_Structure	-0.12	0.35	0.20	-0.21	0.23	0.21	0.25	0.05	0.11
Deep_Content	0.03	0.27	0.08	0.04	0.17	0.22	-0.07	0.11	0.18
Deep_Process	0.30	-0.09	0.11	-0.04	0.15	-0.03	-0.15	-0.04	0.23
Self	0.15	0.08	0.28	0.08	-0.24	-0.18	0.20	-0.19	0.02
Body_language	0.22	0.19	0.11	0.01	0.26	0.21	0.11	0.15	0.23
Process_orientation	0.23	0.02	0.16	0.08	0.26	0.09	-0.17	0.20	0.19
Adaptive_error_correction	0.72	0.01	0.31	0.37	0.12	-0.03	0.06	-0.05	0.15
Dialogue_orientation		0.08	0.21	0.40	0.00	-0.11	0.00	0.07	0.16
Quality_feed_back	0.08		0.36	0.25	0.12	0.41	0.42	0.29	0.32
Quality_feed_forward	0.21	0.36		0.35	0.27	0.26	0.45	0.02	0.14
Checking_understanding	0.40	0.25	0.35		0.18	0.20	0.19	0.24	0.23
Balanced_information	0.00	0.12	0.27	0.18		0.79	-0.01	0.38	0.16
Target_oriented_error_selection	-0.11	0.41	0.26	0.20	0.79		0.10	0.47	0.17
Detailedness	0.00	0.42	0.45	0.19	-0.01	0.10		0.19	0.24
Language_accuracy	0.07	0.29	0.02	0.24	0.38	0.47	0.19		0.37
Professional_accuracy	0.16	0.32	0.14	0.23	0.16	0.17	0.24	0.37	

H.4. Test Score Frequencies

Table 85: *RobS test score frequencies.*

RobS Score	Frequency	Percent	Cumulative Percent
10	1	2.0	2.0
15	1	2.0	3.9
17	2	3.9	7.8
18	1	2.0	9.8
19	1	2.0	11.8
21	2	3.9	15.7
22	1	2.0	17.6
24	1	2.0	19.6
25	2	3.9	23.5
26	1	2.0	25.5
27	2	3.9	29.4
28	1	2.0	31.4
29	1	2.0	33.3
30	1	2.0	35.3
31	2	3.9	39.2
32	4	7.8	47.1
33	2	3.9	51.0
34	5	9.8	60.8
35	5	9.8	70.6
36	3	5.9	76.5
37	2	3.9	80.4
38	3	5.9	86.3
39	1	2.0	88.2
40	2	3.9	92.2
41	1	2.0	94.1
42	1	2.0	96.1
45	1	2.0	98.0
47	1	2.0	100.0
Total	51	100.0	

H.5. Normal Distribution in the Actor-specific Groups

Table 86: Statistics on the normal distribution of RobS test scores in the actor-specific subsamples.

	<i>n</i>	Skewness			Kurtosis			Shapiro-Wilk-Test	
		<i>g</i>	<i>SE</i>	<i>z</i>	<i>k</i>	<i>SE</i>	<i>z</i>	<i>W</i>	<i>p</i>
Actor 1	9	0.13	0.72	0.18	0.69	1.40	0.49	0.98	.983
Actor 2	22	-0.43	0.49	.88	1.20	0.95	1.26	0.96	.508
Actor 3	20	-0.66	0.51	1.29	0.55	0.99	0.56	0.96	.473

H.6. Further Statistical Information on SPK, PCK and CK /LP

Table 87: Statistics on the normal distribution of the external tests of participants in their Bachelor program.

	<i>n</i>	Skewness			Kurtosis			Shapiro-Wilk-Test	
		<i>g</i>	<i>SE</i>	<i>z</i>	<i>k</i>	<i>SE</i>	<i>z</i>	<i>W</i>	<i>p</i>
SPK	31	0.01	0.42	0.02	-0.94	0.82	1.15	0.95	.18
PCK	31	-0.04	0.42	0.10	-0.63	0.82	0.77	0.98	.71
CK / LP	31	0.07	0.42	0.17	-0.25	0.82	0.30	0.98	.85

Table 88: Statistics on the normal distribution of the external tests of participants in their Master of Education program.

	<i>n</i>	Skewness			Kurtosis			Shapiro-Wilk-Test	
		<i>g</i>	<i>SE</i>	<i>z</i>	<i>k</i>	<i>SE</i>	<i>z</i>	<i>W</i>	<i>p</i>
SPK	20	-0.36	0.51	0.71	-0.64	0.99	0.64	0.95	.33
PCK	20	1.47	0.51	2.88	2.75	0.99	2.77	0.87	.01
PCK (without outlier)	19	0.60	0.52	1.15	-0.56	1.01	0.55	0.92	.09
CK / LP	20	0.03	0.51	0.06	-0.82	0.99	0.82	0.98	.91

Table 89: Statistics on the normal distribution of the external tests.

	<i>n</i>	Skewness			Kurtosis			Shapiro-Wilk-Test	
		<i>g</i>	<i>SE</i>	<i>z</i>	<i>k</i>	<i>SE</i>	<i>z</i>	<i>W</i>	<i>p</i>
SPK	51	-0.27	0.33	0.82	-0.81	0.66	1.23	0.99	.059
PCK	51	0.84	0.33	2.55	2.16	0.66	3.27	0.96	.040
PCK (without outlier)	50	0.07	0.34	0.21	-0.30	0.66	0.45	0.98	.582
CK / LP	51	-0.05	0.33	0.15	0.02	0.66	0.30	0.95	.948

Table 90: Descriptive statistics of the three external tests of participants in the Bachelor program.

	<i>n</i>	<i>M</i>	<i>SE</i>	<i>SD</i>	95 % CI		Median	<i>Min</i>	<i>Max</i>
					lower	upper			
SPK	31	22.35	1.00	5.54	20.32	24.39	22	12	31
PCK	31	5.45	0.51	2.84	4.41	6.49	6	0	11
CK / LP	31	78.79	1.96	10.90	74.79	82.79	77.50	55.00	100.00

Table 91: Descriptive statistics of the three external tests of participants in the Master of Education program.

	<i>n</i>	<i>M</i>	<i>SE</i>	<i>SD</i>	95 % CI		Median	<i>Min</i>	<i>Max</i>
					lower	upper			
SPK	20	26.05	0.95	4.24	24.07	28.03	26	17	32
PCK	20	7.60	0.79	3.51	5.95	9.25	7	4	18
PCK (without outlier)	19	7.05	0.59	2.59	5.80	8.30	7	4	12
CK / LP	20	80.56	1.56	6.97	77.30	83.83	80	67.50	92.50

Table 92: Statistics on the normal distribution of the external tests of participants in their Bachelor (BA) and Master of Education (MEd) program.

	<i>n</i> (BA)	Shapiro-Wilk-Test		<i>n</i> (MEd)	Shapiro-Wilk-Test	
		<i>W</i>	<i>p</i>		<i>W</i>	<i>p</i>
SPK	31	0.95	.176	20	0.95	.333
PCK	31	0.98	.711	20	0.87	.010
PCK (without outlier)				19	0.92	.091
CK / LP	31	0.98	.851	20	0.98	.908

Table 93: Bivariate correlations for the external tests and graduation grade.

		Graduation grade	
		Pearson's <i>r</i>	Spearman's ρ
PK	<i>r</i>	-.31	-.31
	Sig. (2-tailed)	.026	.026
PCK	<i>r</i>	-.35	-.31
	Sig. (2-tailed)	.012	.025
CK / LP	<i>r</i>	-.15	-.11
	Sig. (2-tailed)	.295	.450

I. Digital Supplement

The digital supplement includes the SPSS syntax which is used for the statistical analyses of the data in the main survey.

Traditional assessment formats in university-based EFL teacher education programs usually focus on cognitive dispositions rather than on the actual performance of pre-service EFL teachers in everyday teaching situations. This assessment gap is addressed in this thesis by developing and validating a role-play-based simulation (RobS) designed for the summative assessment of pre-service EFL teachers' feedback competence on writing.

Drawing on theories from multiple disciplines, such as higher education, medical education, teacher education, educational psychology, and EFL-specific didactics, the RobS is developed as a performance-oriented assessment format. In the RobS, pre-service EFL teachers engage in a feedback conversation with trained actors who portray a standardized learner.

Following an argument-based approach to validation, the extent to which the RobS can be considered valid is investigated. Data from multiple studies, focusing on aspects such as authenticity, fairness, reliability, and external validity, are presented. The discussion in the validity argument indicates that the RobS can elicit and assess the performative facet of feedback competence on writing with sufficient confidence.

This work contributes a novel, empirically supported assessment framework to teacher education research. Moreover, it presents an approach to address the assessment gap, enabling pre-service EFL teachers to *show how* they provide feedback, rather than just *tell* their lecturers about it.