UNIVERSITÄT PADERBORN

Strukturanalysen des physikdidaktischen Wissens mithilfe von Machine Learning

Dissertation zur Erlangung eines Doktorgrades (Dr. rer. nat.)
der Fakultät für Naturwissenschaften
an der Universität Paderborn

vorgelegt von

Jannis Zeller

aus
Aachen

April 2025

Gutachter:

Prof. Dr. Josef Riese (Erstgutachten)

Prof. Dr. Oliver Tepner (Zweitgutachten)

Datum der mündlichen Prüfung: 03.04.2025

Danksagung

„It's over. It's done."

– Frodo Baggins in *The Lord of the Rings: The Return of the King* (Movie)

Meine Promotion war eine tolle Zeit und ich möchte sie hier (vielleicht bis auf einige Scharmützel mit bürokratischen Hürden) absolut nicht mit der Reise, die Frodo bis zu dieser Stelle in der Geschichte hat hinter sich bringen müssen, vergleichen. Trotzdem konnte ich mir diese Eröffnung hier nicht verkneifen.

Diese Arbeit wäre ohne die Unterstützung einer Vielzahl an Personen nicht möglich gewesen. Als erstes bedanken möchte ich mich bei meinem Doktorvater Prof. Josef Riese. Lieber Josef, danke, dass Du erst mir einen großen Datensatz gegeben und dann mich auf diesen losgelassen hast. Danke, dass Du mich stets unterstützt und mir die Möglichkeit gegeben hast, mich entsprechend meiner Interessen methodisch weiterzubilden. Danke, dass Du einerseits das Vertrauen in mich hattest, geplante Analysen erwartungsgemäß durchzuführen, aber trotzdem immer ein offenes Ohr hattest, wenn ich mich (häufig) ob meines Vorgehens rückversichern wollte. Danke vor allem auch, für die Auseinandersetzung mit den oben bereits angesprochenen „herausfordernden" bürokratischen Situationen, die für mich primär emotional belastend waren, für Dich aber auch echte Arbeit verursacht haben. Vielen Dank auch an Prof. Oliver Tepner, der sich bereit erklärt, das Zweitgutachten zu dieser Arbeit anzufertigen. Vielen Dank für Dein Interesse an meinem Projekt, den immer anregenden Austausch und die Arbeit, die Du in die Begutachtung und Teilnahme an meiner Disputation investiert hast.

Darüber hinaus möchte ich mich stellvertretend für alle Wissenschaftler:innen, die Teil der Projekte ProfiLe-P und ProfiLe-P+ waren, bei Dr. Yvonne Webersen bedanken. Danke, dass Du und Ihr die notwendigen Testinstrumente entwickelt und anschließend den Datensatz erhoben hast / habt, ohne den mein Projekt nicht möglich gewesen wäre. Nur so hatte ich den zeitlichen Freiraum, tief in methodische und technische Elemente einzutauchen. Weiherhin möchte ich mich bei der Gesellschaft für Didaktik der Chemie und Physik bedanken, bei der ich mich stets willkommen gefühlt habe und der ich die Vernetzung mit vielen anderen Forschenden zu verdanken habe. Ein großer Dank gilt auch der Studienstiftung des deutschen Volkes, die mich während der Promotion gefördert und mir somit maximale Flexibilität ermöglicht hat. Hierbei möchte ich mich besonders bei meinem Vertrauensdozent Prof. Dominik Groß für vertrauensvolle Ratschläge und gesellige Abende sowie bei der für mich zuständigen Referentin Dr. Anne-Sophie Käsbauer für die Unterstützung bei der Bewältigung bürokratischer Hürden bedanken. Danke auch an die Physikdidaktik-Arbeitsgruppen aus Aachen und Paderborn für spannende Gespräche, Projekteinblicke und Feedback. Danke insbesondere an die jeweiligen Mitarbeiterinnen des Sekretariats, die mich bei mir besonders unliebsamen verwaltungstechnischen Anliegen stets nach Kräften unterstützt haben. Danke vor allem an Doro, Rike und Melanie, die mich herzlich in der Physikdidaktik aufgenommen haben. Danke für viele lustige Abende auf Tagungen. Ob noch mehr solche Gelegenheiten folgen? Schauen wir mal, *dass* wird.

An meine Familie vielen Dank für mein Leben im Allgemeinen und die Unterstützung im Studium im Speziellen. Natürlich wäre ich ohne Euch und Euer Werk nicht an dem Punkt, an dem ich jetzt stehe. Ein besonderer Dank gilt meinem Opa der mir stets ein Beispiel an Geradlinigkeit, Zielstrebigkeit und Direktheit war. Danke, dass Du mich stets unterstützt und auch als kritischer Geist begleitet hast, mich aber trotzdem meine eigenen Entscheidungen hast treffen (und möglicherweise auch den ein oder anderen notwendigen Fehler hast machen) lassen.

Zuletzt vielen Dank an Klara dafür, dass Du auch in stressigeren Phasen meine – wenn man den Gerüchten glauben mag – verbissene Arbeitsweise ohne Klagen und mit höchstens minimalen Belehrungen ausgehalten hast. Danke, dass Du mir stattdessen eine Schulter zum Anlehnen und Runterkommen geboten hast.

Danke an Euch und Sie alle!

Eidesstattliche Erklärung

Name / Anschrift

Jannis Zeller
[redacted]

Erklärung zu meiner Dissertation mit dem Titel: „Strukturanalysen des physik-didaktischen Wissens mithilfe von Machine Learning“

Ich, Jannis Zeller, erkläre hiermit, dass ich die beigefügte Dissertation selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel genutzt habe. Alle wörtlich oder inhaltlich übernommenen Stellen habe ich als solche gekennzeichnet.

Ich versichere außerdem, dass ich die beigefügte Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und, dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.

Paderborn, 08.04.2025	J. Zeller
Ort, Datum	Unterschrift

Vorbemerkung

Zu dieser Arbeit existiert umfangreiches digitales Ergänzungsmaterial, welches die Datenbasis, den vollständigen Analysecode sämtlicher Analyseschritte sowie entwickelte Modelle enthält.

Diese kumulative Arbeit basiert auf den folgenden veröffentlichten Publikationen:

Artikel 1: Zeller, J., Schiering, D., Kulgemeyer, C., Neumann, K., Riese, J. & Sorge, S. (2024). Empirisch-kriterienorientierte Analyse des fachdidaktischen Wissens angehender Physiklehrkräfte. Welche inhaltlichen Strukturen zeigen sich über unterschiedliche Projekte hinweg? *Unterrichtswissenschaft*. https://doi.org/10.1007/s42010-024-00200-w

Artikel 2: Zeller, J. & Riese, J. (2025). Competency profiles of PCK using unsupervised learning: What implications for the structures of pPCK emerge from non-hierarchical analyses? *Journal of Research in Science Teaching.* https://doi.org/10.1002/tea.70001

Artikel 3: Zeller, J. & Riese, J. (2025). Machine-Learning-basierte Analyse von latenten Profilen des physikdidaktischen Wissens. *Zeitschrift für Didaktik der Naturwissenschaften, 31,* Artikel 5. https://doi.org/10.1007/s40573-025-00181-y

Die jeweiligen Beiträge aller beteiligten Autoren werden für jeden Artikel im Rahmen des Abschnitts „**Beiträge der Autoren**“ bzw. „**Author Contributions**“ am Ende des jeweiligen Artikels kenntlich gemacht.

Ein Teil bzw. Teile dieser Arbeit sind neben den Artikeln im Rahmen des kumulativen Formats bereits in den folgenden Tagungsbandbeiträgen enthalten:

Zeller, J. & Riese, J. (2023). Datenbasierte Fähigkeitsprofile im Physikdidaktischen Wissen. In H. van Vorst (Hrsg.), *Lernen, Lehren und Forschen in einer digital geprägten Welt, Tagungsband der GDCP Jahrestagung 2022* (S. 794–797). Gesellschaft für Didaktik der Chemie und Physik.

Zeller, J. & Riese, J. (2024). Fähigkeitsprofile im Physikdidaktischen Wissen mithilfe von Machine Learning. In H. van Vorst (Hrsg.), *Frühe naturwissenschaftliche Bildung, Tagungsband der GDCP Jahrestagung 2023* (S. 122–125). Gesellschaft für Didaktik der Chemie und Physik.

Zeller, J. & Riese, J. (im Druck). Assessment des physikdidaktischen Wissens mithilfe von Machine Learning. In H. van Vorst (Hrsg.), *Entdecken, lehren und forschen im Schülerlabor, Tagungsband der GDCP Jahrestagung 2024*. Gesellschaft für Didaktik der Chemie und Physik.

Kurzfassung

Das Fachdidaktische Wissen (FDW) zählt zu den zentralen Elementen des Professionswissens (angehender) Lehrkräfte und seine Relevanz ist sowohl theoretisch angenommen als auch empirisch belegt. In der fachdidaktischen Forschung liegt daher bereits seit längerem ein Fokus auf der Analyse des FDW, wobei mittlerweile vor allem Auswirkungen auf die Handlungsqualität und auf Lernergebnisse in den Blick genommen werden. Nach wie vor stellt aber auch die empirisch fundierte inhaltliche Beschreibung des FDW sowie der Transfer entwickelter FDW-Testverfahren auf Basis von Testinstrumenten mit offenem Antwortformat in die Ausbildungspraxis ein Forschungsdesiderat dar. In diesem Dissertationsprojekt werden daher auf Basis eines Datensatzes von 846 Bearbeitungen eines FDW-Testinstruments im Fach Physik (1) projektübergreifende FDW-Kompetenzniveaus auf Basis von Item-Response-Modellierungen exploriert, (2) nicht-hierarchische FDW-Kompetenzprofile auf Basis von (probabilistischen) Cluster- und Textanalysen beschrieben und (3) ein vollständig automatisiertes FDW-Assessment-System auf Basis von Machine Learning entwickelt. Dabei wurden insbesondere kognitive Anforderungskategorien als Subskalen des verwendeten Testinstruments betrachtet. Das Assessment-System wurde dabei auf Basis dieser und weiterer Subskalen sowie anhand der Zuordnung von Proband:innen zu den Kompetenzprofilen evaluiert und zeigte sowohl relativ zur Interrater-Übereinstimmung als auch absolut betrachtet hohe Performanzwerte.

Structural Analyses of Physics Pedagogical Content Knowledge using Machine Learning

Abstract

Pedagogical Content Knowledge (PCK) is one of the central elements of the professional knowledge of (prospective) teachers. Its relevance is theoretically established and empirically shown multiple times. PCK has therefore been analyzed continuously in education research, currently with a particular emphasis on its impact on the quality of teaching and directly on learning outcomes. However, there is still a lack of detailed empirically backed descriptions of the intricacies of PCK and of methodologies for translating developed PCK assessment procedures based on open-ended questionnaires into educational practice. In this dissertation project, therefore, three objectives are pursued, based on a dataset of 846 responses to a physics PCK test instrument. First, cross-project PCK competency levels are explored based on item response modeling. Second, non-hierarchical PCK competency profiles are described based on (probabilistic) cluster and text analyses. Third, a fully automated FDW assessment system based on Machine Learning is developed. In particular, cognitive requirement categories were considered as subscales of the test instrument used. The assessment system was evaluated based on these and other subscales, as well as the assignment of respondents to the competency profiles, and demonstrated high performance values both in relation to inter-rater agreement and in absolute terms.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AIC	Akaike Information Criterion
ANOVA	Analysis of Variances
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
SBERT	Sentence-BERT
SciEdBERT	Science-Education-BERT
BIC	Bayersian Information Criterion
BMBF	Bundesministerium für Bildung und Forschung
BoW	Bag of Words
CGT	Computational Grounded Theory
CI	Konfidenzintervall / Confidence Interval
CK	Content Knowledge
COACTIV	Verbundforschungsprojekt „Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung mathematischer Kompetenz“
CV	Cross-Validierung / Cross Validation
DEFT	Distributing epistemic functions and tasks
EAP	Expected-a-posteriori Schätzung
FDW	Fachdidaktisches Wissen
FF(s)	Forschungsfrage(n)
FW	Fachwissen
GMM	Gaussian Mixture Model
GPT	Generative Pretrained Transformer
GPU	Grafikkarte / Graphics Processing Unit
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HTML	HyperText Markup Language
IRT	Item-Response-Theorie
IQ	Intelligenzquotient
KeiLa	Verbundforschungsprojekt „Kompetenzentwicklung in mathematischen und naturwissenschaftlichen Lehramtsstudiengängen“
KI	Künstliche Intelligenz
KiL	Verbundforschungsprojekt „Messung professioneller Kompetenzen in mathematischen und naturwissenschaftlichen Lehramtsstudiengängen“
KMK	Kultusministerkonferenz
LDA	Latent Dirichlet Allocation
LLaMA	Large Language Model Meta AI

LL – UL	Unterer – Oberer Grenzwert / Lower – Upper Limit
LLM	Large Language Model
LM	Language Model
LPA	Latente Profilanalyse / Latent Profile Analysis
LR / LogReg	Logistische Regression
MW, M	Mittelwert / Mean *(in Tabellen und in „$M = x$, $SD = y$“-Datenangaben)*
MAP	Maximum-a-posteriori Schätzung
MC	Multiple Choice
MINT	Zusammenfassende Bezeichnung für die Wissenschaftsbereiche bzw. Schulfächer Mathematik, Informatik, Naturwissenschaften und Technik
ML	Machine Learning
MLE	Maximum Likelihood Schätzung
MoC	Model of Competence / Kontinuumsmodell
NLP	Natural Language Processing
NN	Neuronales Netzwerk / Neural Network
FCNN	Fully Connected NN
OECD	Organization for Economic Co-operation and Development
PCK	Pedagogical Content Knowledge
cPCK	collective PCK
ePCK	enacted PCK
pPCK	personal PCK
PISA	Program for International Student Assessment
PK	Pedagogical Knowledge
ProfiLe-P	Verbundforschungsprojekt „Professionswissen in der Lehramtsausbildung Physik“ (1. Projektphase)
ProfiLe-P+	2. Projektphase des ProfiLe-P-Projekts
ProwiN	Verbundforschungsprojekt „Professionswissen von Lehrkräften in den Naturwissenschaften“
PTR (Cycle)	Plan-Teach-Reflect (Cycle)
PW	Pädagogisches Wissen
RCM	Refined Consensus Model (of PCK)
SD	Standardabweichung / Standard Deviation
STM	Structural Topic Model
SVM	Support Vector Machine
TF-IDF	Term frequency – Inverse document frequency
TIMSS	Trends in International Mathematics and Science Study

1. Einleitung

Die Lehrkraft ist ein wesentlicher Einflussfaktor für den schulischen Erfolg von Schülerinnen und Schülern (z. B. Hattie, 2012). Bereits seit geraumer Zeit steht dementsprechend das Professionswissen von Lehrkräften im Fokus fachdidaktischer und bildungswissenschaftlicher Forschung (Baumert & Kunter, 2006; Riese, 2009; Shulman, 1986, 1987). Dabei wird angenommen, dass die Kompetenz von Lehrkräften im Rahmen einer Wirkungskette (indirekten) Einfluss auf den Unterrichtserfolg hat (Terhart, 2012). Orientiert an den Pionierarbeiten von insbesondere Shulman (1986) wird dabei das Professionswissen meist in die Bereiche Fachwissen (FW), Pädagogisches Wissen (PW) und Fachdidaktisches Wissen (FDW) unterteilt (mehr dazu in Abschnitt 2.1). Das FDW wird dabei als spezifisches Wissen von Lehrkräften verstanden, das notwendig ist, um konkretes Fachwissen konkreten Schülerinnen und Schülern zu vermitteln (Neumann et al., 2019; Shulman, 1987).

FDW spielt somit schon aus konzeptionellen bzw. theoretischen Gründen eine besondere Rolle. Auch empirische Ergebnisse belegen die Bedeutung von FDW für (a) die anderen beiden Professionswissensdomänen und deren Entwicklung (Hume et al., 2019; Riese et al., 2017; Sorge et al., 2019) sowie (b) den Unterrichtserfolg (Förtsch et al., 2016; Großmann & Krüger, 2022; Schröder et al., 2020). Zur Messung des FDW, um entsprechende Studien überhaupt zu ermöglichen, wurden bisher (im deutschsprachigen Raum) Leistungstests mit offenen und geschlossenen Aufgabenformaten eingesetzt (Gramzow, 2015; Kröger, 2019; Tepner et al., 2012). Dabei werden die Testaufgaben meist unter Nutzung von Aufgabenentwicklungs-modellen erstellt, die einerseits fachdidaktische Inhaltsbereiche oder „Facetten“ und andererseits kognitive Anforderungen oder Wissensarten beinhalten (ebd.). Solche Konzeptualisierungen umfassen zwar auch implizit Beschreibungen der angenommenen inhaltlichen (Fein-) Struktur des FDW, allerdings sind entsprechende Unterteilungen meist primär theoretisch motiviert und somit von eher normativem Charakter. Es bleibt bis auf eher technische Argumente wie statistische Item-Response-Modellvergleiche (Riese et al., 2017) unklar, inwieweit diese Konzeptualisierungen auch empirisch abgesichert werden können.

Empirische Untersuchungen von inhaltlichen Strukturen des FDW sind bislang im Bereich der Naturwissenschaften, genauer der Physik, primär mithilfe von hierarchischen Ansätzen auf Basis von Item-Response-Modellierungen durchgeführt worden (Schiering et al., 2023; Schiering et al., 2019). Die Ergebnisse dieser Untersuchungen sind inhaltliche Beschreibungen von FDW-Leistungsniveaus, die allerdings eng in den Kontext der jeweiligen Projekte eingebettet sind und sich direkt auf die Inhalte der Aufgaben der entsprechenden Testinstrumente beziehen (ebd.). Es ist also unklar, ob und inwiefern diese Niveau-beschreibungen zu projektunabhängigen Aussagen verallgemeinert werden können. Die genutzte Methodik (Mullis & Fishbein, 2020) lässt sich aber direkt oder in abgewandelter Form auch auf andere Datensätze übertragen, wodurch eine vergleichende Betrachtung der sich ergebenden Niveaubeschreibungen unterschiedlicher Projekte ermöglicht wird. Eine solche vergleichende Analyse ist das erste Kernziel dieses Projekts. Zur Erreichung dieses Ziels werden aufbauend auf Item-Response-Modellen von FDW-Score-Datensätzen aus zwei Projekten ($N_1 = 427$, $N_2 = 779$) Niveaumodelle des FDW mithilfe zweier unterschiedlicher Methoden erstellt und vergleichend analysiert (Kapitel 4).

Neben der erweiterten Betrachtung hierarchischer Strukturen des FDW durch einen projektübergreifenden Ansatz ist eine Ausweitung der Analyse der inneren Struktur des FDW mithilfe nicht-hierarchischer Analysen (z. B. MacQueen, 1967; McInnes et al., 2017; Spurk et al., 2020) wünschenswert. Eine solche Betrachtung wird zudem durch Ergebnisse der hierarchischen Analysen nahegelegt, in denen sich zeigt, dass mit hierarchischen Analysemethoden Unterschiede hinsichtlich interessanter Teilkompetenzen wie dem Evaluieren oder Kreieren im Kontext des FDW nicht abgebildet werden können (siehe Kapitel 4). Nicht-hierarchische Beschreibungen sind zudem nützlich, um in einem Assessment zu Feedbackzwecken auch nützliche empirisch fundierte Informationen zum Stand des FDWs liefern zu können, die über eine „bessere“ oder „schlechtere“ Gesamteinschätzung hinaus gehen. Eine solche nicht-hierarchische Untersuchung ist das zweite Kernziel dieses Projekts. Zur Erreichung dieses Ziels werden nicht-hierarchische Strukturen des FDW durch eine Cluster-Analyse von FDW-Score-Daten ($N = 846$) ermittelt und mithilfe einer explorativen Sprachanalyse der zugehörigen authentischen Sprachproduktionen der Proband:innen ausgeschärft (Kapitel 5). Aufbauend auf diesen Ergebnissen wird zudem eine Latente Profilanalyse (LPA, Spurk et al., 2020) durchgeführt, um die Ergebnisse stärker empirisch abzusichern (Kapitel 6). Die erhaltenen Strukturen werden auch als „Kompetenz*profile*“ bezeichnet, um sie von den hierarchischen Kompetenzniveaus aus dem ersten Zielpaket abzugrenzen.

Die Untersuchung der inneren Struktur des FDW ist neben (eher theoriebildenden) Forschungszwecken auch für die Erstellung von reichhaltigem inhaltlichem Feedback (Hattie & Timperley, 2007) nützlich bzw. notwendig. Für diesen Zweck ist es zudem naheliegend, die bereits existierenden Testinstrumente für ein Assessment zu nutzen. Die Auswertung der als besonders authentisch geltenden Aufgaben mit offenem Aufgabenformat (z. B. Krüger & Krell, 2020; Kulgemeyer et al., 2023) solcher Testinstrumente ist allerdings mit hohem manuellem Aufwand durch trainierte Kodierer:innen verbunden. Um ein solches Assessment auch skalierbar in die Ausbildungspraxis zu überführen und zudem eine effektive Nachnutzung der Testinstrumente für weitere Forschungs- und Monitoring-Zwecke zu ermöglichen, ist es also notwendig, die Auswertungsprozesse zu automatisieren. Eine Überführung der Aufgaben in geschlossene Antwortformate zu diesem Zweck ist allerdings hinsichtlich der Authentizität der entstehenden geschlossenen Aufgaben sowie der Übereinstimmung der durch die offenen bzw. geschlossenen Testinstrumente abgebildeten Konstrukte nicht unproblematisch (Kulgemeyer et al., 2023). Moderne Methoden aus dem Bereich des Machine Learning (ML) und Natural Language Processing (NLP) bieten aber alternativ die Möglichkeit bei Verfügbarkeit eines geeigneten Datensatzes statistische Modelle zu erstellen, die die Bepunktung offener Aufgaben automatisiert durchführen können (Zhai et al., 2020a; Zhai et al., 2020b; siehe auch Abschnitt 2.6 sowie Kapitel 6). Die Entwicklung und Evaluierung eines solchen Modells für ein konkretes FDW-Testinstrument ist das dritte Kernziel dieses Projekts. Zu diesem Zweck wird ein (vergleichsweise kleines) BERT[1]-Sprachmodell (Devlin et al., 2019) zur automatisierten Bepunktung des Testinstruments entwickelt und dessen Nutzbarkeit für ein informatives

[1] **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Assessment auf Basis von FDW-Subskalen und der Kompetenzprofile aus dem zweiten Zielpaket evaluiert (Kapitel 6). In diesem Rahmen werden auch weitere mögliche Modelle vorgestellt und evaluiert (Abschnitt 6.7.6 & 6.7.8).

Insgesamt wird in diesem Projekt das FDW von (angehenden) Physiklehrkräften einer empirisch-datenbasierten Detailanalyse zur inhaltlichen Beschreibung innerer Strukturen unterzogen. Dabei wird zudem ein automatisiertes Assessment-System für das FDW auf Basis eines etablierten Testinstruments (Gramzow, 2015) aus dem Projekt ProfiLe-P[2](+) (Vogelsang et al., 2019) entwickelt. Zunächst werden sowohl hierarchische FDW-„Kompetenzniveaus" mithilfe von Item-Response-Modellen (Kapitel 4) als auch nicht-hierarchische FDW-„Kompetenzprofile" mithilfe von Clustermodellen (Kapitel 5 & 6) genauer inhaltlich untersucht. Neben theoriebildenden Erkenntnissen aus diesen Analysen können insbesondere die Ergebnisse der nicht-hierarchischen Analyse genutzt werden, um ein Feedback reliabel und valide mit inhaltlichen Aussagen anzureichern. Im letzten Teil des Projekts wird dann ein BERT-Sprachmodell zur automatischen Bepunktung der offenen Aufgaben des verwendeten Testinstruments trainiert und die Performanz dieses Modells unter Rückgriff auf die zuvor gefundenen Kompetenzprofile sowie bestehende Subskalen u. Ä. evaluiert (Kapitel 6).

[2] Akronym ProfiLe-P: „**Prof**essionskompetenz **i**m **Le**hramtsstudium **P**hysik", gefördert durch das Bundesministerium für Bildung und Forschung. In der ersten Projektphase (ProfiLe-P, siehe z. B. Riese & Reinhold, 2012) wurde auf die Modellierung und Operationalisierung der Domänen des Professionswissens für das Fach Physik fokussiert. In der zweiten Projektphase (ProfiLe-P+ siehe z. B. Vogelsang et al., 2019) wurde die längsschnittliche Entwicklung sowie der Zusammenhang der Domänen des Professionswissens zur Performanz in prototypischen Handlungssituationen in den Blick genommen. Für die hier vorgestellte Analyse sind primär die Daten aus dem in ProfiLe-P entwickelten und in ProfiLe-P+ verwendeten FDW-Testinstruments (Gramzow, 2015) relevant.

2. Theoretische und Methodische Grundlagen

Die Analysen dieses Projekts sind wesentlich durch theoretische Konzeptualisierungen und bereits bestehende Ergebnisse zur professionellen Kompetenz und insbesondere zum FDW von (Physik-)Lehrkräften vorbereitet und strukturiert. Zur Analyse werden einerseits „klassische" Item-Response-Theorie (IRT)-basierte Niveauanalysen und andererseits ML- bzw. NLP-Methoden eingesetzt. Im Folgenden werden daher sowohl inhaltsbezogene theoretische als auch methodische Grundlagen der Analysen dargestellt.

2.1. Professionswissen von (Physik-)Lehrkräften

Die professionelle Kompetenz von (angehenden) Lehrkräften ist bereits lange zentraler Gegenstand fachdidaktischer und bildungswissenschaftlicher Forschung – sowohl international (z. B. Gess-Newsome, 1999; Hume et al., 2019; Neumann et al., 2019; Shulman, 1986, 1987) als auch im deutschsprachigen Raum (Baumert & Kunter, 2006; Kirschner et al., 2017; Kleickmann et al., 2014; Riese et al., 2015; Sorge et al., 2019). Das Professionswissen als kognitive Komponente der professionellen Kompetenz (siehe auch Abbildung 2.1) wird dabei als wesentlich für die Handlungsqualität im Unterricht und den Unterrichtserfolg aufgefasst (Ball et al., 2001; Harms & Riese, 2018; Terhart, 2012). In den frühen Konzeptualisierungen nach Shulman (1987) wurde sich wesentlich auf das Professionswissen fokussiert, das in die sieben Bereiche (1) *Content Knowledge*, (2) *General Pedagogical Knowledge*, (3) *Pedagogical Content Knowledge*, (4) *Curriculum Knowledge*, (5) *Knowledge of Learners and Their Characteristics*, (6) *Knowledge of Educational Contexts*, (7) *Knowledge of Educational Ends, Purposes, and Values*[3] unterteilt wurde. Im deutschsprachigen Raum hat sich vor allem das Modell für die professionelle Kompetenz von Lehrkräften nach Baumert und Kunter (2006) aus dem COACTIV-Projekt (Baumert & Kunter, 2011) durchgesetzt. Dieses ursprünglich für den Bereich der Mathematik entwickelte Modell wurde seitdem für unterschiedliche Fachrichtungen adaptiert.

In der Physik wird dieser Entwicklung folgend häufig die Adaption des Modells professioneller Handlungskompetenz nach Riese (2009) verwendet (Abbildung 2.1). Dieses Modell umfasst insbesondere nicht nur Professionswissen, sondern auch motivationale, volitionale und soziale Aspekte. Darunter fallen beispielsweise sog. „Belief Systems", also Wertesysteme und Rollenbilder der Lehrkräfte. Auch wenn diese Aspekte wichtige Elemente professioneller Kompetenz sind, spielen sie für die Analysen in diesem Projekt eine untergeordnete Rolle. Im Folgenden wird sich daher auf die Beschreibung des Professionswissens bzw. der kognitiven Aspekte des Modells professioneller Handlungskompetenz beschränkt.

Das Professionswissen wird im deutschsprachigen Raum meist den Modellen von Baumert und Kunter (2006) sowie Shulman (1987) folgend in die drei Domänen Fachwissen (FW),

[3] Englischsprachige Begriffe werden hier im Original benannt, um Vermischungen von Konstrukten, wie dem Fachdidaktischen Wissen und dem Pedagogical Content Knowledge (siehe Abschnitt 2.2) zu vermeiden.

Pädagogisches Wissen (PW) und Fachdidaktisches Wissen (FDW) gegliedert. Zusätzliche Bereiche wie beispielsweise das *Curriculum Knowledge* (Shulman, 1987) werden dabei entweder einer dieser drei übergeordneten Professionswissensdomänen oder einem anderen Bereich des Modells der Professionskompetenz untergeordnet. Unterschiede in den Modellen entstehen durch im Detail unterschiedliche Konzeptualisierungen der einzelnen Wissens- und Kompetenzbereiche[4].

Abbildung 2.1 Modell der Professionellen Handlungskompetenz nach Riese (2009) in Anlehnung an Baumert und Kunter (2006) sowie Blömeke et al. (2008b).

Das FW beschreibt fachliches Wissen, zunächst ohne expliziten Bezug zum Lehrberuf (Baumert & Kunter, 2006; Riese, 2009; Shulman, 1987). Es wird allerdings davon ausgegangen, dass dieses Wissen über den in der Schule behandelten Umfang hinausgehen muss, damit die Lehrkräfte fachliche Inhalte im Rahmen eines größeren Kontextes einordnen können. Erst dadurch sind sie befähigt, die Entwicklungen ihrer Schülerinnen und Schüler zu antizipieren und sie auf eine potenzielle spätere Vertiefung ihrer Kenntnisse in Studium, Ausbildung oder Beruf vorzubereiten (Blömeke et al., 2008b; Krauss et al., 2008). Die konkreten Inhaltsbereiche des FW, in denen Lehrkräfte entsprechende Kenntnisse erwerben sollen, sind Teil des gesellschaftlichen und wissenschaftlichen Diskurses, für die Physik besteht aber weitestgehender Konsens (Schiering, 2021; Sorge et al., 2019). Dabei sind in Deutschland die Bereiche Mechanik, Elektrodynamik, Optik, Thermodynamik, Festkörperphysik, Atom- und Kernphysik, spezielle Relativitätstheorie sowie Quantenphysik festgelegt (Kultusministerkonferenz [KMK], 2024). Es existieren zudem empirisch fundierte Niveaumodelle des FW für die Physik, die in Abschnitt 2.3 noch einmal thematisiert werden.

Das PW wird allgemein als fachunabhängiges Wissen über allgemeindidaktische und pädagogische Konzepte verstanden (Baumert & Kunter, 2006; Voss et al., 2015). Es existieren unterschiedliche Konzeptualisierungen zu sog. „Facetten“ (~ Unterkategorien) dieser Professionswissensdomäne. Beispielsweise unterteilen König und Seifert (2012) das PW in die

[4] Genaueres zu diesen Unterschieden insbesondere für das FDW werden in Abschnitt 2.2 genauer erläutert.

drei Bereiche (1) *Erziehung und Bildung*, (2) *Unterricht und allgemeine Didaktik* sowie (3) *Schulentwicklung und Gesellschaft*. Mit einem ähnlichen Grundansatz aber einer feineren Unterteilung differenzieren Kunter et al. (2017) PW in die sechs inhaltlichen Bereiche (1) *Unterrichtsgestaltung*, (2) *Schulorganisation*, (3) *Bildungstheorie*, (4) *Lernen und Entwicklung*, (5) *Diagnostik und Evaluation* sowie (6) *Lehrerberuf als Profession* und heben dabei außerunterrichtliche Aspekte der Tätigkeit von Lehrkräften hervor.

Das FDW wird als das Wissen zu Vermittlung von konkretem Fachwissen an eine konkrete Zielgruppe verstanden (Shulman, 1987). Es wird angenommen, dass gerade das FDW diejenige Wissenskategorie ist, in der sich Lehrkräfte von reinen Fachwissenschaftler:innen bzw. reinen Pädagog:innen unterscheiden (Hume et al., 2019; Neumann et al., 2019; Shulman, 1987). Das FDW ist die in diesem Projekt primär betrachtete Professionswissensdimension, weshalb ihm hier ein eigener umfangreicherer Abschnitt gewidmet ist (Abschnitt 2.2).

2.2. Fachdidaktisches Wissen

Die frühen Arbeiten zum Professionswissen bzw. der professionellen Kompetenz von (Shulman, 1986, 1987) können ebenso als Pionierarbeiten zur Konzeptualisierung des FDW angesehen werden. In seinem als *Pedagogical Content Knowledge* (PCK) bezeichneten Konstrukt fasst Shulman (1987) Wissen zusammen, das nötig ist, um bestimmtes Fachwissen einer bestimmten Zielgruppe zu vermitteln. Dabei wird eine grundsätzliche Abhängigkeit vom thematisierten Fachinhalt angenommen. Parallel dazu, sowohl in Fortführungen bzw. Adaptionen dieses Ansatzes als auch unter Einbezug hiesiger Bildungstraditionen, hat sich im deutschsprachigen Raum das Konstrukt des FDW entwickelt (Baumert & Kunter, 2006; Gramzow, 2015; Kröger, 2019; Riese, 2009). Unter anderem aufgrund von Unterschieden zwischen den Bildungstraditionen des englischsprachigen und des deutschsprachigen Raums sind FDW und PCK zwar nah verwandt, aber nicht deckungsgleich (z. B. Gramzow et al., 2013; Vollmer & Klette, 2023).

Verortung des Fachdidaktischen Wissens in Rahmenmodellen

Um diese Unterschiede zu verdeutlichen und das in diesem Projekt zugrundeliegende Verständnis des Konstrukts des FDW klar darzulegen, wird FDW in diesem Abschnitt im Kontext zweier Rahmenmodelle verortet.

Rahmenmodell A). International hat sich in den letzten Jahren vor allem das sog. *Refined Consensus Model* (RCM) *of PCK* (Carlson et al., 2019; Hume et al., 2019) durchgesetzt. Dieses Modell konzeptualisiert PCK im Rahmen von drei Domänen (Carlson et al., 2019, 83–91):

1. *Collective PCK* (cPCK): Das cPCK stellt die kollektive Wissensbasis von fachdidaktischen Communities dar, umfasst also einen Korpus an explizierbarem, eher deklarativem Wissen, das beispielsweise (aber nicht exklusiv) in fachdidaktischer Fachliteratur zu finden ist.
2. *Personal PCK* (pPCK): Das pPCK stellt die persönliche Wissensbasis der einzelnen Lehrkraft dar. Wie auch das cPCK wird pPCK als explizierbar betrachtet. Der Transfer von cPCK zu pPCK wird dabei von unterschiedlichen (z. T. äußeren) Rahmen-

bedingungen beeinflusst, wie beispielsweise Eigenheiten der Schule, des Bildungssystems oder von Schülerinnen und Schülern. Die Gesamtheit dieser „Filter" zwischen dem cPCK und pPCK wird im RCM auch als *Learning Context* bezeichnet. Umgekehrt kann pPCK auch das cPCK beeinflussen, indem sich beispielsweise in einer fachdidaktischen Community innerhalb einer Schule bestimmte Erfahrungen zu einer kollektiven Wissensbasis verfestigen.

3. *Enacted PCK* (ePCK): Das ePCK umschreibt das PCK, das konkreten gezeigten Handlungen im fachdidaktischen Kontext zugrunde liegt. Solche Handlungen sind beispielsweise Unterrichtsvorbereitungen oder konkret für die Physik die Erklärung physikalischer Phänomene. Dieses Wissen ist in der Regel nicht mehr explizierbar. Es wird angenommen, dass das ePCK sich im Rahmen des sog „*Plan-Teach-Reflect-Cycle*" (PTR-Cycle) in einem zirkulären Prozess entwickelt (Alonzo et al., 2019) und, dass ePCK in wechselseitiger Beziehung zum pPCK steht.

Die Filter zwischen den einzelnen PCK-Domänen, ihre Auswirkung auf die Entwicklung von PCK und Professionswissen und die innere Struktur der einzelnen PCK-Domänen genauer empirisch zu untersuchen ist herausfordernd (z. B. Behling et al., 2022a; Kulgemeyer et al., 2023). Zum ePCK gibt dabei Ansätze, die Auswirkung einer explizit dem PTR-Cycle folgenden Ausbildungsmaßnahme auf andere Komponenten des PCK bzw. professioneller Kompetenz zu untersuchen (Behling et al., 2022b). Dabei zeigten sich positive Auswirkungen auf das pPCK und motivationale Orientierungen. Es bleibt allerdings unklar, ob und inwieweit sich die im PTR-Cycle für das ePCK-beschriebene Feinstruktur in Form einer Unterscheidung zwischen „ePCK-plan", „ePCK-teach" und „ePCK-reflect" (Alonzo et al., 2019) in ähnlicher Weise auch für die anderen PCK-Domänen zeigt. Zudem ist unklar, ob potenzielle trennbare pPCK-Komponenten auch verstärkt mit einzelnen ePCK-Komponenten zusammenhängen. Neben den beschriebenen Filtern zwischen den einzelnen Domänen des PCK werden im RCM zudem äußere Einflussfaktoren sog. *Professional Knowledge Bases* wie beispielsweise Wissen über Assessment, Wissen über das Curriculum oder auch Fachwissen beschrieben (Carlson et al., 2019). Diese Wissensbereiche stehen wiederum in wechselseitiger Beziehung zum PCK. Im RCM wird PCK also schichtweise modelliert, wie in Abbildung 2.2 dargestellt.

Das deutschsprachige Konstrukt des FDW ist eng verwandt zum PCK, umfasst aber insbesondere weniger die konkret gezeigten fachdidaktischen Handlungen, bzw. das durch diese Handlungen implizit gezeigte Wissen (Gramzow, 2015; Sorge et al., 2019). Die konkrete Handlung wird also im Konstrukt des FDW „weniger mitgedacht". In mehreren Projektverbunden (Kulgemeyer et al., 2023; Schiering et al., 2023) und theoretischen Arbeiten (z. B. Vollmer & Klette, 2023) wird daher FDW im Wesentlichen als vergleichbar bis deckungsgleich mit dem pPCK beschrieben. Auch augenscheinlich liegt es nahe, dass das in schriftlichen FDW-Leistungstests, wie denen nach Gramzow (2015) oder Kröger (2019), explizierbares persönliches Wissen der Proband:innen abgefragt wird.

Rahmenmodell B). Ein weiteres prominentes Modell zur Konzeptualisierung von Kompetenzen, welches für das FDW weit verbreitet angewendet wird, ist das sog. *Model of Competence* (MoC) bzw. *Kontinuumsmodell* nach Blömeke et al. (2015). Dieses Modell beschreibt (professionelle) Kompetenz als Ganzes in Form eines Kontinuums zwischen

einerseits latenten kognitiven Dispositionen und andererseits Performanz in für die Profession prototypischen Anforderungssituationen (siehe Abbildung 2.3 Model of Competence / Kontinuumsmodell nach). Dazwischen werden situationsspezifische Fähigkeiten wie Interpretationsfähigkeiten, Wahrnehmung und Entschei-dungsfindung positioniert. Im Kontext des PCK bzw. FDW kann man das RCM als diskretisierte Form des MoCs verstehen, bei denen konkrete Stufen zwischen Dispositionen und Performanz explizit inhaltlich voneinander abgegrenzt werden. PCK ist dann eher als ein Konstrukt zu verstehen, das sich über die gesamte Bandbreite des MoCs erstreckt, währen FDW eher auf der Seite der kognitiven Dispositionen zu verorten ist (Kulgemeyer et al., 2023).

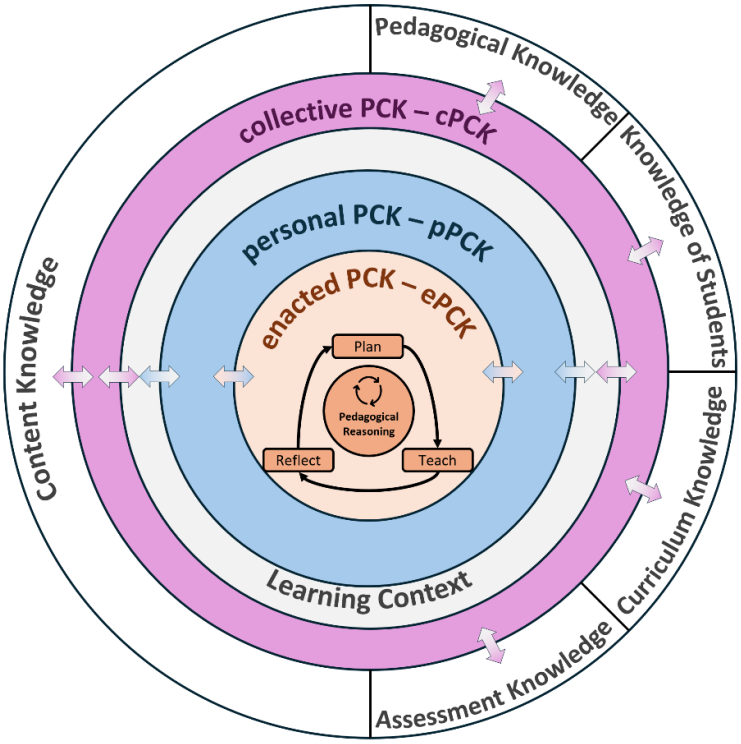


Abbildung 2.2 Schematische Darstellung des Refined Consensus Model of PCK (vereinfacht nach Carlson et al., 2019, S. 83).

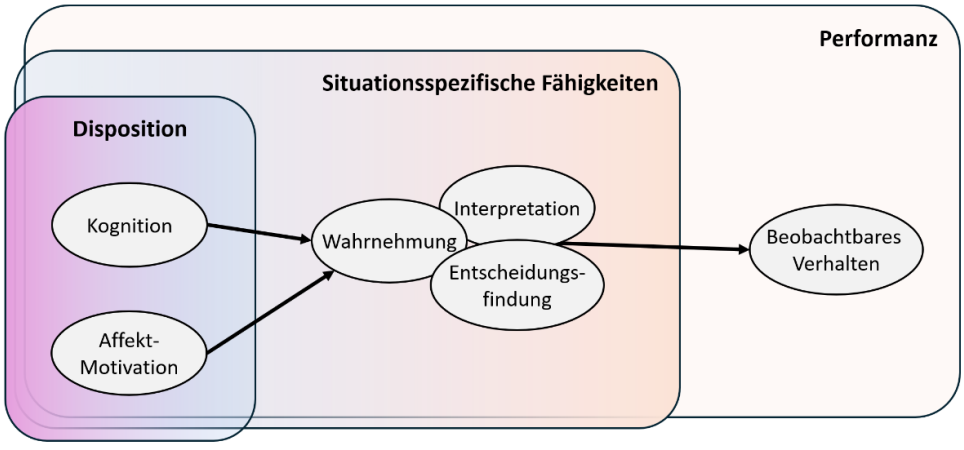


Abbildung 2.3 Model of Competence / Kontinuumsmodell nach Blömeke et al. (2015, S. 7)

Ähnlich zu den Überlegungen zu potenziellen pPCK-Komponenten und deren Zusammenhang zu den ePCK-Komponenten im RCM stellt sich auch im Kontext des MoC die Frage, inwieweit einzelne Dispositionen (FDW-Komponenten) unterschieden werden können, die mit einzelnen Fähigkeiten bzw. der Performanz in unterschiedlichen prototypischen Anforderungssituationen, wie der Unterrichtsplanung (Schröder et al., 2020), dem Erklären physikalischer Phänomene (Kulgemeyer et al., 2020) oder der Reflexion (Kulgemeyer et al., 2021) zusammenhängen.

Zusammengefasst wird in dieser Arbeit PCK als Konstrukt aufgefasst, welches neben explizierbarem Wissen (cPCK und pPCK) auch Performanz in Handlungssituationen, bzw. das für solche Handlungen notwendige implizite Wissen (ePCK) umfasst. Das FDW wird primär als explizierbares Wissen verstanden und schließt implizites Handlungswissen nicht direkt mit ein. Aufgrund der Konzeption der Erhebung der in diesem Projekt genutzten Daten wird daher in Übereinstimmung mit ähnlichen Ansätzen (Kulgemeyer et al., 2023; Schiering et al., 2023) davon ausgegangen, dass das im entsprechenden Testinstrument (Gramzow, 2015) primär erfasste FDW am ehesten mit dem pPCK vergleichbar ist und primär eine kognitive Komponente eines größeren Kompetenzbegriffs abdeckt. Im Folgenden wird daher hauptsächlich der Begriff „FDW“ im dargestellten Verständnis des Konstrukts weiterverwendet[5]. Zudem wurde dargestellt, dass die gebräuchlichsten Rahmenmodelle zur Konzeptualisierung des FDW Desiderate zur genaueren Beschreibung der inneren Struktur des FDW aufweisen.

Operationalisierung und Messung des Fachdidaktischen Wissens

In frühen Arbeiten wurde das FDW meist anhand von Selbsteinschätzungen oder über die Erfassung distaler Merkmale gemessen (vgl. Baumert & Kunter, 2006). In aktuellerer Forschung werden (zumindest im deutschsprachigen Raum) jedoch eher Leistungstests verwendet. Das FDW wird dabei (nicht nur zu Testzwecken) meist dreidimensional modelliert (z. B. Gramzow, 2015; Kröger, 2019; Tepner et al., 2012). Die erste Dimension ist der adressierte fachliche Inhalt, also in der Physik beispielsweise Inhaltsfelder wie Mechanik oder Optik (KMK, 2024). Darüber hinaus sind sog. *fachdidaktische Inhalte* oder *Facetten* eine zentrale Dimension dieser Modelle. Die Facetten beschreiben dabei unterschiedliche inhaltliche Bereiche des FDW und es existieren viele Konzeptualisierungen, in denen jeweils eine unterschiedliche Auswahl von Facetten eingeschlossen wird (z. B. Park & Oliver, 2008; Riese, 2009; übersichtsartig dargestellt bei Schmelzing, 2010, S. 23 Kirschner, 2013, S. 32). Zentrale Facetten, die bereits in Shulmans (1986) ursprünglichen Arbeiten implizit berücksichtigt wurden und auch als Minimalkonsens angesehen werden können (Kirschner, 2013, S. 32; Schmelzing, 2010, S. 23) sind *Instruktions- und Vermittlungsstrategien* sowie *Schüler und Schülerkognitionen*[6]. Darüber hinaus sind im Testinstrument nach Gramzow

[5] In den deutschsprachigen Artikeln 1 und 3 (Kapitel 4 und 6) wird der Begriff „FDW“ im selben Verständnis genutzt. Im englischen Artikel 2 (Kapitel 5) wird primär der Begriff „pPCK“ genutzt, um das Konstrukt auf eine international geläufige Weise zu bezeichnen.

[6] Die Facette *Schüler und Schülerkognition* sowie die Begriffe „Schülerkognition“ und „Schülervorstellungen“ werden der Standardliteratur (z. B. Schecker et al., 2018) folgend nicht geschlechtsneutral umformuliert.

(2015), das zur Erhebung der im Rahmen der hier vorgestellten Analysen verwendeten Daten genutzt wurde, die Facetten *Experimente und Vermittlung eines angemessenen Wissenschaftsverständnisses* (kurz *Experimente*) sowie *Fachdidaktische Konzepte* abgebildet. Dabei handelt es sich explizit um eine aus Gründen der Testökonomie getroffene Auswahl von Facetten und nicht um eine vollständige Liste (Gramzow, 2015). Für eine ausführlichere Beschreibung der Inhalte der einzelnen Facetten sei auf (Gramzow, 2015, S. 96–102) verwiesen. Während die bisherigen Dimensionen fachlicher Inhalt und fachdidaktische Facetten allgemein zur Modellierung des FDW dienen, wird zur Entwicklung von Testinstrumenten meist noch eine dritte Dimension ergänzt. Diese dient der Argumentation von Klieme et al. (2003) folgend der Anreicherung der Testinstrumente mit Aufgaben aus unterschiedlichen kognitiven Anforderungsbereichen. So beschreiben Kröger (2019) und Tepner et al. (2012) die Dimension *Wissensart(en)*, während Gramzow (2015) in der Dimension der *kognitiven Anforderungen* die Kategorien *Reproduzieren*, *Anwenden* und *Analysieren* berücksichtigt. Für die genaue Beschreibung des Verständnisses dieser Kategorien sei erneut auf Gramzow (2015, S. 111–112) verwiesen.

Empirische Untersuchungen zum FDW

Mithilfe der beschriebenen Modellierungen und Operationalisierungen konnten empirisch Zuwächse des FDW im Studium und im Vorbereitungsdienst nachgewiesen werden (Kirschner, 2013; Kröger, 2019; Sorge et al., 2019; Vogelsang et al., 2019). Darüber hinaus zeigten sich sowohl direkte Zusammenhänge zwischen FDW und FW sowie FDW und PW (Kirschner, 2013; Kirschner et al., 2017; Kulgemeyer et al., 2012; Riese et al., 2015; Sorge et al., 2019; Tepner & Dollny, 2014) als auch die Bedeutsamkeit des FDW für die Entwicklung von FW und PW (Sorge et al., 2018). Mittlerweile liegen zudem Ergebnisse vor, die Effekte des FDW auf die Performanz in prototypischen Anforderungssituationen wie beispielsweise (1) der Unterrichtsplanung (Behling et al., 2022b; Riese et al., 2022b; Schröder et al., 2020), (2) dem Erklären physikalischer Phänomene (Kulgemeyer et al., 2020; Kulgemeyer & Riese, 2018), (3) dem Reflektieren über Unterricht (Kulgemeyer et al., 2021), (4) der kognitiven Aktivierung (Förtsch et al., 2016; She et al., 2024), (5) der Nutzung von physischen Modellen (Förtsch et al., 2018) sowie (6) diagnostischen Handlungen (Kramer et al., 2021) zeigen. Strübe (2020) konnte allerdings keine bzw. nur schwache Zusammenhänge zwischen dem FDW und der Arbeit mit Modellen bzw. Experimenten bei Chemielehrkräften nachweisen. Detail-Betrachtungen deuteten hier darauf hin, dass primär bestimmte Facetten des FDW Auswirkungen auf diese Handlungsaspekte haben könnten (Strübe, 2020, S. 208).

Weiterhin zeigten sich auch (indirekte) Einflüsse auf die kognitive Aktivierung und Leistung von Schülerinnen und Schülern (Blömeke et al., 2022; Förtsch et al., 2016), wobei die Studienlage zu den Auswirkungen von Professionswissen auf Schüler:innen nicht eindeutig ist (Cauet et al., 2015; Liepertz & Borowski, 2019). In diesem Kontext konnten Tröger et al. (2017) mithilfe von Item-Response-Modellierungen (s. u.) und darauf aufbauenden Regressionsanalysen zeigen, dass Schüler:innen mit geringem Vorwissen von hohem FDW über Sprachnutzung ihrer Lehrkräfte profitieren. Weitere Analysen verdeutlichen, dass die Entwicklung des Professionswissens und des FDW im Speziellen durch Studienstrukturen bedingt wird (Schiering et al., 2021).

Auch, wenn Studien wie die genannten die Bedeutsamkeit des FDW weiter unterstreichen, so liefern sie doch keine weitere empirische Beschreibung der inneren Struktur des FDW. Die Dimensionalisierung mit den Achsen *fachlicher Inhalt*, *kognitive Anforderungen* und *Facetten* (oder ähnliche Modelle aus den anderen genannten Projekten) sind eher normativ motiviert. Auch die zentrale Dimension der fachdidaktischen Facetten ist zwar je nach Studie argumentativ, durch Experteninterviews und Curriculumsanalysen von Lehrerbildungsprogrammen fundiert (z. B. Kulgemeyer et al., 2020; Magnusson et al., 1999; Park & Oliver, 2008; Schiering et al., 2023), allerdings somit immer noch von eher theoretisch-normativem Charakter. Für ihr Testinstrument konnten Riese et al. (2017) bzw. Gramzow (2015) mithilfe von IRT-Modellvergleichen allerdings zeigen, dass sowohl die drei abgebildeten kognitiven Anforderungen als auch die vier eingeschlossenen Facetten als empirisch trennbare Subskalen aufgefasst werden können. Dies konnte aber nur für eine (relativ kleine) Stichprobe von fortgeschrittenen Studierenden gezeigt werden und auch die verwendeten statistischen Informationskriterien (AIC und BIC) zeigten nur eine schwache Bevorzugung der Modelle mit Subskalen. Zur empirisch fundierten Beschreibung der inneren Struktur des FDW sind also weitere Untersuchungen nötig. Einen dazu bereits erprobten Ansatz stellen hierarchische Niveaumodelle dar, die im nächsten Abschnitt genauer beschrieben werden.

2.3. Hierarchische Niveaumodelle auf Basis von Item-Response-Modellen

Um Niveaus in den Ausprägungen von Kompetenzen inhaltlich auf Basis von Testdaten zu modellieren, haben sich IRT-basierte Niveaumodelle etabliert (Hartig, 2007; Mullis et al., 2016; Organisation for Economic Cooperation and Development [OECD], 2018). Solchen Ansätzen liegt meist ein IRT-Modell der Testdaten zugrunde, welches Personenfähigkeiten und Aufgabenschwierigkeiten auf einer gemeinsamen Skala abbildet (s. u.). Das „ursprüngliche“ IRT-Modell nach Rasch (1960) basiert auf der folgenden Annahme: Die Wahrscheinlichkeit, dass eine Person p mit der „Personenfähigkeit“ θ_p die Aufgabe i mit der „Aufgabenschwierigkeit“ σ_i korrekt löst beträgt

$$P(X_{pi} = 1) = \frac{1}{1 + \exp(\sigma_i - \theta_p)}.$$

Aufgabenschwierigkeit und Personenfähigkeit sind somit relativ zueinander zu interpretieren. Ist $\sigma_i \gg \theta_p$ (bzw. $\sigma_i \ll \theta_p$) so strebt diese Funktion gegen 0 (bzw. 1), d. h. die Wahrscheinlichkeit, dass Person p die Aufgabe i löst ist gering (bzw. hoch). Ist $\sigma_i = \theta_p$, so ist die Wahrscheinlichkeit, dass Person p Aufgabe i korrekt löst gerade gleich $50\,\%$. Da nur die Differenz der Parameter σ_i und θ_p für die Wahrscheinlichkeit relevant sind, werden die Parameter demnach auf einer gemeinsamen Skala abgebildet, was in den Verfahren zur Niveauanalyse ausgenutzt wird (s. u.). Es gibt unterschiedliche Möglichkeiten aus der Datentabelle der Punktzahlen der Personen in den Aufgaben diese Schwierigkeits- bzw. Fähigkeitsparameter zu schätzen.

Es existieren unterschiedliche Erweiterungen und Verallgemeinerungen des Rasch-Modells, die zumeist alle unter der Bezeichnung „Item-Response-Modell“ bzw. IRT-Modell

zusammengefasst werden (z. B. Moosbrugger & Kelava, 2020). Für das hier vorgestellte Projekt ist insbesondere das sog. „Partial Credit Modell“ (Masters, 1982) zu nennen, welches das ursprüngliche Rasch-Modell auf mehrstufige (z. B. 0, 1 und 2 Punkte) Aufgaben verallgemeinert. Die zur Modellierung der Aufgabenschwierigkeiten genutzten Parameter sind dann allerdings nicht mehr so unmittelbar interpretierbar wie die Aufgaben-Parameter im Rasch-Modell. Die analog zum Rasch-Modell interpretierbaren „Thurstone-Threshold“-Parameter (auch nur „Thurstone-Thresholds“) lassen sich allerdings nach dem Modell-Fit ebenfalls berechnen (Linacre, 1998). Somit liegt auch im Partial Credit Modell eine Möglichkeit vor, Aufgabenschwierigkeiten und Personenfähigkeiten auf einer gemeinsamen Skala abzubilden und zu modellieren. Die gemeinsame Darstellung dieser Parameter in einem Plot wird auch „Wright Map“ genannt (siehe Abbildung 4.6 & Abbildung 4.7).

Zur Bildung von Niveaus auf Basis von Wright Maps und den ihnen zugrundeliegenden Parameterschätzungen aus IRT-Modellen existieren im Wesentlichen drei unterschiedliche Methoden, die Woitkowski (2020) in seiner Adaption eines dieser Ansätze zur Niveaumodellierung des physikalischen Fachwissens gegenüberstellt: (1) das Scale-Anchroing-Verfahren, (2) der regressionsanalytische Ansatz und (3) die Bookmark-Methode. Von diesen Methoden sind für das hier vorgestellte Projekt vor allem das Scale-Anchoring-Verfahren und der regressionsanalytische Ansatz relevant.

Im Scale-Anchroing-Verfahren (Mullis et al., 2016) werden zunächst drei[7] Personengruppen gebildet, wobei jeweils eine Gruppe niedrige, eine mittlere und eine hohe Fähigkeitsparameter aufweist. Anhand der Anteile von Personen aus diesen Gruppen, die eine Aufgabe gelöst bzw. (im Falle von polytomen Aufgaben) eine bestimmte Punktzahl in der Aufgabe erreicht haben, werden die Aufgaben bzw. die Punkteschwellen der Aufgaben wiederum in Gruppen eingeteilt. Die Mittelwerte der Schwierigkeitsparameter bzw. Thurstone-Thresholds dieser Aufgabengruppen dienen anschließend als Niveaugrenzen und die inhaltliche Beschreibung der Niveaus folgt aus den inhaltlichen Beschreibungen der Aufgaben, die sich nahe an diesen Grenzen befinden. Der genaue Ablauf des Verfahrens wird in der konkreten Anwendung in Abschnitt 4.4.3 noch einmal genauer erläutert. Das Scale-Anchoring-Verfahren ist durch ein hohes Maß an Datengetriebenheit gekennzeichnet und kann in diesem Sinne im Vergleich zu den anderen Verfahren als besonders objektiv aufgefasst werden. Für besonders aussagekräftige Ergebnisse ist bei der Verwendung des Scale-Anchoring-Verfahrens jedoch eine hohe Anzahl an Aufgaben im Testinstrument und Proband:innen optimal, weshalb das Scale-Anchroing-Verfahren bisher zumeist in den großen Schulleistungsstudien wie TIMSS (Mullis et al., 2016) und PISA (OECD, 2018) angewendet wurde.

Eine Alternative zum Scale-Anchroing-Verfahren stellen regressionsanalytische Ansätze dar (z. B. Blömeke et al., 2008a; Nold et al., 2008). Anders als beim Scale-Anchoring-

[7] Das Scale-Anchoring-Verfahren ermöglicht auch feinere Unterteilungen der Fähigkeiten, d. h. die Beschreibung einer größeren Anzahl an Niveaus. Dazu müssen im ersten Schritt die Personen in eine größere Anzahl an Leistungsgruppen unterteilt werden. Werden die Personen in n Gruppen eingeteilt, so erhält man anschließend $n + 1$ Niveaus (und ein Niveau „< 0“, über das keine weiteren inhaltlichen Aussagen getroffen werden können). Der Übersicht halber, wird hier aber das Verfahren für lediglich drei Personengruppen bzw. 4 Niveaustufen erläutert, da eine feinere Unterteilung mit der verfügbaren Datenbasis in diesem Projekt nicht angestrebt wurde (siehe Kapitel 4).

Verfahren findet hier eine Re-Analyse aller Aufgaben des Testinstruments bereits zu Beginn statt. Dazu werden geeignete schwierigkeitserzeugende Merkmale theorie- oder literaturbasiert ermittelt und niveauartig beschrieben. Die Aufgaben werden dann diesen Stufen zugeordnet. Die Passung und Eignung dieses Modells schwierigkeitserzeugender Merkmale zu den Daten wird mithilfe der Varianzaufklärung einer linearen Regression bzw. ANOVA bezüglich der IRT-Aufgabenschwierigkeiten evaluiert. Liegt eine hohe Varianzaufklärung und dementsprechend ausreichende Passung vor, können die Mittelwerte der Aufgabenschwierigkeiten der somit entstandenen Aufgabengruppen als Niveaugrenzen genutzt werden und die Personen gemäß ihrer Fähigkeitsparameter zu den Niveaus zugeordnet werden. Die inhaltlichen Beschreibungen der Niveaus folgen dann direkt aus dem Modell schwierigkeitserzeugender Merkmale. Im Vergleich zum Scale-Anchoring-Verfahren hat ein regressionsanalytischer Ansatz den Vorteil, dass von vorneherein alle Aufgaben zur Niveaubeschreibung genutzt werden können und somit auch kleinere Testinstrumente mit geringerer Aufgabenanzahl ggf. besser ausgeschöpft werden können. Allerdings stellt die Entwicklung eines geeigneten Modells schwierigkeitserzeugender Merkmale einen zusätzlichen aufwändigen Prozess dar, der zudem als weniger objektivierbar angesehen werden kann als das weitestgehend datengetriebene Vorgehen beim Scale-Anchoring-Verfahren.

Regressionsanalytische Ansätze wurden bereits mehrfach im Kontext der Bildungsforschung im deutschsprachigen Raum genutzt. König (2009) verwendete eine Kombination von drei Stufen sprachlicher Komplexität und zwei kognitiven Anforderungsstufen als schwierigkeitserzeugende Merkmale, um ein Niveaumodell des PW zu entwickeln. Dabei hat sich insbesondere das kognitive Anforderungsniveau als bedeutsam herausgestellt. Bernholt (2010) entwickelte ein Niveaumodell für Fachwissen in der Chemie. Dabei leitete er vier Stufen Komplexitätsstufen orientiert am inhaltsunabhängigen Modell hierarchischer Komplexität (Commons et al., 2014; Commons et al., 1998) ab. Dieser Ansatz wurde von Woitkowski und Riese (2017) für die Physik mit Erfolg adaptiert und genutzt, um den Fachwissenserwerb im Studienanfängerbereich (Woitkowski, 2019) zu untersuchen und Entwicklungstrajektorien über den Studienverlauf abzuleiten (Woitkowski, 2020).

Für das FDW wurden bereits Analysen mithilfe des Scale-Anchoring-Verfahrens durchgeführt. Schiering et al. (2019) bzw. Schiering et al. (2023) wendeten hierbei das Scale-Anchoring-Verfahren auf ihr Testinstrument als Ganzes an. Sie konnten keine Systematik bzgl. des Auftretens bestimmter fachdidaktischer Inhalte (bzw. Facetten) feststellen. Zeller et al. (2022) führten ähnliche Analysen mithilfe einer geschlossenen Version des Testinstruments von Gramzow (2015) durch, wobei die fachdidaktischen Facetten von vorneherein getrennt voneinander betrachtet wurden. Unabhängig von den konkreten Projektkontexten, den in den Instrumenten abgebildeten Facetten oder dem jeweils adressiertem Fachwissen zeigte sich in der Tendenz eine Parallele zwischen den Ergebnissen von Zeller et al. (2022) und Schiering et al. (2019) bzw. Schiering et al. (2023): Bezüglich kognitiver Prozesse scheint sich das FDW in niedrigen Niveaus auf reproduktive Aspekte zu beschränken und in höheren Niveaus auch analytische und anwendungsorientierte Elemente mit einzuschließen. Eine projektübergreifende Betrachtung von FDW-Niveaustufen ist dementsprechend das erste Zielpaket dieses Projekts. Da die Ergebnisse von Studien der Projekte, aus denen die Analysen von Schiering et al. (2019) bzw. Schiering et al. (2023) und Zeller et al. (2022) hervorgegangen

sind, bisher aufgrund der im Detail unterschiedlichen Operationalisierung des FDW isoliert stehen, verspricht hier eine projektübergreifende Betrachtung von inhaltlichen FDW-Niveaus auch die Vorbereitung einer Herstellung von Vergleichbarkeit anderer Ergebnisse aus den jeweiligen Projektkontexten. Für die projektübergreifende Analyse wird zunächst aufgrund der bereits bestehenden vielversprechenden Ansätze das Scale-Anchoring-Verfahren auf beide Datensätze angewandt und ein Modellvergleich durch die sich ergebenden Niveauformulierungen angestrebt. Darüber hinaus wird orientiert an der erfolgreichen Nutzung des Modells hierarchischer Komplexität für einen regressionsanalytischen Ansatz zur Niveaubildung für das physikalische Fachwissen (Woitkowski & Riese, 2017) auch ein gemeinsames Modell hierarchischer Komplexität für das FDW entwickelt und dessen Passung zu den Datensätzen überprüft. Die vollständige Analyse ist in Kapitel 4 ausführlich dargestellt; insbesondere wird die Entwicklung bzw. Adaption des Modells hierarchischer Komplexität nach Commons et al. (1998; siehe auch Commons et al., 2014) für das FDW in Abschnitt 4.2.3 ausführlicher dargestellt und daher aus Platzgründen hier nicht noch einmal wiederholt.

Insgesamt werden somit im ersten Zielpaket inhaltliche Beschreibungen des FDW in Form von Niveaustufen entwickelt (siehe Abschnitt 4.5). Anders als bei den eher normativ-theoretischen Beschreibungen im Rahmen der ursprünglichen Operationalisierungen (z. B. Gramzow et al., 2013; Kröger, 2019; Park & Oliver, 2008), sind diese Ergebnisse induktiv aus vorhandenen (quantitativen) empirischen Daten abgeleitet. Die Ergebnisse dieser Analysen des ersten Zielpakets deuten aber auch darauf hin, dass wesentliche interessante inhaltliche Strukturen des FDW mit strikt hierarchischen Methoden, wie eben Niveaumodellen, nicht modelliert werden können. Im zweiten Zielpaket des Projekts werden daher nicht-hierarchische Analysen in den Blick genommen, die auf explorativ(er)en Machine-Learning-Methoden basieren. Diese werden daher in den nächsten Abschnitten ausführlicher dargestellt.

2.4. Machine Learning

Da sich in den Ergebnissen der Analysen zum ersten Zielpaket des hier vorgestellten Projektes zeigt, dass hierarchische Ansätze wesentliche interessante innere Strukturen des FDW nicht auflösen können (siehe Kapitel 4), werden darüber hinaus explorative Machine-Learning-Methoden zur nicht-hierarchischen Analyse des FDW in den Blick genommen (siehe Kapitel 5 & 6). Zudem ist das Leitziel des Projekts die Ermöglichung eines informativen Assessments des FDW auf Basis der Ergebnisse zu dessen inneren Struktur (siehe Kapitel 1 & 3). Dieses Assessment soll möglichst skalierbar sein und somit vor allem ohne hohen manuellen Aufwand bei der Bepunktung der offenen Aufgaben des genutzten FDW-Testinstruments auskommen. Zu diesem Zweck werden ebenfalls ML-basierte Methoden genutzt (siehe Kapitel 6). Im Folgenden werden daher der Grundansatz von ML-Workflows sowie konkrete Methoden zur Erreichung der genannten Ziele vorgestellt.

Definitionsansätze

Auch wenn die Begriffe *künstliche Intelligenz* (KI) und *Machine Learning* bereits seit Jahrzehnten im wissenschaftlichen und wirtschaftlichen Bereich genutzt werden (z. B. Samuel, 1959), so ist eine genaue Beschreibung ihrer eigentlichen Bedeutung nicht trivial. Früh

bezeichnete Samuel (1959) ML als „das Forschungsfeld, das Computern die Fähigkeit gibt zu lernen, ohne explizit programmiert zu sein“ (zit. nach Géron, 2019, S. 2, übers. JZ). Damit ist z. B. gemeint, dass keine Regel-Systeme nach dem „wenn-dann“-Prinzip zur Lösung eines Problems in den Computer eingegeben werden, sondern, dass das Problem eben durch „Lernen“ angegangen wird. Eine praktische Definition, um die eigentliche Bedeutung des Begriffs „Lernen“ in diesem Kontext zu fassen, stammt von Mitchell (1997):

> „*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.*“
>
> – Mitchell (1997, S. 2)

Als einfaches Beispiel kann hier eine einfache lineare Regression $x \propto y$ dienen: Der Computer lernt durch das Verarbeiten von Datenpaaren $(x_i, y_i), i = 1 \ldots N$ („experience E“), die Zielvariable y in Abhängigkeit von der unabhängigen Variable x vorherzusagen („task T“). Je mehr Datenpaare der Computer verarbeitet hat, umso besser ist die Vorhersagequalität („performance measure P“) – vorausgesetzt, die Annahme $x \propto y$ ist zutreffend. Wie die „Verarbeitung“ dieser Daten stattfindet, kann unterschiedlich sein (s. u.).

Der Begriff KI umfasst üblicherweise jegliche Methodik, die darauf ausgerichtet ist, menschliche Aufgaben durch Computer zu automatisieren und schließt im Unterschied zum ML-Begriff beispielsweise auch regelbasierte Systeme mit ein (z. B. Géron, 2019). Darüber hinaus wird häufig zudem der Begriff *Data Science* genutzt, der allgemein erkenntnisgewinnende, aber auch produktive Methoden unter Datennutzung zusammenfasst. Im Data Science Bereich wird sich dabei unter anderem der Methoden aus den KI- und ML-Bereichen bedient. Man kann die angesprochenen Forschungs- und Entwicklungsfelder in einem Schema wie in Abbildung 2.4 dargestellt miteinander in Beziehung setzen.

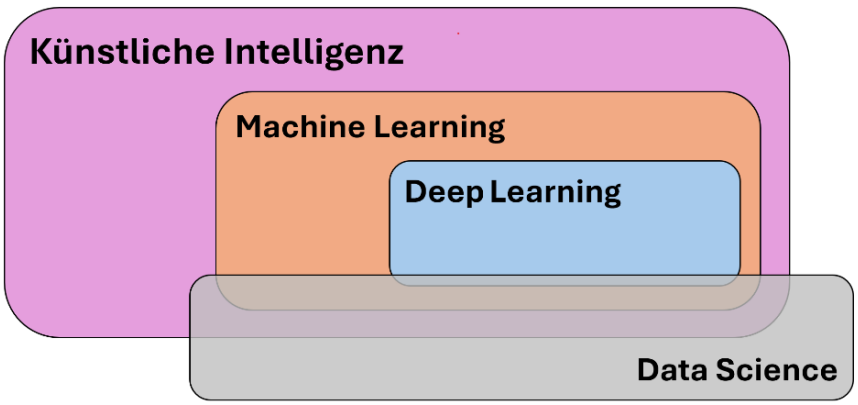


Abbildung 2.4 Darstellung der Bereiche Künstliche Intelligenz, Machine Learning, Deep Learning und Data Science. Der Bereich Deep Learning wird im Abschnitt 2.7 noch einmal aufgegriffen[8].

[8] Bei dieser Abbildung handelt es sich um eine weit verbreitete Standarddarstellung, die vielfach in unterschiedlichen Kontexten genutzt wird, und daher nicht einer expliziten Literaturquelle zugeordnet werden kann.

Neben diesen grundlegenden eher konzeptionellen Ansätzen zur Beschreibung des ML-Themenfeldes, gibt es auch praktischere Unterteilungen, die eher konkrete Ziele von entsprechenden Modellen und Verfahren in den Mittelpunkt stellen. Entscheidend ist hier insbesondere die Unterteilung in das sog. *Supervised Learning* und das sog. *Unsupervised Learning* (z. B. Duda et al., 2001; Géron, 2019). Methoden des Supervised Learnings haben das Ziel, aus bestimmten verfügbaren Variablen (*Features*) eine Zielvariable (*Target*) vorherzusagen. Features können beispielsweise demographische Merkmale sein und ein mögliches Target könnte schulischer Erfolg sein. Ein Supervised-Learning-Setting erfordert also die Verfügbarkeit eines Datensatzes, in dem diese Target-Daten auch vorhanden sind, d. h. üblicherweise manuell durch Menschen generiert wurden oder historisch vorliegen. Insbesondere bei manuell generierten Target-Daten spricht man auch von *Labels*. Im Gegensatz dazu zielt man im Unsupervised Learning darauf ab, Muster und Strukturen in Daten zu finden. Es geht also darum in einem Datensatz Gruppen von Datenpunkten zu finden, die sich auf eine gewisse Weise ähnlich sind. Das können beispielsweise Probanden einer Interviewstudie sein, die durch eine ähnliche Wortwahl charakterisiert sind.

Paradigmen

Man kann zudem zwischen algorithmischen Modellen, die häufig primär heuristisch motiviert sind, und probabilistischen Modellen (manchmal auch „bayesianische“ Modelle genannt) unterscheiden. Bei probabilistischen Modellen wird eine Wahrscheinlichkeitsverteilung, der die Daten folgen sollten, angenommen und ausgehend von den tatsächlich beobachteten Daten die Parameter dieser Wahrscheinlichkeitsverteilung mithilfe von mathematischen Methoden ermittelt (McElreath, 2020; Murphy, 2022; Ng & Jordan, 2001; siehe auch ein Beispiel in Anhang A). Das Rasch-Modell (Abschnitt 2.3) ist in diesem Sinne ein Beispiel für ein probabilistisches Modell. Häufig können algorithmische Modelle auch ausgehend von probabilistischen Modellen hergeleitet werden (Bishop & Lasserre, 2007). Probabilistische Ansätze zur Beschreibung und Analyse von Daten haben den Vorteil, dass sie häufig unmittelbarer interpretierbar sind als vergleichbare klassische Ansätze und direkt Schätzungen für die Unsicherheit der Ergebnisse liefern. Sie werden mittlerweile auch in der Naturwissenschaftsdidaktik angewendet (z. B. Kubsch et al., 2021b).

Bisher wurden nun bereits zwei Beispiele für ML-Methoden genannt – Lineare Regression und Rasch-Modell – die man dem ML-Bereich vielleicht eher weniger zuordnen würde, wenn man eher einen Hintergrund in der Tradition der „klassischen“ Hypothesen-testenden Statistik hat. Der grundsätzliche Unterschied zwischen den Ansätzen der Hypothesen-testenden Statistik und ML-Methoden sind aber nicht die genutzten mathematischen Modelle – auch, wenn es durchaus Modelle gibt, die eher einem der beiden Ansätze zugeordnet werden. Der grundlegende Unterschied ist vielmehr die Herangehensweise und die Art der Ergebnisevaluation und -interpretation wie Breiman (2001) darstellt.

In der klassischen Hypothesen-testsenden Statistik („*Data Modelling*“ bei Breiman, 2001) ist das Ziel die Beschreibung von Phänomenen und Zusammenhängen durch mathematisch-theoretische Modelle, über die im Rahmen der schließenden Statistik Aussagen bzgl. ihrer Gültigkeit, Unsicherheit und Bedeutung getroffen werden können, z. B. mithilfe von

Signifikanzen und Effektstärken. Um solche Aussagen treffen zu können, beispielsweise durch die Berechnung von p-Werten, müssen die genutzten Modelle wahrscheinlichkeitstheoretisch wohldefiniert und händelbar sein.

Im ML-Ansatz („*Algorithmic Modelling*“ bei Breiman, 2001) ist das Ziel zwar auch die Modellierung von Phänomenen und Zusammenhängen, allerdings nicht, indem die Gültigkeit oder die Bedeutung der erhaltenen Modelle durch mathematisch-theoretische Sätze abgeleitet wird. Stattdessen ist die *Generalisierung der Modelle auf neue, ungesehene Daten* das zentrale Anliegen. Wenn der Zweck eines Modells also beispielsweise die Vorhersage des Studienerfolgs für Studienanfänger ist, dann wird das Modell daran bewertet, wie genau die Vorhersagen für Studienanfänger ist, deren Daten während des Lernprozesses des Modells (auch *Training* genannt, s. u.) nicht genutzt wurden. Um diese Generalisierung eines Modells einzuschätzen, also das Modell zu *evaluieren*, wird der für die Analyse verfügbare Datensatz in einen Trainings- und einen Evaluierungsdatensatz unterteilt (z. B. Géron, 2019). Das Modell wird dann mithilfe der Trainingsdaten erstellt und anschließend mithilfe der Evaluierungsdaten evaluiert. Die Vorhersagekraft oder Performanz des Modells für die Evaluierungsdaten, ggf. im Vergleich zur Performanz für die Trainingsdaten, ist dann das Gütekriterium anhand dessen das Modell bewertet wird. Die dazu nutzbaren Modelle müssen also auf mathematisch-theoretischer Ebene nicht so wohldefiniert und händelbar sein, wie Modelle, die in der schließenden Statistik genutzt werden. Sind sie es doch, ist das allerdings selbstverständlich auch kein Hindernis, sie trotzdem im Sinne eines ML-Ansatzes zu nutzen – das Ziel und die Herangehensweise sind das Entscheidende.

In den letzten Jahren haben sich diese beiden „Kulturen“, die Breiman (2001) versucht zu umreißen, allerdings deutlich angenähert und es existieren viele übergreifende Ansätze (z. B. Murphy, 2022). Trotzdem ist es hilfreich die grundsätzlichen Herangehensweisen zu kennen, um die Zielsetzungen konkreter Projekte und Ansätze besser einordnen und nachvollziehen zu können. Auch in der Naturwissenschaftsdidaktik finden solche eher algorithmischen Modellierungen bzw. ML-Methoden zunehmend Anwendung (z. B. Estrellado et al., 2020; Zhai et al., 2021b; Zhai et al., 2020b).

Loss-Funktionen und Training

Es wurden nun Ansätze, ML zu konzeptualisieren und von „klassischer“ Statistik abzugrenzen, dargestellt. Es ist allerdings noch nicht geklärt, *wie* die Modelle eigentlich aus den Trainingsdaten „lernen“. Um dies zu erläutern, müssen zunächst einige (wenige) Notationen eingeführt werden. Als *Modell* wird hier eine Funktion (im mathematischen Sinne) verstanden, die Inputs x auf Outputs $f_w(x)$ abbildet:

$$\text{Modell(funktion):} \quad f_w: X \to Y, \;\; x \mapsto f_w(x)\,.$$

Diese Funktion hängt von Parametern w ab. Für das „Lernen“ dieser Parameter der Funktion liegt ein *Datensatz* aus Features und Targets vor:

$$\mathcal{D} = \{(x_i, y_i \mid i = 1, \dots, N\}\,.$$

Die Features x und Targets y können dabei (ggf. mehrdimensionale) Zahlen, Kategorien o. Ä. sein. Das Ziel ist nun, die Parameter w so anzupassen, dass die Ausgaben der Modellfunktion

möglichst genau bei den jeweiligen Targets liegen, d. h., dass

$$f_{\hat{w}}(x_i) \approx y_i$$

gilt, wobei $\hat{w}$ die optimale Parameterwahl bezeichnet. Üblicherweise wird zu diesem Zweck eine sog. *Loss-Funktion* (auch einfach nur *Loss*) verwendet. Im Falle einer Regression ist die Loss-Funktion zum Beispiel typischerweise die Least-Squares-Funktion (z. B. Géron, 2019)

$$\mathcal{L}(w) = \sum_{i=1}^{N} (y_i - f_w(x_i))^2.$$

Für Klassifikationsmodelle, die anstelle eines kontinuierlichen einen diskreten Output bzw. diskrete Targets y haben, wird stattdessen zumeist die sog. Cross-Entropy Loss-Funktion verwendet (Géron, 2019). Im Unsupervised Learning hat typischerweise jeder Ansatz eine eigene Loss-Funktion. Loss-Funktionen sind häufig heuristisch motiviert, lassen sich aber häufig aus wahrscheinlichkeitstheoretischen Überlegungen herleiten (siehe ein Beispiel in Anhang A).

Das Ziel der Entwicklung eines ML-Modells lautet dann, die jeweilige Loss-Funktionen zu minimieren. Loss-Funktionen sind also stets so konstruiert, dass ihre Minimierung dem Ziel der Modellbildung entspricht. Den gesamten Prozess nennt man dann auch *Training*. Bei der oben exemplarisch dargestellten Least-Squares-Funktion wird also die Summe der quadratischen Abstände der Datenpunkte zur Ausgleichsgerade minimiert, sodass die Ausgleichsgerade „möglichst nah" an allen Datenpunkten liegt, was dem Ziel der Regression genau entspricht. Diese Optimierungen werden typischerweise mit dem sog. *Gradient Descent* Algorithmus vorgenommen (z. B. Géron, 2019). Dabei werden die Parameter iterativ gemäß der Vorschrift

$$w_{k+1} \leftarrow w_k - \alpha \frac{\partial \mathcal{L}(w)}{\partial w}$$

aktualisiert, sodass $\mathcal{L}(w)$ minimiert wird[9]. Der frei wählbare Parameter α wird dabei auch die *Learning Rate* genannt. In modernen Anwendungen werden aber zumeist Erweiterungen dieses Basisalgorithmus genutzt, wie beispielsweise der sog. Adam-Optimizer (Kingma & Ba, 2014). Diese Erweiterungen reagieren weniger sensibel auf suboptimale Wahlen der Learning Rate und sind zudem robuster bezüglich potenzieller lokaler Minima der Loss-Funktionen[10]. Bei Methoden des Unsupervised Learning und probabilistischen Ansätzen sind auch noch andere Optimierungsalgorithmen in Fällen gebräuchlich, in denen Gradient Descent-Varianten nicht angewendet werden können, wie beispielsweise die Expectation-Maximization- oder Variational-Inference-Methoden. Um diese darzustellen sind umfangreichere Beschreibungen auf Basis der Wahrscheinlichkeitstheorie notwendig, die hier aus Platzgründen nicht

[9] Bei „einfachen" Loss-Funktionen wie dem Least-Squares-Loss im Falle einer linearen Regression lässt sich das Optimum teilweise sogar noch analytisch bestimmen, indem $\partial \mathcal{L}(w)/\partial w = 0$ nach w aufgelöst wird. Das ist jedoch für komplexere Loss- bzw. Modell-Funktionen und große Datensätze nicht mehr möglich.

[10] Gerade im Deep Learning Bereich hat sich aber unter anderem Aufgrund der hohen Dimensionalität des Parameterraums gezeigt, dass lokale Minima ein geringeres Problem sind als intuitiv angenommen (z. B. Choromanska et al., 2015). Das liegt unter anderem daran, dass in hochdimensionalen Parameterräumen Nullstellen der Loss-Gradienten in den meisten Fällen lediglich Sattelpunkte sind.

vorgenommen werden können. Es sei daher auf einschlägige Literatur verwiesen (z. B. Murphy, 2022).

Batching und Learning Curves

Mit dem Modell $f_w(x)$, einem geeigneten Loss $\mathcal{L}(w)$ und einem Optimierungsalgorithmus wie Gradient Descent sind somit alle Bausteine für das Training eines ML-Modells vorhanden. In modernen Anwendungen gibt es allerdings noch eine weitere Hürde: Die Menge an Trainingsdaten ist meist zu groß, als dass alle Datenpunkte in jeder einzelnen Iteration der Optimierung verwendet werden können. In diesen Fällen geht man dazu über, den Datensatz randomisiert in kleinere Segmente, sog. *Batches*, zu unterteilen und diese dann nacheinander in den Iterationen zu nutzen (z. B. Géron, 2019). Im Falle des Gradient Descent Algorithmus nennt man dieses Vorgehen aufgrund der Zufälligkeit der Zuordnung zu den Batches auch *Stochastic* Grandient Descent. Die Optimierung in Batches hat sich zum de facto Standard entwickelt, weshalb bei den meisten anderen Optimierungsalgorithmen der Zusatz „Stochastic" gar nicht genutzt wird. Das batchweise Training kann zwar etwas instabiler sein, ist aber üblicherweise deutlich schneller und ab einem gewissen Verhältnis zwischen verfügbarer Rechenleistung und Trainingsdaten unvermeidlich.

Sind einmal alle Daten des Datensatzes (möglicherweise batchweise) durchlaufen worden, so spricht man auch von einer *Epoch* absolviertem Training. Je nach Modell wird üblicherweise nur für wenige Epochs oder aber auch mehrere hundert Epochs trainiert. Um zu erkennen, ob die Optimierung konvergiert, werden üblicherweise während oder nach dem Training sog. *Learning Curves* erstellt (Géron, 2019). Dabei wird der jeweilige (Batch-)Loss gegen die absolvierten Trainingsschritte oder Epochs aufgetragen. Um bereits während des Trainings zu überwachen, wie gut das Modell auf ungesehene Daten generalisierbar ist, kann hier auch der Loss für die Evaluierungsdaten mit aufgetragen werden, allerdings ohne für diese Loss-Berechnung auch einen Optimierungsschritt durchzuführen. Eine beispielhafte Learning Curve ist in Abbildung 2.5 dargestellt. Ergänzend zum Loss kann man vor allem bei Klassifikationsmodellen auch leichter interpretierbare Metriken wie die prozentuale Übereinstimmung (*Accuracy*) oder Cohens κ ergänzend auftragen.

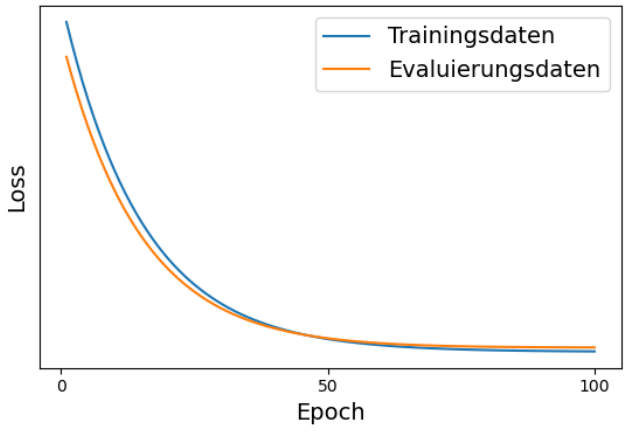


Abbildung 2.5 Beispielhafte prototypische Learning Curves. Hier sind absichtlich keine Werte für den Loss auf der y-Achse angegeben, da diese häufig nicht absolut interpretierbar sind.

Es ist hilfreich, während des Trainings die Learning Curves bezüglich der Evaluierungsdaten zu betrachten, um bereits im Prozess zu überwachen, ob das Modell auf ungesehene Daten generalisierbar ist. Das ist allerdings nicht unproblematisch. Es gibt viele Modelle, in denen einige Parameter bereits vor dem eigentlichen Training gewählt werden müssen (z. B. Regularisierungsterme, s. u.) und auch das Training selbst ist von Parametern wie der Learning Rate oder Größe der Batches abhängig. Solche Parameter werden *Hyperparameter* genannt (Géron, 2019). Nutzt man die Evaluierungsdaten bereits während des Trainings, so kann man sich nicht (ganz) sicher sein, ob die Performanz bezüglich der Evaluierungsdaten für eine andere Wahl von Hyperparametern nicht anders (geringer) sein könnte. Das heißt, man kann unbeabsichtigt die Hyperparameter manuell so optimieren, dass eine höhere Performanz bezüglich des Evaluierungsdatensatzes erreicht wird, die für „tatsächlich ungesehene" Daten nicht vollständig repräsentativ ist. Es gibt im Wesentlichen zwei Möglichkeiten, um dem zu begegnen. Der erste Ansatz wäre, noch einen dritten Datensatz aus den Gesamtdaten abzuspalten, mit dessen Hilfe als vollständig unangetasteter *Test*-Datensatz am Ende des Trainings das Modell erneut evaluiert wird. Sind nicht genügend Daten vorhanden, um eine solche weitere Unterteilung durchführen zu können, bietet sich das Verfahren der sog. *Cross-Validierung* (CV) an (z. B. Géron, 2019). Dazu wird der Datensatz in k bis auf Rundung gleich große Segmente unterteilt. Nun wird das Modell k-mal trainiert, wobei jeweils eines dieser Segmente zur Evaluierung zurückgehalten wird. Die Evaluierung erfolgt somit einmal auf Basis aller verfügbarer Daten und ist somit deutlich robuster gegenüber Schwankungen und es ist unwahrscheinlicher hier durch Hyperparameter-optimierung tatsächlich nicht-repräsentative Performanzzuwächse zu erzeugen (siehe auch Kapitel 6). Selbstverständlich kann man auch CV nutzen und trotzdem zusätzlich noch mithilfe eines vollständig separatem Test-Datensatz arbeiten.

Overfitting

Nachdem der Grundansatz und die Grundmethodik von ML-basierten Analysen vorgestellt wurden, soll nun noch ein zentrales Phänomen beschrieben werden, welches auch für das hier vorgestellt Projekt relevant ist. Komplexe ML-Modelle, die über eine große Anzahl trainierbarer Parameter verfügen, sind häufig in der Lage, sämtliche Spezifika eines Trainingsdatensatzes zu „erlernen". Dieses Phänomen ist als *Overfitting* (z. B. Géron, 2019) bekannt und zeigt sich in einem deutlichen Unterschied zwischen dem Loss bezüglich der Trainings- und Evaluierungsdaten. Bei einem Regressionsmodell, d. h., wenn die Target-Variable kontinuierlich ist, lässt sich dies leicht mithilfe eines entsprechenden Kurven-Fits verdeutlichen (Abbildung 2.6). Auch bei Klassifikationsproblemen lässt sich Overfitting zumindest im Falle von zweidimensionalen Features noch gut visualisieren (Abbildung 2.7, links). Man beachte, dass es sich bei allen Abbildungen in diesem Abschnitt um reale Modellfits handelt. Es gibt einige sog. Regularisierungsmethoden (z. B. Géron, 2019; Murphy, 2022), die man nutzen kann, um Overfitting zu verringern, von denen eine in Abbildung 2.7 (rechts) angewendet wurde.

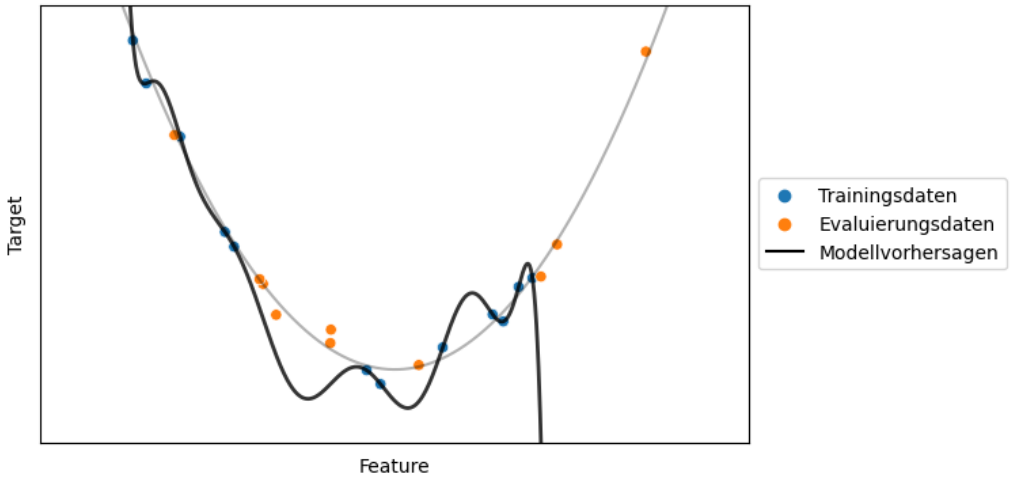


Abbildung 2.6 Darstellstellung von Overfitting bei einem Regressionsproblem. Man erkennt, wie die Modellfunktion (schwarze Linie) alle Trainingsdatenpunkte exakt trifft, aber jeden Evaluierungsdatenpunkt verfehlt. Die den Daten zugrundeliegende Funktion ist eine quadratische Funktion, die in hellgrau dargestellt ist. Die Daten folgen der Vorschrift $Y = X^2 + \epsilon$, wobei ϵ normalverteilt ist.

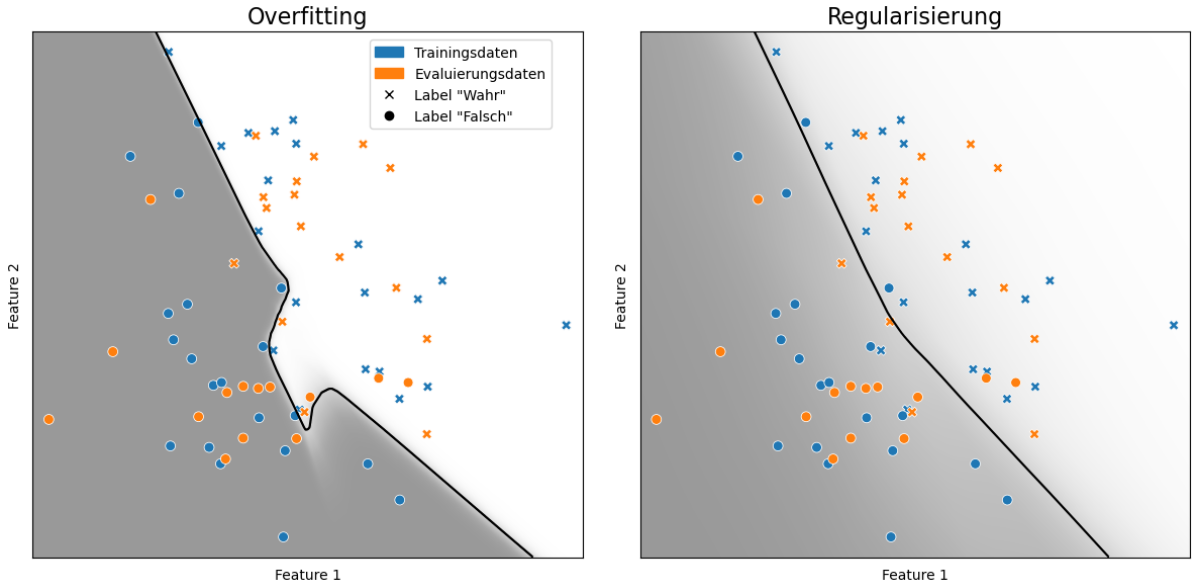


Abbildung 2.7 Darstellung von Overfitting und Regularisierung bei einem Klassifikationsproblem. Man erkennt, wie die sog. *Decision Boundary*, d. h. die Grenzlinie, ab der das Modell zwischen den Zuordnungen wechselt, sich beim linken Modell förmlich um einzelne Trainingsdaten „herumlegt“. Die Einfärbung zeigt die Wahrscheinlichkeit, dass das Modell die Zuordnung „wahr“ vornimmt. Je heller ein Bereich hinterlegt ist, umso wahrscheinlicher ist es, dass das Modell einen dort liegenden Datenpunkt als „wahr“ klassifiziert. Die Datenpunkte wurden gleichverteilt generiert und die tatsächliche Zuordnung wurde gemäß: $x_1 + x_2 > 0$ vorgenommen. Allerdings wurde anschließend normalverteiltes Rauschen zu x_1 und x_2 addiert, um einen realen Datensatz zu simulieren und das Overfitting deutlicher hervorzuheben. Rechts wurde die sog. L_2-Regularisierung (auch „weight decay“, siehe z. B. Géron, 2019) angewendet.

Wenn die Feature- oder Target-Variablen allerdings höherdimensional sind, kann die Vorhersage nicht mehr so leicht visualisiert werden. Man greift stattdessen häufig auf die Learning Curves zurück, um einzuschätzen, ob Overfitting vorliegt (Géron, 2019). Ein sicheres Indiz für das Vorliegen von Overfitting ist, dass der Evaluierungsloss ab einem bestimmten

Trainingsschritt anfängt anzuwachsen, wie in Abbildung 2.8 links zu sehen. Das heißt, das Modell erlernt aktiv Spezifika des Trainingsdatensatzes und generalisiert immer schlechter auf die ungesehenen Evaluierungsdaten. In Abbildung 2.8 rechts dargestellt ist der Fall, indem zwar auch ein deutlicher Unterschied zwischen dem Loss bezüglich Evaluierungs- und Trainingsdaten sichtbar ist, aber nicht unbedingt Overfitting vorliegt. Hier ist auch möglich, dass der Trainingsdatensatz schlicht nicht die gesamte Varianz abbildet, die in der Grundgesamtheit vorhanden ist.

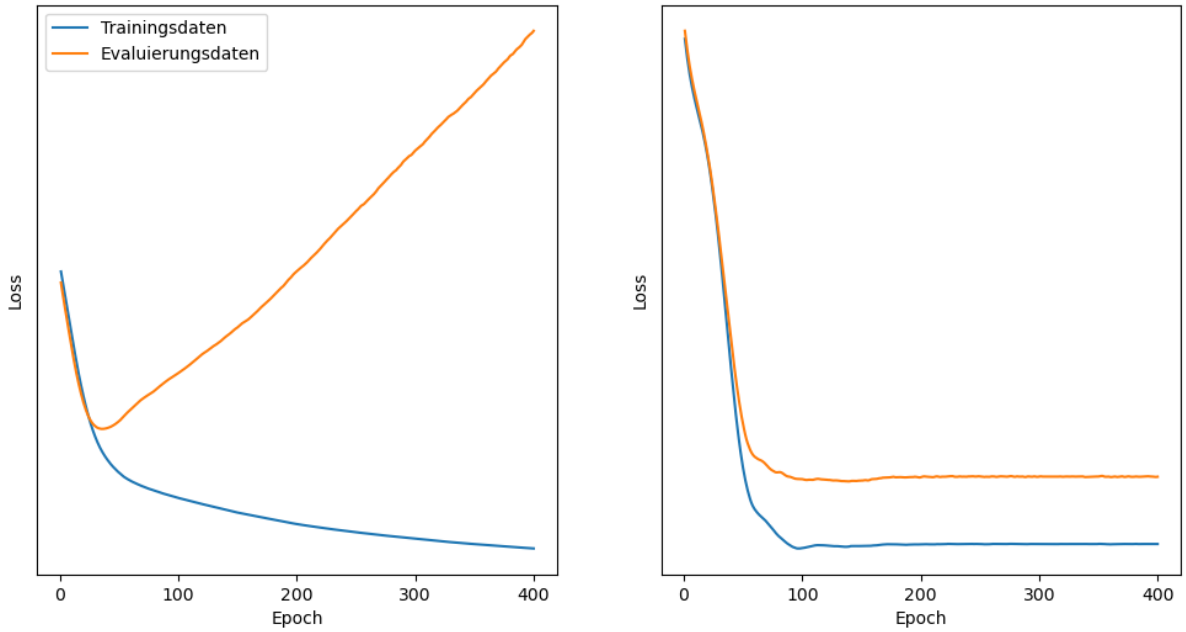


Abbildung 2.8 Overfitting sichtbar in Learning Curves. Dargestellt sind die zu den Modellen, deren Decision Boundaries in Abbildung 2.7 dargestellt sind, gehörigen Learning Curves.

Neben dem Overfitting existiert auch das gegenteilige Phänomen des *Underfitting*. Underfitting liegt vor, wenn das Modell an sich schon zur Beschreibung des Datensatzes nicht ausreichend variieren kann. Dieses Problem würde beispielsweise vorliegen, wenn man an die in Abbildung 2.6 dargestellten Daten eine lineare Funktion anpassen würde. Für die Sprachmodelle (siehe Abschnitt 2.6), die in dem hier vorgestellten Projekt die größte Rolle spielen, ist Underfitting bei der verfügbaren Anzahl an Modellparametern (mehrere 10 Mio. bis mehrere Mrd.) unwahrscheinlich. Bei den anderen hier verwendeten, einfacheren Modellen, wie linearen und logistischen Regressionsmodellen, könnte Underfitting zwar eine Rolle spielen, die Evaluierungen zeigten hier jedoch keine Problematik. Insgesamt wird daher hier auf ausführlichere Visualisierungen u. Ä. zum Underfitting verzichtet.

2.5. Machine-Learning-Rahmenmodelle für naturwissenschaftsdidaktische Forschung

Es wurden nun einige grundlegende Begrifflichkeiten, Workflows und Ansätze aus dem Bereich des ML vorgestellt. Diese dienen einerseits zur Vorbereitung der explorativen, nicht-hierarchischen Analysen der inneren Struktur des FDW im zweiten Zielpaket dieses Projekts

(Kapitel 5). Andererseits sind sie aber auch die Grundlage für die Automatisierung der Auswertung des genutzten Testinstruments auf Basis der vorherigen Ergebnisse im Rahmen des dritten Zielpakets (Kapitel 6). In diesem Abschnitt werden nun Ansätze erläutert, die zur Strukturierung insbesondere der explorativen Analyse der inneren Struktur des FDW herangezogen werden. Erst im nächsten Abschnitt (2.6) werden dann einige weitere ML- und NLP-Methoden vertieft thematisiert.

ML-Methoden und insbesondere Methoden des Unsupervised Learning in sozial- und bildungswissenschaftlichen Forschungsvorhaben anzuwenden, bringt Herausforderungen sowohl in der technischen Umsetzung als auch bei der Interpretation und Deutung der Ergebnisse mit sich (z. B. Nelson, 2020; Zhai et al., 2020b). Zhai et al. (2020b) arbeiten im Rahmen ihres systematischen Reviews bestehender ML-Anwendungen im Kontext des Assessments im naturwissenschaftlichen Bereich heraus, dass ein Großteil bisheriger ML-Anwendungen primär zur Unterstützung und Entlastung menschlicher Rater bei basalen Aufgaben dient. Zhai et al. (2020a) entwickelten parallel ein Framework, welches die Anwendung von ML auf einem Kontinuum zwischen reinem „Ersetzen“ (*Substitution*) bis hin zu echter „Transformation“ (*Redefinition*) von Assessmentprozessen systematisiert. Zhai (2021) hebt darüber hinaus die Potenziale der Anwendung von ML-Methoden für Assessmentzwecke noch einmal explizit hervor.

Unsupervised-Learning-Methoden sind gerade für Erkenntnisgewinnung und echte Transformationen von Assessmentprozessen interessant, da sie (anders als Supervised-Learning-Methoden) dazu in der Lage sind, neue bzw. bisher unerkannte Strukturen in den untersuchten Konstrukten sichtbar zu machen (Nelson, 2020; Zhai et al., 2020b). Besonders bei Unsupervised-Learning-Methoden gestalten sich Interpretation und Deutung der Analyseergebnisse aber häufig als komplex. Sherin (2013) schlug daher schon früh vor, die Interpretationskraft und Sachkenntnis menschlicher Experten direkt in explorative Analyseprozesse mit einzubinden. Ein prominenter Ansatz, um diese Verschränkung von menschlicher Expertise und computergestützter Modellierung zu systematisieren und organisieren, ist die sog. Computational Grounded Theory (CGT) nach Nelson (2020). Die CGT wird im vorgestellten Projekt intensiv zur Konzeption der Untersuchungen zur nicht-hierarchischen Struktur des FDW, d. h. im Rahmen des zweiten Zielpakets, genutzt (Kapitel 5). Sie dient insbesondere zur Strukturierung explorativer Analysen unter der Nutzung von Unsupervised-Learning-Methoden und hat das Ziel, die Interpretation der Ergebnisse zu erleichtern und ihre Verlässlichkeit zu erhöhen. Nelson (2020) schlägt dafür die folgenden drei Schritte vor:

1. *Pattern Detection*: Explorative Methoden werden zur Identifikation von neuen Mustern und Strukturen in den Daten genutzt. Im Falle von Daten zu psychometrischen Testinstrumenten können das Clusteranalysen der Scores sein. Im Falle von Interview- oder Freitextdaten können explorative Textanalysemethoden (siehe Abschnitt 2.6) angewendet werden.
2. *Pattern Refinement*: Die identifizierten Muster werden durch Tiefenanalysen ausgeschärft. Dabei fließen menschliches Expertenwissen und Interpretationskraft in die Analyse ein. Im Falle von psychometrischen Testinstrumenten können dabei

beispielsweise Informationen über die Subskalen des Testinstruments genutzt werden. Im Falle von Sprachdaten können Zusammenhänge zwischen Sprachnutzung und Kovariaten untersucht werden oder besonders charakteristische Texte einer erneuten manuellen Untersuchung unterzogen werden.

3. *Pattern Confirmation*: Um ein Argument für die Stabilität und in diesem Sinne auch Validität der identifizierten Muster und Strukturen zu bieten, wird die Vorhersagekraft von ML-Modellen bei der Klassifizierung der zuvor ermittelten Kategorien evaluiert. Im Falle von psychometrischen Testinstrumenten können typische Klassifikationsmodelle genutzt werden, um Proband:innen anhand ihrer Scores den Kategorien zuzuordnen. Im Falle von Sprachdaten können für denselben Zweck NLP-Modelle unterschiedlicher Komplexität (siehe Abschnitt 2.6) genutzt werden. Die Pattern Confirmation dient zur Bestätigung der in der Pattern Detection gefundenen Muster in folgendem Sinne: Eine (ausreichend) hohe Performanz von ML-Modellen bei Verortung von Datensätzen im Rahmen der gefundenen Muster (z. B. Zuordnung zu Clustern) dient als Nachweis der Existenz latenter Strukturen in den Daten, die mit diesen Mustern korrespondieren. ML-Modelle können hierbei (anders als viele „klassische“ Verfahren wie lineare Regressionsmodelle oder Strukturgleichungsmodelle, siehe Moosbrugger & Kelava, 2020) auch nicht-lineare Zusammenhänge modellieren. Der Nachteil ist, dass diese latenten Strukturen dann nicht unbedingt greifbar sind. Trotzdem liefert eine erfolgreiche Pattern Confirmation somit ein Argument für die Robustheit, (bei der Nutzung von Evaluierungsdaten auch) die Generalisierbarkeit und die Validität der beschriebenen Muster. Welche Validitätsaspekte (z. B. Messick, 1995; Schaper, 2014) dabei adressiert werden, hängt von der Beziehung zwischen Feature- und Target-Daten der Pattern-Confirmation-Modelle ab (siehe auch Ende Abschnitt 6.7.3).

Auch, wenn die bei Nelson (2020) beschriebene Form der CGT stark auf die Analyse von Textdaten ausgerichtet ist, lässt sich das Verfahren auch auf andere Datentypen bzw. mehrere Datenquellen übertragen. In diesem Fall werden dann insbesondere die Pattern Detection und das Pattern Refinement als ein iterativer Prozess verstanden, in dem menschliches Expertenwissen an unterschiedlichen Stellen im Analyseprozess genutzt werden kann. Die CGT bietet somit eine Möglichkeit, Unsupervised-Learning-Ansätze für eine echte Transformation von Assessmentprozessen im Sinne von Zhai et al. (2020a) zu nutzen, indem die ermittelten Kategorien als Zielkonstrukte bzw. Targets eines Assessment-Systems genutzt werden. So kann sowohl neu ermitteltes als auch bestehendes Wissen über die innere Struktur der betrachteten Konstrukte direkt im Assessmentprozess genutzt werden. Die CGT wurde in der naturwissenschaftsdidaktischen Forschung bereits zur Untersuchung von Erklärprozessen (Rosenberg & Krist, 2021) und Argumentationsmustern von Schülerinnen und Schülern (Tschisgale et al., 2023) erfolgreich eingesetzt.

Die CGT bietet zudem einen Rahmen, den Daten-Mix, der für die Analyse der inneren Struktur des FDW in diesem Projekt vorliegt, gesamtheitlich in den Blick zu nehmen. Dabei werden insbesondere die zuvor durch trainierte Kodierer manuell erstellten Scores für die Pattern Detection und die authentischen Sprachproduktionen der Proband:innen in den offenen Aufgaben des Testinstruments im Pattern Refinement genutzt. Das genaue Vorgehen und die

genaue Anwendung der CGT bei den Analysen zu diesem zweiten Zielpaket des Projekts werden in den Kapiteln 5 und 6 genauer vorgestellt.

Darüber hinaus haben Kubsch et al. (2022) das *Distributing Epistemic Functions and Tasks* (DEFT)-Framework vorgeschlagen[11], mit dem sie die Anwendung von ML- und Data-Science-Methoden nicht nur für Assessmentzwecke, sondern für sozial- und bildungswissenschaftliche Forschungsvorhaben im Allgemeinen systematisieren. Sie schlagen dafür eine zweidimensionale Strukturierung vor, wobei auf der ersten Achse zwischen *Supervised Settings* (Targets bzw. Label sind im Voraus bekannt) und *Unsupervised* bzw. *Grounded Settings* (Targets bzw. Label sind nicht vordefiniert) unterschieden wird. Damit folgen sie der grundlegenden Unterscheidung zwischen Supervised und Unsupervised Learning, schließen aber „Mischformen" nicht aus. In der hier vorliegenden Untersuchung der inneren Struktur des FDW sind beispielsweise Scores aus der Bepunktung des genutzten Testinstruments als „Labels" vorhanden, allerdings (noch) nicht die tatsächlichen Zielkonstrukte zur inneren Struktur des FDW. Es kann also im Sinne des DEFT-Frameworks eher von einem Grounded Setting gesprochen werden. Auf der zweiten Achse des Frameworks unterscheiden Kubsch et al. (2022) zwischen sog. *High Inference* und *Low Inference*. In Low-Inference-Settings werden einfache Konstrukte bzw. unmittelbar zugängliche Kategorien, wie die reine Bepunktung eines Testinstruments, in den Blick genommen, während in High-Inference-Settings komplexere Konstrukte wie Kompetenzprofile untersucht werden.

In diesem Abschnitt wurden die CGT und das DEFT-Framework zur Strukturierung von ML-basierten Analysen in (u. A.) naturwissenschaftsdidaktischer Forschung dargestellt. Dabei wurde insbesondere die Anwendung der CGT für die Analysen im Rahmen des zweiten und dritten Zielpakets des vorliegenden Projekts bereits angedeutet. Zur Analyse des vorliegenden Daten-Mix aus Scores und Sprachproduktionen in den offenen Aufgaben des verwendeten Testinstruments werden dabei explorative Sprachanalysemethoden, sog. Topic Models (Blei, 2012), genutzt. Zur Automatisierung der Auswertung des Testinstruments im Rahmen des dritten Zielpakets werden Deep-Learning-Sprachmodelle (z. B. Devlin et al., 2019) verwendet. Aufbauend auf den ML-Grundlagen aus Abschnitt 2.4 werden daher im folgenden Abschnitt weitere Begriffe und Methoden aus dem Bereich des Deep Learning und der Sprachanalyse vorgestellt. Insbesondere die Deep-Learning-basierte Sprachanalyse ist eine zentrale Grundlage für das vorgestellte Projekt und ermöglicht die angestrebte Automatisierung des FDW-Assessments.

2.6. Machine-Learning-basierte Sprachanalyse

In Abschnitt 2.4 wurden einige ML-Grundbegriffe eingeführt und in Abschnitt 2.5 wurde insbesondere die CGT zur Strukturierung explorativer Analysen im (u. A.) bildungswissenschaftlichen Kontext vorgestellt. Nun werden aufbauend auf Abschnitt 2.4 weitere ML-(basierte) Methoden beschrieben, die zur Analyse der authentischen Sprachprodukte (Antworten auf die offenen Testaufgaben) von Proband:innen zu den

[11] Für eine verwandte aber anders ausgerichtete Systematisierung sei hier auch auf die jüngst erschienenen Arbeiten von Nehring et al. (2025) verwiesen.

Aufgaben des für dieses Projekt verwendeten FDW-Testinstruments genutzt werden (Kapitel 5 & 6).

Encodings

Natural Language Processing (NLP) kann als das Teilgebiet der Data Science aufgefasst werden, das sich mit der (computerbasierten bzw. automatisierten) Verarbeitung menschlicher Sprache befasst (Jurafsky & Martin, 2024). Frühe NLP-Methoden basierten teilweise auf expliziten Regel-Systemen („wenn-dann“) oder Word-Count Tabellen (z. B. TF-IDF, Mladenić et al., 2016). Moderne Methoden umfassen unter anderem probabilistische Ansätze (z. B. Blei, 2012; Roberts et al., 2019) und Deep-Learning-basierte Sprachmodelle, die *Language Models* (LM) oder *Large Language Models* (LLM) genannt werden (z. B. Devlin et al., 2019; Übersicht bei Naveed et al., 2024). Der Grundansatz von NLP-Methoden besteht darin, die in den einzelnen *Dokumenten* des Datensatzes auftretenden Worte systematisiert zu erfassen und in eine mathematische Repräsentation zu überführen (Jurafsky & Martin, 2024). Die Gesamtheit der in den Dokumenten des Datensatzes auftauchenden Worte wird auch *Vokabular* genannt. Jedem Wort des Vokabulars wird typischerweise zunächst eine natürliche Zahl als Index zugeordnet. Da menschlicher Sprache aber keine hierarchische Dimension innewohnt, nach der man die Worte sinnvoll ordnen könnte, muss diese Indexdarstellung in eine Darstellung übertragen werden, die keine Reihenfolge der Worte impliziert. Üblicherweise werden dazu in erster Instanz sog. *One-Hot-Encodings* verwendet, d. h. die Worte werden als Vektoren dargestellt, die nur an ihrem jeweiligen Index „1“ und sonst überall „0“ sind. Diese Vektoren müssen allerdings so viele Einträge haben, wie Worte im Vokabular sind, und sind dadurch üblicherweise hochdimensional (mehrere tausend Dimensionen). Die Überführung von Dokumenten in One-Hot-Encodings ist in Abbildung 2.9 schematisch dargestellt. Dabei werden zwei „Dokumente“ (einzelne Sätze) verarbeitet. Das Vokabular enthält alle Worte im Korpus der Dokumente und wird genutzt, um die Dokumente in die Indexdarstellung, bei der jedes Wort durch den entsprechenden Index abgebildet wird, zu überführen. Die One-Hot-Encodings können dann genutzt werden, um weitere Repräsentationen des Textes abzuleiten.

Das Verarbeiten solcher One-Hot-Encodings ist umständlich und ineffizient, weshalb unterschiedliche Methoden genutzt werden, um diese Darstellungen weiter zu reduzieren bzw. zu verdichten (z. B. Géron, 2019 Géron, 2019; Jurafsky & Martin, 2024). Ein Standard-Ansatz ist der sog. *Bag-of-Words*-Ansatz (BoW-Ansatz), indem die Reihenfolge der Worte in den Dokumenten ignoriert wird und die Dokumente durch die Summierung der Encodings ihrer jeweiligen Worte repräsentiert werden (siehe Abbildung 2.9). Um die Dimensionalität dieser Vektoren weiter zu verringern, werden zudem typischerweise einige Schritte zur Reduktion des Vokabulars genutzt, darunter

- *Lowercasing*: Alle Worte werden klein geschrieben.
- *Puncutation-Removal*: Satzzeichen u. Ä. werden vernachlässigt.
- *Stopword-Removal*: Worte, die in der vorliegenden Sprache der Dokumente sehr häufig auftauchen, aber wenig inhaltliche Bedeutung tragen, z. B. „und“, „der“, „die“, etc.,

werden vernachlässigt. Für solche Stopwords existieren in Sprachverarbeitungssoftware (z. B. Bird et al., 2009[12]) typischerweise vordefinierte Listen.

- *Word-Frequency*: Worte, die in einem sehr großen oder sehr geringen Anteil an Dokumenten auftauchen, werden vernachlässigt. Ähnlich wie beim Stopword-Removal tragen diese Worte oft nur geringe inhaltliche Bedeutung für die Analyse, entweder, weil sie die Dokumente nicht voneinander unterscheiden, oder, weil sie so selten auftreten, dass die Bedeutung ihrer An- oder Abwesenheit in Dokumenten nicht (ausreichend) systematisch beschrieben werden kann.
- *Tokenization*: Die Texte werden in sog. *Token* unterteilt. Dabei ist ein Token ein Wort oder Teil eines Wortes. Moderne Sprachverarbeitungsmodelle nutzen nicht-triviale Methoden, um nützliche Unterteilungen in Token vorzunehmen (z. B. Mistral AI, o. D.). Im Schnitt entspricht ein Token ca. ¾ eines Wortes.

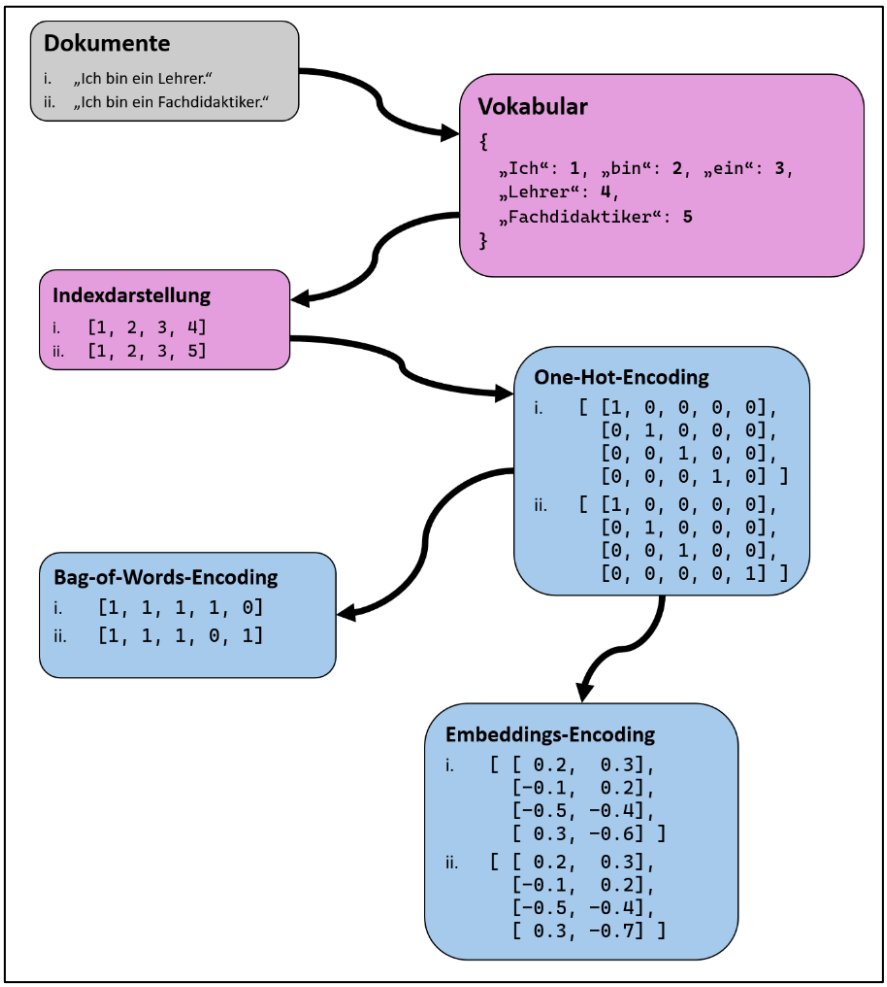


Abbildung 2.9 Darstellung der Überführung von Dokumenten in unterschiedliche Encodings. Die Berechnung von Embeddings-Encodings (unten rechts), wird am Ende des Abschnitts 2.7 beschrieben.

[12] Auch, wenn diese Quelle schon etwas älter ist, wird das dort eingeführte Python-Paket „Natural Language Toolkit (NLTK)" (https://www.nltk.org/, zugegriffen 17. Januar 2025) nach wie vor aktiv erweitert und genutzt.

Topic Models

Mithilfe der BoW-Darstellung von Texten lassen sich bereits eine Reihe von Analysen durchführen. Einerseits können einfache Klassifikationsmodelle wie logistische Regressionsmodelle auf Basis dieser Darstellung zur Vorhersage bestimmter Labels zu den Dokumenten trainiert werden (z. B. Géron, 2019). Andererseits können Unsupervised-Learning-Methoden zur Untersuchung von Mustern im Sprachgebrauch der Dokumente angewendet werden. Für solche explorativen Untersuchungen werden häufig sog. *Topic Models* verwendet (Chen & Liu, 2017). Dabei handelt es sich um probabilistische BoW-Modelle, die die Wahrscheinlichkeit des Auftretens der Wörter in den Dokumenten über den Zwischenschritt der sog. *Topics* modellieren. Das ursprüngliche Topic Model (Blei, 2012), oder auch *Latent Dirichlet Allocation*-Modell (LDA, Blei et al., 2003), nimmt dabei folgenden Prozess für die Erzeugung der Dokumente an:

1. Jedes Dokument beschäftigt sich anteilig mit jedem Topic[13].
2. Jedes Topic besitzt eine Wahrscheinlichkeit für jedes Wort des Vokabulars. Der englische Begriff *Topic* wird hier als ein feststehender Begriff des Topic Modelling genutzt, um Verwechslungen zu vermeiden. Trotzdem lassen sich die Topics als „latente Themen", um die es in den Dokumenten geht, verstehen.
3. Jeder „Wort-Platz" in einem Dokument wird mit der Dokument-Topic-Verteilung („1.") einem Topic zugeordnet.
4. Für jeden dieser „Wort-Plätze" wird gemäß der jeweiligen Topic-Wort-Verteilung („2.") ein Wort generiert.

Dieser Prozess lässt sich mathematisch durch eine Wahrscheinlichkeitsverteilung modellieren (Blei et al., 2003). Es wird dann ein Schätzalgorithmus konstruiert, der ausgehend von den tatsächlichen Dokumenten bzw. deren Worten im Datensatz die wahrscheinlichsten Dokument-Topic- und Topic-Wort-Verteilungen berechnet. Die Details dieses Vorgehens lassen sich nicht ohne umfangreiche Vorarbeiten im Bereich der Wahrscheinlichkeitstheorie bzw. des probabilistischen MLs darstellen, weshalb hier auf entsprechende Literatur verwiesen wird (Blei, 2012; Blei et al., 2003; Murphy, 2022). Das Ergebnis der Analyse sind dann die folgenden Werte[14]:

- $\theta_{dk} \approx$ Der Anteil des Dokuments d der dem Topic k gewidmet ist. Ist θ_{dk} groß, ist das Dokument d also stark auf das Topic k fokussiert.
- $\varphi_{vk} \approx$ Die Wahrscheinlichkeit, dass das Wort v in Topic k auftritt. Ist φ_{vk} groß, spielt das Wort v also eine zentrale Rolle in Topic k.

[13] Dabei wird angenommen, dass diese Dokument-Topic-Verteilung eine Dirichlet-Verteilung (z. B. Murphy, 2022) ist. Dasselbe gilt für die Topic-Wort-Verteilung. Daher stammt auch die Bezeichnung Latent *Dirichlet* Allocation.

[14] Es sind $d = 1 \dots M$, $k = 1 \dots K$ und $v = 1 \dots V$, wobei M die Anzahl an Dokumenten, K die Anzahl an Topics und V die Größe des Vokabulars ist.

Die Topic-Wort-Anteile φ_{vk} dienen dann zur inhaltlichen Beschreibung der bis dahin inhaltlich nicht näher charakterisierten Topics. Dabei werden meist die wahrscheinlichsten Worte der Topics genutzt, um die Topics zu interpretieren.

Um Wissen über Kovariaten im Topic Model zu berücksichtigen, gibt es einige Erweiterungen des LDA-Ansatzes (Blei & Lafferty, 2005; Hennig et al., 2012; Roberts et al., 2016; Roberts et al., 2019). Solche erweiterten Modelle können insbesondere zusätzliche Informationen über die Dokumente, wie z. B. Autoren oder Entstehungszeit berücksichtigen. Dafür werden im angenommenen probabilistischen Prozess, der die Dokumente erzeugt (s. o.), entsprechende Zwischenschritte eingefügt. Die Schätzalgorithmen werden dadurch allerdings ebenfalls komplexer. In den Analysen zum zweiten Zielpaket dieses Projekts werden zunächst die Bearbeitungen des FDW-Testinstruments über eine Cluster-Analyse der Scores einer von vier Gruppen zugeordnet. In der explorativen Analyse der Sprachproduktionen der Proband:innen zur Ausschärfung der Beschreibung dieser Cluster (hin zu Kompetenzprofilen) werden dann die Cluster der jeweiligen Proband:innen als Kovariaten aufgefasst. Ein Dokument (im NLP-Sinne) sind dann alle Antworten, die eine Person in einer Bearbeitung des Testinstruments niedergeschrieben hat, zusammengenommen. Die Kovariate ist das Cluster, dem diese Person zugeordnet ist. In dieser Konfiguration wird im Rahmen der Analysen zum zweiten Zielpaket dann ein sog. *Structural Topic Model* (STM, Roberts et al., 2019) erstellt, bei dem die Clusterzuordnung einen Einfluss auf die Dokument-Topic-Verteilung haben kann. Zusätzlich zu den Dokument-Topic-Anteilen und den Topic-Wort-Verteilungen erhält man bei einem STM als zusätzliches Ergebnis Schätzwerte über die Stärke der Zusammenhänge zwischen den Topics und den Score-Clustern. Das genaue Vorgehen und die Ergebnisse zu dieser Untersuchung sind in Kapitel 5 dargestellt.

2.7. Deep-Learning-basierte Sprachanalyse

Neben den beschriebenen „klassischen“ Machine-Learning basierten Methoden zur Sprachanalyse haben sich (auch für die naturwissenschaftsdidaktische Forschung und Entwicklung) in den letzten Jahren vor allem Deep-Learning basierte Sprachmodelle als vielversprechender Ansatz erwiesen (z. B. Camus & Filighera, 2020; Zhai et al., 2020b; sowie Wulff et al., 2023 und darauf aufbauend Mientus et al., 2023). Im Folgenden werden daher einige Begriffe und Konzepte des Deep Learning mit besonderem Fokus auf Sprachmodellierung eingeführt. Die angesprochenen Modelle und Methoden stellen vor allem die Basis des automatisierten Assessment-Systems (Kapitel 6) dar.

Deep Learning

Unter *Deep Learning* wird letztlich „normales“ Machine Learning mit einer bestimmten, sehr umfangreichen Klasse von Modellen verstanden (Géron, 2019; siehe auch Abbildung 2.4). Diese Modellklasse wird allgemein als *Neural Network* (NN) bezeichnet und ist dadurch charakterisiert, dass schichtweise mathematische Operationen in sog. *Layern* hintereinander geschachtelt werden. In einem einfachen sog. *Fully-Connected-NN* (FCNN, Abbildung 2.10) sind diese Layer Matrixmultiplikationen. Die jeweiligen Outputs der einzelnen Layer sind in ihrer Dimensionalität variabel und werden auch *Nodes* genannt. In einem FCNN sind alle

Nodes einer Layer mit allen Nodes der vorherigen und folgenden Layer verbunden. Die Berechnung der Outputs der ersten Layer des in Abbildung 2.10 dargestellten Netzwerkes lautet also:

$$y^{(1)} = g\big(W^{(1)}x\big), \qquad W^{(1)} \in \mathbb{R}^{4\times 3}.$$

Dabei ist g eine nicht-lineare Funktion, die Element-weise auf die Matrix-Vektor-Produkte $W^{(1)}x$ angewendet wird und auch *Activation Function* oder kurz *Activation* genannt wird. Jedes Matrix-Element ist in Abbildung 2.10 durch einen Pfeil visualisiert. Die Matrix-Elemente sind dabei die trainierbaren Parameter des Modells (siehe Abschnitt 2.4). Die gesamte Modellfunktion des in Abbildung 2.10 dargestellten Netzwerkes lautet dementsprechend:

$$f_W(x) = g\left(W^{(3)}g\left(W^{(2)}g\big(W^{(1)}x\big)\right)\right).$$

Zur Übersichtlichkeit wurde hier angenommen, dass jede Layer dieselbe Activation besitzt. Verbreitete Activation Functions sind die Sigmoid-, tanh- oder ReLU-Activation:

$$\text{sigmoid(x)} = \frac{1}{1+e^{-x}}, \qquad \tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}, \qquad \text{ReLU}(x) = \max(0,x).$$

Für die Output-Layer von Regressionsnetzwerken wird meist auf eine Activation verzichtet, damit alle reellen Zahlen abgebildet werden können. Für die (K-dimensionale) Output-Layer von Klassifikationsnetzwerken wird meistens die Softmax-Activation genutzt:

$$\text{softmax}(x)_j = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}}, \qquad j = 1 \dots K\,.$$

Die Outputs dieser Funktion sind praktikabel als Wahrscheinlichkeiten interpretierbar[15]. Die vielseitige Einsetzbarkeit der Modellklasse der NNs beruht wesentlich auf dem sog. *Universal Approximation Theorem* (Hornik et al., 1989): Unter der Annahme sehr (beliebig) großer Breite (Anzahl an Nodes) und Tiefe (Anzahl an Layern), sowie unter Nutzung von nicht-linearen Activations kann gezeigt werden, dass ein FCNN eine große Menge unterschiedlicher Funktionen approximieren kann. Die Nicht-Linearität der Activations ist dabei essenziell, da das Modell ansonsten nur eine lineare Funktion aus vielen Matrix-Multiplikationen wäre. Flexible NNs sind daher sehr breit und tief und weisen somit große Anzahlen an trainierbaren Parametern auf ($\sim 10^5$ bis 10^{12}).

Das Training eines NN unterscheidet sich prinzipiell nicht vom Training anderer Modelle. Die Schwierigkeit besteht darin, dass die Ableitungen der Modellfunktion nach den Parametern durch die Kettenregel der Differentialrechnung sehr komplex werden und manuell kaum

[15] Mathematisch betrachtet liegt das daran, dass diese Funktion den Vektor $x \in \mathbb{R}^K$ auf die Menge (sog. K-*Einheitsimplexrand*) $\{y \in \mathbb{R}^K \mid y_i \geq 0,\ \sum_{i=1}^{K} y_i = 1\}$ abbildet. Somit sind die einzelnen Output-Werte auf das Intervall [0,1] beschränkt und können also die Wahrscheinlichkeit für die Zuordnung zur jeweiligen Kategorie interpretiert werden.

gefasst werden können. Daher werden sog. „Autodiff"-Verfahren[16] auf Basis des Backpropagation-Algorithmus (Rumelhart et al., 1986) verwendet, die diese Berechnungen automatisieren. Moderne Verfahren optimieren dabei zudem je nach verfügbarer Hardware die Balance zwischen verfügbarer Leistung zur Berechnung der Ableitungen und dem für ihre Speicherung verfügbaren Arbeitsspeicher (z. B. Chen et al., 2023; Dao et al., 2024).

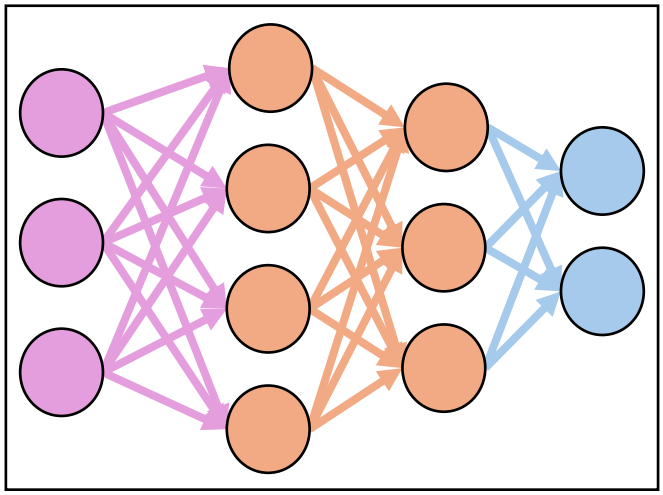


Abbildung 2.10 Schematische Darstellung eines Fully-Connected-NNs. Dabei sind die Input Layer in Rosa, die Hidden Layers in Orange und die Output Layer in Blau dargestellt.

Embeddings

Zur Verarbeitung unterschiedlicher Datenstrukturen wie Bilder und Sequenzen (beispielsweise Text) wurden NN-Sonderformen wie Convolutional NNs (LeCun et al., 1998) oder Recurrent NNs (Amari, 1972) entwickelt. Solche NN-Varianten sind dadurch gekennzeichnet, dass die Matrix-Gewichte bei der Verarbeitung einzelner Segmente der Daten (bei Text: Worte) wiederverwendet werden. Einerseits wird dadurch die Anzahl an Parametern reduziert, andererseits können so wiederkehrende Strukturen erkannt und genutzt werden. Typischerweise bei Sprachdaten, mittlerweile aber vermehrt auch bei anderen Datenformaten wie z. B. Bildern (Dosovitskiy et al., 2021), werden zudem üblicherweise sog. *Embedding*-Layer als erste Schicht des NNs genutzt (z. B. Géron, 2019). Bei Sprachdaten weist die Embedding-Layer jedem einzelnen Wort des Vokabulars des Datensatzes (siehe Abschnitt 2.6 sowie Abbildung 2.9) einen Vektor mit fixer Dimension d (üblicherweise einige 100 bis 1000) zu[17]. Die Elemente dieser Vektoren sind trainierbare Modellparameter. Allgemein werden solche numerischen Repräsentationen von Worten und Texten (aber auch anderen Daten-Segmenten) *Embeddings* genannt (z. B. Liu et al., 2020). Sie werden häufig mithilfe von Semi-

[16] Die üblicherweise für die Erstellung, Erprobung und Deployment von NNs genutzten Programmbibliotheken TensorFlow (Abadi et al., 2016) und PyTorch (Paszke et al., 2019) stellen neben einem Autodiff-Framework zudem Funktionen bereit, um die Rechenleistungen von Grafikkarten (Graphics Processing Units, GPUs), die auf die Berechnung von Matrixoperationen hin optimiert sind, für das Deep Learning zu nutzen.

[17] Eine Embedding-Layer führt letzten Endes eine Matrixmultiplikation der Embedding-Matrix mit der One-Hot-Encoding-Repräsentation des Textes durch.

Supervised-Learning-Ansätzen[18], wie der Vorhersage des folgenden Wortes bei gegebenem Satzanfang (Mikolov et al., 2013), der Vorhersage von umgebenden Worten (Pennington et al., 2014) oder der Vorhersage des Gemeinsam-Auftretens von Worten (Peters et al., 2018) ermittelt. Dadurch können Embeddings erzeugt bzw. Embedding-Modelle trainiert werden, die semantische und syntaktische Bedeutung tragen. Ein klassisches Beispiel ist hier der bei Mikolov et al. (2013) beobachtete Zusammenhang:

$$\text{Embedding(Paris)} - \text{Embedding(France)} + \text{Embedding(Italy)} \approx \text{Embedding(Rome)}\,.$$

Man hat also mithilfe der allgemeinen Sprachmodellierung in einem Self-Supervised-Learning-Ansatz eines großen Text-Datensatzes ein Embedding-Modell geschaffen, welches die semantische Bedeutung des Wortes „Hauptstadt" implizit erlernt hat. Solche Embeddings und Embedding-Layers sind dementsprechend nicht nur für sprachverarbeitende NNs als Input-Layer interessant, sondern können auch direkt selbst beispielsweise im Rahmen von Klassifikationsmodellen oder Cluster-Analysen (z. B. Grootendorst, 2022) genutzt werden. Viele Embedding-Tabellen bzw. Embedding-Modelle stehen dafür auch open-source zur Verfügung (z. B. Reimers & Gurevych, 2019).

Transformer Sprachmodelle

Ein Durchbruch in der Entwicklung von NNs war insbesondere die Entwicklung sog. *Attention*-Layers (Bahdanau et al., 2014) bzw. der sog. *Transformer*-Modellarchitektur (Vaswani et al., 2017). Diese Modelle sind auf die Verarbeitung von Sequenzen mehrdimensionaler Inputs $x_1, x_2, \dots, x_T \in \mathbb{R}^d$ ausgelegt. Das kann beispielsweise ein Text sein, bei dem die einzelnen Worte zuvor eine Embedding-Layer passiert haben. Das wesentliche Merkmal einer Attention-Layer ist, dass die einzelnen Inputs im Rahmen eines sog. Attention-Mechanismus aufeinander bezogen werden. Ein Attention-Mechanismus kann dabei jede Funktion sein, die ausschließlich von den Skalarprodukten (engl.: „Dot-Product") der Inputs abhängt:

$$a_{ij} = f\left(x_i \cdot x_j\right), \qquad i,j = 1 \dots T\,.$$

Ohne hier zu sehr ins Detail gehen zu können ermöglicht dieses Aufeinander-Beziehen der Inputs den Attention-Modellen die entsprechende Bezugsstruktur der Daten stärker zu berücksichtigen als andere Modelle es können. Die eigentlich trainierbaren Parameter-Matrizen werden typischerweise vor und nach den Attention-Layern angewandt (z. B. Devlin et al., 2019; Vaswani et al., 2017). Mittlerweile hat sich die Transformer-Architektur nicht nur für Sprachverarbeitung, sondern auch für andere Datentypen wie Bilder durchgesetzt (z. B. Dosovitskiy et al., 2021). Schlüssel für den Erfolg der Transformer-Architektur sind also einerseits die Repräsentation der Daten in Form von Embeddings und andererseits das Aufeinander-Beziehen dieser Embeddings in den Attention-Layers.

[18] Solche Ansätze werden als „*Semi-Supervised"* bezeichnet da das Training zwar einem Supervised-Learning-Workflow entspricht, allerdings Teile der „Features" (hier Text) selbst als „Target" genutzt werden und somit keine durch Menschen generierten Label benötigt werden – es sei denn, man fasst das Schreiben eines Textes als kontinuierliches „Labeln" des bisher geschriebenen Textes auf.

Die meisten aktuell verwendeten Sprachmodelle z. B. GPT2[19] (Radford et al., 2019), GPT3 (Brown et al., 2020) oder BERT (Devlin et al., 2019), die großen Sprachmodelle hinter Tools wie ChatGPT (OpenAI, 2024c) oder Konkurrenten wie Claude (Anthropic, 2024) und auch große Open-Source Sprachmodelle (z. B. LLaMA[20], Touvron et al., 2023a bzw. Touvron et al., 2023b oder OpenAssistant, Köpf et al., 2024) sind Transformer-Modelle[21]. Neben der Modellarchitektur ist für ein Sprachmodell auch das Training entscheidend. Typischerweise werden Sprachmodelle mithilfe der sog. Next-Token-Prediction trainiert, d. h. es wird der nächste Token Basis einer gewissen Menge vorangegangener Token trainiert. Man spricht auch von *autoregressivem* Training. Durch dieses Training mithilfe großer Datenmengen (bei Text-Daten typischerweise einige Gigabyte bis mehrere Terabyte) „erlernen" die Modelle eine umfassende Repräsentation von Sprache und können somit (je nach Größe) natürliche bzw. natürlich wirkende Sprache erzeugen[22]. Gleichzeitig kann dies aber auch genutzt werden, um für konkrete Anwendungsfälle höhere Performanz bzgl. anderer Aufgaben, wie dem automatisierten Scoren von Antworten (z. B. Camus & Filighera, 2020), zu erzielen. Beim expliziten Training von Sprachmodellen für konkrete Anwendungsfälle spricht man auch von *Finetuning* während man das autoregressive Training im Vorfeld auch als *Pretraining* bezeichnet. Im Rahmen der huggingface Python-Paket-Familie werden viele open-source Modelle und Methoden für Pretraining und Finetuning bereitgestellt (z. B. Wolf et al., 2020).

Die Nutzung und Verbreitung immer größerer Sprachmodelle (viele Mrd. Parameter) ist allerdings vor dem Hintergrund ihres hohen Energieverbrauchs und der damit entstehenden Kosten und Umwelteinflüsse kritisch zu betrachten[23] (Dhar, 2020). Zudem können solche Modelle nicht auf üblicher Consumer-Grade Hardware betrieben werden, da sie aufgrund ihrer Größe meist auf mehrere Recheneinheiten mit großen Arbeitsspeichermengen aufgeteilt werden müssen. Die Nutzung von Angeboten wie ChatGPT und den dazugehörigen Online-Schnittstellen (auch *Application Programming Interface*, API) ist zwar möglich, allerdings aus zwei Gründen problematisch. Erstens müssen die auszuwertenden Text-Daten dafür an einen fremden Server geschickt werden. Auch wenn es sich häufig nicht um personenbezogene Daten handelt, ist dies datenschutztechnisch nicht immer unbedenklich, da meistens keine manuelle Kontrolle zwischen der Dateneingabe durch Nutzer:innen eines Tools und dem Übergeben an die API stattfinden kann. Zweitens werden solche APIs und auch die dort angebotenen Modelle kontinuierlich weiterentwickelt und befinden sich somit in stetigem Wandel. Es kann also nicht davon ausgegangen werden, dass die Performanz des genutzten Systems für den jeweiligen

[19] Die Abkürzung „GPT" wird in unterschiedlichen Modellbezeichnungen verwendet und steht für „**G**enerative **P**retrained **T**ransformer".

[20] **L**arge **L**anguage **M**odel **M**eta **AI** (Touvron et al., 2023a)

[21] Aufgrund der hohen benötigten Rechenleistung der Attention-Mechanismen werden aktuell aber auch alternative Ansätze exploriert (Gu & Dao, 2023; Peng et al., 2023; Zhai et al., 2021a; Zhu et al., 2024).

[22] Um einen Dialog führen zu können, wie bekannte Chatbots wie ChatGPT es tun, werden die Modelle typischerweise noch mithilfe des sog. *Reinforcement Learning from Human Feedback* (z. B. Christiano et al., 2017; Ouyang et al., 2024) bzw. *Instruction-Funetuning* (z. B. Chung et al., 2024) weitertrainiert, um typische Gesprächsstrukturen abbilden zu können.

[23] Es gibt allerdings auch zunehmend Bestrebungen die notwendige Rechenleistung und Speichernutzung der Modelle zu Reduzieren (z. B. Dettmers et al., 2024).

Anwendungsfall konstant bleibt. Auch die Bedienung der APIs selbst kann sich ändern. Häufig reichen aber auch kleinere Modelle wie das BERT-Modell[24] (Devlin et al., 2019) aus, um für spezifische Anwendungsfälle eine ausreichende bis gute Performanz zu erreichen (Camus & Filighera, 2020). Im vorliegenden Projekt wird daher auch primär der Ansatz des Trainings eines eigenen, kleineren Modells mit vollständiger Kontrolle über das Modell und die Daten gewählt, auch wenn Alternativen ergänzend exploriert und evaluiert werden (siehe Kapitel 6, insbesondere Abschnitt 6.7).

Insgesamt werden ML- und Deep-Learning-basierte Methoden und Modelle sowohl für explorative Sprachanalysen (Blei, 2012; Grootendorst, 2022; Roberts et al., 2019) als auch in Supervised-Settings wie beispielsweise automatisiertem Scoring (Gamieldien et al., 2023; Lee et al., 2019; Ludwig et al., 2021; Maestrales et al., 2021; Mayfield & Rosé, 2012; Sawatzki et al., 2022; Yan et al., 2020) eingesetzt. Es werden auch über diese konkreten Ansätze hinaus große Potenziale für die Nutzung dieser Technologien im Kontext der Bildung und Bildungsforschung sowie konkret im naturwissenschaftsdidaktischen Bereich identifiziert, darunter insbesondere die Nutzung von ML-Methoden zu Forschungszwecken (Hilbert et al., 2021; Kubsch et al., 2023; Kubsch et al., 2021a) und für automatisiertes Assessment und Feedback über basales Scoring hinaus (Fütterer et al., 2023; Zhai et al., 2023).

[24] Ca. 110 Mio. Parametern, was bei einer Gleitkommapräzision von 32-Bit (`float32`) ca. 440 MB an Arbeitsspeicher entspricht.

3. Projektstruktur und Forschungsfragen

Im hier vorgestellten Dissertationsprojekt sollen (a) die innere Struktur des FDW von Lehramtsstudierenden empirisch basiert detaillierter inhaltlich beschrieben werden sowie (b) Möglichkeiten zur Automatisierung des Assessments des FDW unter anderem auf Basis der gefundenen Strukturen exploriert werden. Da es sich um ein kumulatives Dissertationsprojekt handelt, werden vor allem konkrete theoretische Grundlagen eher im Rahmen der jeweiligen (auch in dieser Rahmung enthaltenen) Artikel (Kapitel 4, 5 & 6) detaillierter erläutert. Übergreifende theoretische Grundlagen zum Professionswissen und insbesondere zum FDW von Lehrkräften mit dem Fokus auf der Physik und den Naturwissenschaften wurden daher in den Abschnitten 2.1 und 2.2 einleitend eher knapp eingeführt.

Da in den Artikeln oft nicht genügend Raum für ausführliche methodische Ausführungen verfügbar war bzw. ist, wurden folgende methodische Aspekte detaillierter erläutert:

- Item-Response-Kompetenzniveaumodelle (Abschnitt 2.3)
- ML mit Fokus auf Anwendungen in der naturwissenschaftsdidaktischen Forschung (Abschnitt 2.4 & 2.5)
- ML-basierte und Deep-Learning-basierte Sprachverarbeitung (Abschnitt 2.6 & 2.7)

In diesem Rahmen wurden insbesondere der allgemeine Workflow und die Herangehensweise bei der Nutzung von ML-Methoden und die Unterscheidung in Supervised- und Unsupervised-Learning-Methoden vorgestellt. Unsupervised-Learning-Methoden werden in diesem Projekt zur explorativen Untersuchung der inneren Strukturen des FDW eingesetzt. Dabei wird die CGT (Nelson, 2020) genutzt, um die Analysen zu strukturieren. Supervised-Learning-Methoden werden in diesem Projekt primär für das automatisierte Assessment der gefundenen FDW-Strukturen eingesetzt. Entsprechende Ansätze für automatisiertes Assessment existieren bereits länger, beispielsweise auf Basis von Embedding-Methoden wie der Semantic Analysis (z. B. Andersen & Zehner, 2021; Leacock & Chodorow, 2003; Zehner et al., 2016). In diesem Projekt wird sich primär auf die vielversprechende (z. B. Camus & Filighera, 2020) Nutzung von Transformer-Sprachmodellen für das Assessment fokussiert.

In diesem Kapitel folgt nun ein Überblick über die Ziele und Forschungsfragen des Projekts (Abschnitt 3.1). Dabei werden die in dieser Arbeit enthaltenen Artikel bereits grob in den Gesamtkontext eingeordnet. Anschließend wird der verwendeten Gesamtdatensatz vorgestellt (Abschnitt 3.2).

3.1. Forschungsziele und -fragen

Im Theorieteil wurden zwei wesentliche Forschungsdesiderate dargestellt. Erstens ist eine detailliertere empirisch fundierte inhaltliche Beschreibung der inneren Struktur des FDW notwendig um (a) das FDW im Kontext gängiger Rahmenmodelle wie dem RCM oder MoC weiter auszudifferenzieren (Abschnitt 2.2) und (b) ein Assessment zu ermöglichen, welches über die reine Angabe von Scores hinausgeht. Zweitens wäre es aufgrund des hohen Aufwands bei der Auswertung der als besonders authentisch geltenden offenen Antwortformate

entsprechender Testinstrumente wünschenswert, ein solches Assessment mithilfe moderner Methoden der Datenverarbeitung zu automatisieren. Auf Basis der verfügbaren Daten werden zur Bearbeitung dieser Desiderate im Folgenden die Zielpakete dieses Dissertationsprojekts abgeleitet. Dabei werden die jeweiligen Forschungsfragen, die zu den Zielpaketen in den Artikeln fokussiert werden, hier bereits vorweggegriffen, um einen Überblick darzustellen. Die einzelnen Forschungsfragen unterscheiden sich den Anforderungen der jeweiligen Zeitschriften und Gutachtenden entsprechend teilweise in ihrem Stil und Grad der Konkretisierung. Zwischen den einzelnen Zielpaketen werden diese knapp im Projektkontext verortet und ggf. wird die Genese der entsprechenden Artikel knapp kommentiert. In Abbildung 3.1 ist der gesamte Workflow des Projekts noch einmal (möglichst) übersichtsartig dargestellt.

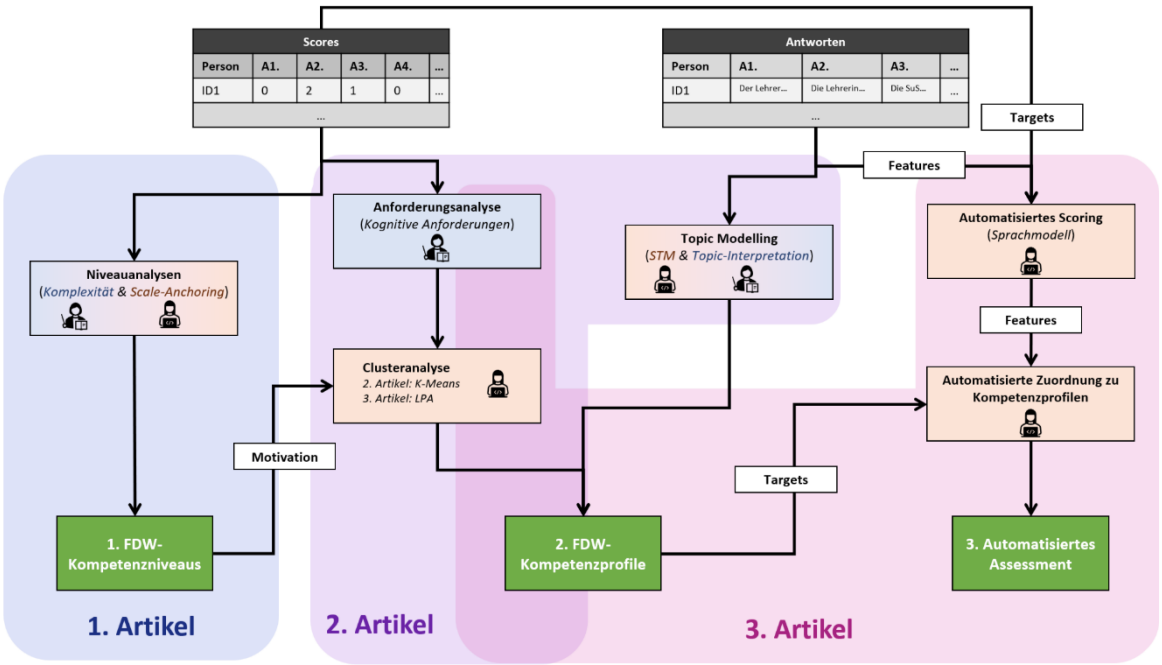


Abbildung 3.1 Übersichtsdarstellung der drei Zielpakete und des Workflows des Projekts. Den Ausgangspunkt der Analysen stellen die Score- und Textdaten dar. Diese werden im Rahmen von sowohl (eher) computerbasierten bzw. quantitativen Methoden (orange Kästen) als auch im Rahmen von eher theoriegeleiteten Schritten (blaue Kästen) genutzt, um die Projektziele (grüne Kästen) zu erreichen. Die farbigen Segmente im Hintergrund repräsentieren die drei Artikel und die jeweiligen Schritte der Analyse, die in ihnen jeweils eine Rolle spielen.

In Abschnitt 2.3 wurde dargestellt, dass zur empirisch basierten inhaltlichen Beschreibung der inneren Struktur des FDW bisher primär hierarchische Ansätze in Form von Niveaumodellen vorliegen. Diese Modelle sind zudem bisher weitestgehend voneinander isoliert, da sie sich auf die in den jeweiligen Projektkontexten unterschiedlichen konkret fokussierten FDW-Inhaltsbereiche bzw. FDW-Facetten beziehen. Im Rahmen des ersten Zielpakets soll dieser Forschungslücke durch eine projektübergreifende Analyse begegnet werden, deren Ergebnisse auch die späteren nicht-hierarchischen Analysen (Zielpaket 2) vorbereiten:

1. Zielpaket – Kompetenzniveaus:

Durchführung einer (projektübergreifenden) Analyse von hierarchischen Kompetenzniveaus des FDW auf Basis von IRT-Modellen

- ***FF1.1*** *(Artikel 1)*: Inwieweit lassen sich mithilfe des Scale-Anchoring-Verfahrens projektübergreifend inhaltliche Strukturen des FDW identifizieren und inhaltlich charakterisieren?
- ***FF1.2*** *(Artikel 1)*: Inwieweit lassen sich Stufen hierarchischer Komplexität des FDW projektübergreifend identifizieren und inhaltlich charakterisieren?

Der vollständig auf die Kompetenzniveauanalysen fokussierte Artikel 1 wurde in Kooperation zwischen den bzw. Teilen der Projektgruppen KiL[25] (z. B. Schiering et al., 2019) bzw. KeiLa[26] (z. B. Schiering et al., 2023) und ProfiLe-P(+) (z. B. Riese et al., 2022b; Vogelsang et al., 2019) erarbeitet. Im Rahmen dieser Analysen zeigten sich zwar projektübergreifende Gemeinsamkeiten bezüglich kognitiver Anforderungen (Reproduzieren, Anwenden, Evaluieren etc., siehe Abschnitt 4.5 & 4.6), allerdings blieben die Beschreibungen der Niveaustufen aufgrund methodischer Limitationen recht allgemein und vor allem auf hierarchische Abstufungen beschränkt. Nicht-hierarchische inhaltliche Strukturen des FDW wurden bisher bis auf konfirmatorische Modellvergleiche (Riese et al., 2017) kaum empirisch fundiert untersucht. Daher wurden im zweiten Zielpaket (potenziell) nicht-hierarchische Strukturen genauer in den Blick genommen:

2. Zielpaket – Kompetenzprofile:

Durchführung von explorativen Analysen des FDW auf Basis von sowohl Scoredaten als auch Sprachdaten zur Beschreibung von Probandengruppen mit prototypischen Antwortverhalten und Kompetenzausprägungen

- ***FF2.1*** *(Artikel 2 ~ Pattern Detection)*: Welche Kompetenzprofile des FDW können mithilfe einer *(K-Means)* Clusteranalyse der Scores gefunden werden?
- ***FF2.2*** *(Artikel 2 ~ Pattern Refinement)*: Zeigen Proband:innen, die zu einem bestimmten Kompetenzprofil gehören, prototypische Sprachnutzung in ihren Antworten zu den offenen Aufgaben des eingesetzten Testinstruments?
- ***FF2.3*** *(Artikel 2 ~ Pattern Confirmation)*: Wie hoch ist die Performanz von ML-Modellen bei der Zuordnung zu den Kompetenzprofilen für ungesehene Daten?

---

[25] Akronym KiL: „Messung professioneller **K**ompetenzen **i**n mathematischen und naturwissenschaftlichen **L**ehramtsstudiengängen“, gefördert durch Leibniz Gemeinschaft. In diesem Projekt wurde das Professionswissen von Lehramtsstudierenden der mathematisch-naturwissenschaftlichen Fächer gemeinsam modelliert. Für die Physik wurde dabei ein FDW-Testinstrument von Kröger (2019) entwickelt.

[26] Akronym KeiLa: „**K**ompetenz**e**ntwicklung **i**n mathematischen und naturwissenschaftlichen **L**ehr**a**mtsstudiengängen“, gefördert durch Leibniz Gemeinschaft. Aufbauend auf den Modellierungen aus KiL wurde in KeiLa die Entwicklung des Professionswissens im Zusammenhang mit Lerngelegenheiten und individuellen Merkmalen der Proband:innen untersucht (z. B. Sorge et al., 2019) Auch hier wurde das FDW-Testinstrument nach Kröger (2019) wieder eingesetzt. In der hier vorgestellten Analyse werden die FDW-Daten aus beiden Projektphasen genutzt.

- ***FF2.4*** *(Artikel 3 ~ „Refined" Pattern Detection)*: Welche latenten FDW-Kompetenzprofile lassen sich durch eine GMM-basierte LPA in den FDW-Score-Daten des Projekts ProfiLe-P bezüglich der kognitiven Anforderungskategorien *Reproduzieren*, *Anwenden-Kreieren* und *Analysieren-Evaluieren* finden*?*

Die Analysen zum zweiten (und dritten) Zielpaket sind aus methodischen Gründen sowie Gründen des Workloads allerdings wieder auf das Projekt ProfiLe-P beschränkt. Die Analysen in Artikel 2 und 3 sind dabei wesentlich durch die CGT strukturiert (vgl. auch Abschnitt 2.5) und umfassen neben den explorativen Analysen der Scores und Textdaten in den Pattern Detection (FF2.1 bzw. FF2.4) bzw. Pattern Refinement Schritten (FF2.2) auch eine Analyse der Performanz von ML-Modellen bei der Zuordnung zu den Kompetenzprofilen für ungesehene Daten (FF2.3). Im Rahmen dieser FF2.3 werden die Kompetenzprofile zunächst ausgehend von den Scores vorhergesagt. Eine tatsächliche automatisierte Auswertung der offenen Aufgaben ist in Artikel 2 aus Platzgründen und Gründen der Projektgenese noch nicht eingeschlossen.

In den Analysen zu Artikel 2 (FFs 2.1 bis 2.3) konnten aufgrund von methodischen Limitationen noch keine echt „latenten" Kompetenzprofile ermittelt werden. Im Rahmen des Artikels 3 bot sich somit die Gelegenheit, die sich abzeichnenden inneren Strukturen des FDW noch einmal aus einer etwas anderen Perspektive mithilfe einer Latenten Profilanalyse (LPA, Spurk et al., 2020) zu untersuchen[27]. Zu diesem Zweck wurden die betrachteten kognitiven Anforderungskategorien entsprechend den vorangegangenen Ergebnissen zu FF1.1 und FF2.1 zusammengefasst, sodass eine LPA ermöglicht wurde. Die LPA liefert ein stärkeres Argument für den prototypischen Charakter der Kompetenzprofile als eine K-Means Analyse, da sie echte latente Strukturen abbildet. Auch Artikel 3 ist als eher inhaltliche Analyse daher wieder mithilfe der CGT strukturiert. FF2.4 kann dabei als eine erweiterte („refined") Pattern Detection auf Basis des vorherigen Analysezyklus (vor allem Artikel 2) aufgefasst werden. Die folgenden Forschungsfragen zum automatisierten Assessment (Zielpaket 3) dienen dann als Pattern Confirmation der neuen gefundenen Strukturen:

3. Zielpaket – Automatisiertes Assessment:

Erprobung der Nutzung von Machine-Learning-Modellen zur automatisierten Auswertung des FDWs

FF3.1 *(Artikel 3 ~ Pattern Confirmation I)*: Welche Maschine-Mensch-Übereinstimmung erreicht ein BERT-Sprachmodell (Devlin et al., 2019) bei der Vorhersage von FDW-Scores unter Nutzung eines typischen Finetuning-Workflows auf Basis von 846 Bearbeitungen des FDW-Testinstruments?

FF3.2 *(Artikel 3 ~ Pattern Confirmation II)*: Wie hoch ist die Maschine-Mensch-Übereinstimmung einer automatisierten Zuordnung von Bearbeitungen des FDW-

[27] Dies wurde auch im Review des dritten Artikels von Seiten der Herausgebenden und Reviewenden im Rahmen der stärkeren Herausarbeitung des inhaltlichen Erkenntnismehrwerts im Vergleich zu einer Fokussierung primär auf das automatisierte Assessment (Zielpaket 3) gewünscht.

Testinstruments zu einem prototypischen FDW-Kompetenzprofil auf Basis der maschinellen Score-Vorhersagen (FF3.1)?

Der Pattern Confirmation Schritt unter Nutzung der tatsächlichen Antworten auf die Testaufgaben als „Rohdaten“ (FF3.1 & FF3.2) liefert ein deutlich stärkeres Argument für die Robustheit und Validität der in FF2.4 gefundenen Strukturen als die Pattern Confirmation auf Basis der Scores im zweiten Artikel (FF2.3), wie in Abschnitt 6.7.3 noch einmal genauer diskutiert wird. Darüber hinaus stellt dieser Schritt auch den wesentlichen Beitrag zum Forschungsdesiderat der Automatisierung eines inhaltlich reichhaltigen FDW-Assessments auf Basis von Kompetenzprofilen und fundierten Subskalen dar. Dabei wird in Artikel 3 jedoch aus Platzgründen nicht auf weitere erprobte Workflows oder Modelle (neben dem BERT-Modell) eingegangen, die allerdings im Kontext des Automatisierungsdesiderats auch für andere Projekte von Interesse sein können. Ergänzend werden solche Betrachtungen daher in Abschnitt 6.7 detaillierter dargestellt.

3.2. Stichprobe und Datenaufbereitung

Der für die Analysen und Erprobungen dieses Projekts hauptsächlich verwendete Datensatz stammt aus dem Projekt ProfiLe-P+[2], in dem PW, FW, FDW, affektive Orientierungen und Beliefs sowie Performanz in prototypischen Handlungssituationen wie dem Planen von Physikunterricht und dem Erklären physikalischer Phänomene von (angehenden) Lehrkräften erhoben wurden. Dabei werden sowohl die quantitativen Score-Daten, d. h. die manuell vergebenen Scores zu den einzelnen Aufgaben, als auch qualitative Text-Daten in Form der Antworten auf die offenen Aufgaben des verwendeten FDW-Testinstruments genutzt.

Für die projektübergreifenden Analysen in Artikel 1 (FF1.1 & FF1.2) stehen zudem die IRT-Thurstone-Thresholds und IRT-Personenfähigkeiten (siehe Abschnitt 2.3) sowie einige demographische Informationen aus dem FDW-Datensatz der Projekte KiL und KeiLa zur Verfügung. Da die Vorbereitung dieses Datensatzes nicht im Aufgaben- und Verantwortungsbereich des Projekts selbst lag, wird dieser Datensatz und auch das Testinstrument hier nicht zusätzlich zu den in Artikel 1 enthaltenen Informationen (Abschnitt 4.4.1) vorgestellt[28]. Eine Beispielaufgabe aus diesem Testinstrument ist in Abbildung 4.4 dargestellt.

Im Folgenden wird nun der zentrale FDW-Datensatz aus dem Projekt ProfiLe-P+ detaillierter beschrieben und es werden einige Informationen zusätzlich zu den Artikeln ergänzt. Das verwendete FDW-Testinstrument nach Gramzow (2015) besteht aus 20 offenen und 4 geschlossenen Multiple-Choice-(MC)-Aufgaben, wobei 3 der offenen Aufgaben aus je 2 einzeln bepunkteten Teilaufgaben bestehen. Daher ist teilweise auch von 23 offenen Aufgaben die Rede. Die MC-Aufgaben bestehen aus 4 bis 6 einzelnen Items, sodass insgesamt je nach Publikation auch 20 MC-Items und somit insgesamt 43 Items berichtet werden. Die Aufgaben und das Kodiermanual zur Bewertung der offenen Aufgaben wurden mithilfe von Curriculumsanalysen, Expertenbefragungen und Think-Aloud-Studien erprobt und validiert

[28] Für ausführlichere Informationen zum dort verwendeten Testinstrument sei auf Kröger (2019) verwiesen. Der Datensatz und das IRT-Modell werden zudem in Schiering et al. (2023) bereits ausführlich vorgestellt.

(Gramzow, 2015). Das Testinstrument umfasst die fachdidaktischen Facetten *Instruktionsstrategien*, *Schülervorstellungen*, *Experimente und Vermittlung eines angemessenen Wissenschaftsverständnisses* (kurz *Experimente*) sowie *Fachdidaktische Konzepte*. Diese Facetten stellen eine begründete Auswahl möglicher Facetten dar. In der ursprünglichen Testentwicklung wurden zudem die kognitiven Anforderungen *Reproduzieren*, *Anwenden* und *Analysieren* abgedeckt. Da im Projekt ProfiLe-P eine möglichst detaillierte Erfassung des Professionswissens bezüglich des Fachinhalts *Mechanik* angestrebt wurde, ist auch das FDW-Testinstrument auf diesen Fachinhalt fokussiert. Das gesamte Item-Entwicklungsmodell des Testinstruments ist in Abbildung 4.1 enthalten. Im Rahmen des hier vorgestellten Projekts wurden die Aufgaben aber mit einem größeren Fokus auf den kognitiven Anforderungen re-analysiert (Kapitel 5 & 6). Beispielaufgaben dieses Testinstruments sind in Abbildung 4.3 und Figure 5.3 dargestellt.

Das FDW-Testinstrument wurde im Rahmen des Projekts ProfiLe-P+ von 2016 bis 2019 in Bachelor- und Masterstudiengängen des Physik-Lehramts an 12 deutschsprachigen Universitäten eingesetzt. Insgesamt umfasst der Datensatz 846 Bearbeitungen dieses Testinstruments in Quer- und Längsschnitten[29]. Die Bearbeitungen werden dabei hier der Methode virtueller Probanden (Davier et al., 2008; siehe auch Wright, 2003) folgend in allen Analysen als unabhängige Bearbeitungen betrachtet. Die diesen Bearbeitungen entsprechenden virtuellen Probanden sind im Mittel ca. 23 ($M = 22.80$, $SD = 4.60$) Jahre alt und befinden sich ca. im 4. Fachsemester ($M = 4.11$, $SD = 3.51$). Darüber hinaus sind 34 % weiblich und 79 % befinden sich im Bachelorstudium, die übrigen 21 % im Masterstudium.

Die Testhefte lagen zu Beginn dieses Dissertationsprojekts analog und teilweise als Scans in PDF-Format vor. Die fehlenden Testhefte wurden ebenfalls gescannt und die Antworten auf die offenen Aufgaben durch Hilfskräfte vollständig digitalisiert, um die späteren computerbasierten Textanalysen etc. zu ermöglichen. Darüber hinaus wurden bereits während des ProfiLe-P+- Projekts alle offenen Aufgaben durch eine trainierte Kodiererin bepunktet. Insgesamt liegen somit 15600 Antwort-Score Paare (454 bis 825 pro Aufgabe) mit Scores zwischen 0 und 3 Punkten vor. Die MC-Aufgaben wurden dem Vorgehen in anderen Teilen des ProfiLe-P-Verbunds folgend (Jordans et al., 2022; Kulgemeyer et al., 2023) entsprechend des K-prim-Schwellensystems (Krebs, 1997) bewertet, sodass die durch Raten erreichbaren Punktzahl und die Übergewichtung der MC-Aufgaben reduziert wurden. Im Rahmen dieses Systems werden die MC-Aufgaben gemäß ihrer Einzelitems mit 0, 1 oder 2 Punkten bewertet. Eine MC-Aufgabe mit beispielsweise 4 Einzelitems wird…

- …mit 0 Punkten bewertet, wenn 2 oder weniger Einzelitems korrekt gelöst wurden.
- …mit 1 Punkt bewertet, wenn 3 Einzelitems korrekt gelöst wurden.
- …mit 2 Punkten bewertet, wenn alle 4 Einzelitems korrekt gelöst wurden.

[29] Da längsschnittliche Daten vorliegen, war zu Beginn des hier vorgestellten Projekts geplant, auch längsschnittliche Analysen im Kontext der Kompetenzprofile durchzuführen. Die genauere Betrachtung des Datensatzes zeigte aber, dass der Anteil an Proband:innen, die tatsächlich auch an mehreren Erhebungen teilgenommen haben, zu klein bzw. der Dropout zu groß ist, um hier belastbare Aussagen ableiten zu können (siehe Anhang D).

Fehlende Antworten werden mit 0 Punkten bewertet. Die Gesamtanzahl an Antworten bzw. vergebenen Punkten ist in Tabelle 3.1 dargestellt.

Tabelle 3.1 Anzahl der Punktzahlen in den einzelnen Aufgaben

	Offene Aufgaben				**Alle Aufgaben und Missings als 0**			
Punktzahl	0	1	2	3	0	1	2	3
Anzahl	8800	5128	1646	26	10071	5780	2976	26

Neben dieser Hauptkodierung, die für die eigentlichen Analysen verwendet wurde, liegt noch eine weitere Kodierung von 267 Testheften (4748 Antworten zu offenen Aufgaben) durch einen anderen trainierten Kodierer vor. Bezüglich der offenen Aufgaben beträgt die Interrater-Übereinstimmung $\kappa = 0.665$ (Cohens κ; z. B. Fleiss & Cohen, 1973) und ist somit als gute Übereinstimmung einzuordnen (Döring, 2023). Schließt man die MC-Aufgaben (nach Anwendung der K-prim Schwellen) mit ein und füllt fehlende Werte mit 0 Punkten, so beträgt die Interrater-Übereinstimmung bezogen auf das gesamte Testinstrument $\kappa = 0.761$, was als sehr gute Übereinstimmung eingeordnet werden kann (Döring, 2023).

Die Datenauswertung und Analyse finden mit den Programmiersprachen R (R Core Team, 2024) und Python (Python Software Foundation, o. D.) statt. Der vollständige Analysecode und sämtliche Ergebnisse etc. sind im digitalen Begleitmaterial zu finden.

4. Empirisch-kriterienorientierte Analyse des fachdidaktischen Wissens angehender Physiklehrkräfte (*Artikel 1*)

Einordnung in das Gesamtprojekt

Ausgangspunkt der ersten Analyse dieses Projekts bilden zwei unabhängige Niveauanalysen des FDW von Physiklehrkräften. Schiering et al. führten zunächst eine Scale-Anchoring-Analyse auf Basis der Daten des KiL-Projekts (Schiering et al., 2019) und später auch des Gesamtdatensatzes aus den Projekten KiL und KeiLa (Schiering et al., 2023) durch. Orientiert an diesem Vorgehen führten Zeller et al. (2022) ebenfalls eine Scale-Anchoring-Analyse auf Basis eines Datensatzes zu einer geschlossenen Version des FDW-Testinstruments aus dem ProfiLe-P-Projekt (Jordans et al., 2022) durch. Dabei fiel auf: Auch wenn die konkreten Bezüge zu fachinhaltlichen Themen und fachdidaktischen Inhalten bzw. Facetten naturgemäß unterschiedlich waren, so deuteten sich doch Gemeinsamkeiten bezüglich lernpsychologisch interpretierbarerer Operatoren wie „nennen“, „kennen“, „bewerten“ etc. an.

Da in der Analyse von Zeller et al. (2022) allerdings nur eine recht kleine Datenbasis verfügbar war und die Übertragbarkeit der Beobachtungen auf das ursprüngliche FDW-Testinstrument nicht uneingeschränkt angenommen werden kann und konnte (Kulgemeyer et al., 2023), wurde die Analyse für den Gesamtdatensatz des ProfiLe-P+ Projekts wiederholt. Dabei wurde zudem vorgeschlagen, diese Analysen in einer vergleichenden Betrachtung mit den Ergebnissen von Schiering et al. (2023) zusammenzuführen. Es konnte allerdings kein gemeinsames IRT-Modell der beiden Datensätze (KiL/KeiLa und ProfiLe-P+) genutzt werden, da die zu diesem Zweck notwendige Verknüpfung durch gemeinsame Aufgaben oder eine Überschneidung in der Stichprobe nicht gegeben war. Um die bisher isoliert stehenden Modelle zusammenzuführen, wurde daher der Weg über die inhaltliche Beschreibung der Niveaus gewählt. Neben den Scale-Anchoring-Analysen wurde zudem eine regressionsanalytische Niveaubildung angestrebt. So sollte der beschränkten Aufgabenanzahl der Testinstrumente, die das Scale-Anchoring-Verfahren erschwert, begegnet werden. Nach dem Vorbild bestehender regressionsanalytischer Niveaumodelle für das Fachwissen (Bernholt, 2010; Woitkowski, 2015; Woitkowski & Riese, 2017) wurde hierfür die Adaption des Modells hierarchischer Komplexität (Commons et al., 2014; Commons et al., 1998) vorgeschlagen.

Die Ergebnisse des Scale-Anchoring-Verfahrens bestätigen die Vermutungen und zeigen, dass sich das FDW unabhängig von den fokussierten fachdidaktischen und fachlichen Inhalten / Facetten in niedrigen Niveaus auf reproduktive Aspekte beschränkt, während in hohen Niveaus analytische, kreative, anwendungsorientierte und bewertende Aspekte hinzukommen. Diese Beobachtungen und insbesondere ihr projektübergreifender Charakter dienen als Grundlage für die vorbereitenden Schritte der nicht-hierarchischen explorativen Analysen in den Artikeln 2 und 3. In diesem Sinne kann die Kompetenzniveauanalyse als erster Pattern Detection Schritt eines das Gesamtprojekt überspannenden CGT-Workflows aufgefasst werden. Die daraus folgende Fokussierung auf die kognitiven Anforderungen stellen in diesem Sinne dann ein theorie- und empiriegeleitetes Pattern Refinement dar.

Bibliographische Angabe

Zeller, J., Schiering, D., Kulgemeyer, C., Neumann, K., Riese, J. & Sorge, S. (2024). Empirisch-kriterienorientierte Analyse des fachdidaktischen Wissens angehender Physiklehrkräfte: Welche inhaltlichen Strukturen zeigen sich über unterschiedliche Projekte hinweg? *Unterrichtswissenschaft*. http://doi.org/10.1007/s42010-024-00200-w

Zusammenfassung

In den letzten Jahren wurde das Professionswissen (angehender) Lehrkräfte intensiv untersucht. Neben Aussagen zur inneren Struktur liegen auch Ergebnisse über den Zusammenhang zwischen Professionswissen, Performanz in prototypischen Handlungssituationen sowie Unterrichtserfolg vor. In diesen Analysen hat sich gezeigt, dass insbesondere dem fachdidaktischen Wissen eine zentrale Rolle zukommt. Es mangelt bisher jedoch an empirisch fundierten Beschreibungen von Niveaustufen des fachdidaktischen Wissens. Zwar liegen einzelne Vorschläge vor, diese sind jedoch entweder empirisch nicht fundiert oder post hoc generiert, so dass unklar ist, inwieweit die Beschreibung der Ausprägungen auch außerhalb der jeweiligen Projektkontexte anwendbar ist. Der vorliegende Artikel stellt eine projektübergreifende Analyse des fachdidaktischen Wissens mithilfe zweier Ansätze zur Bildung von Niveaustufen vor. Dazu werden Niveaumodelle mit Daten zum fachdidaktischen Wissen aus zwei Projekten ($N = 427$ und $N = 779$) mithilfe des Scale-Anchoring-Verfahrens sowie eines regressionsanalytischen Ansatzes auf Basis eines Modells hierarchischer Komplexität erstellt. Das Scale-Anchoring-Verfahren liefert Niveaubeschreibungen, die sich zwar bezüglich fachlicher und fachdidaktischer Inhalte unterschieden, aber Parallelen bezüglich lernpsychologisch interpretierbarer Operatoren zeigten. Projektübergreifend deuteten die Ergebnisse daraufhin, dass sich das fachdidaktische Wissen in niedrigen Ausprägungen auf reproduktive Aspekte beschränkt, in höheren Ausprägungen aber kreative und evaluierende Elemente hinzukommen. Das Modell hierarchischer Komplexität zeigte sich nur für einen der Datensätze als geeignet, um ein Niveaumodell abzuleiten und konnte daher für projektübergreifende Analysen nicht weiter genutzt werden. Nichtsdestotrotz lieferte die projektübergreifende Analyse mithilfe des Scale-Anchoring-Verfahrens kontextunabhängige Beschreibungen von Ausprägungen des fachdidaktischen Wissens und ermöglicht so erste Schritte in Richtung eines empirisch fundierten, inhaltlich reichhaltigen Assessments, welches über eine Einordnung mittels eines Scores hinaus geht.

Schlüsselwörter: Fachdidaktisches Wissen · Niveaumodell · Projektübergreifende Analyse · Physik

Cross-project empirical and criteria-oriented analysis of pre-service physics teachers' pedagogical content knowledge

What content structures emerge in the context of different models?

Abstract

In recent years, the professional knowledge of (pre-service) teachers has been intensively investigated. In addition to statements regarding its internal structure, there are also findings on the relationship between professional knowledge, performance in prototypical action situations, and teaching effectiveness. These analyses have shown that pedagogical content knowledge plays a central role. However, there is still a lack of an empirically grounded description of competency levels of pedagogical content knowledge. There have been some individual proposals, though they are either not empirically grounded or post hoc generated, leaving the extent to which the descriptions of such levels are applicable outside the specific project contexts unclear. This article presents a cross-project analysis of pedagogical content knowledge using two approaches to establish levels of proficiency. Therefore, level models were constructed based on data regarding pedagogical content knowledge from two projects ($N = 427$ and $N = 779$) using the Scale-Anchoring procedure and a regression-analytical approach based on a model of hierarchical complexity. The Scale-Anchoring procedure provided level descriptions that, despite differences in subject matter and pedagogical content, exhibited parallels in terms of operators that are interpretable in terms of learning psychology. Across projects, the results indicated that pedagogical content knowledge in low levels is limited to reproductive aspects but incorporates creative and evaluative elements in higher levels. The model of hierarchical complexity turned out to be properly applicable only for one of the datasets and thus could not be further utilized for cross-project analyses. Nevertheless, the cross-project analysis using the Scale-Anchoring procedure provided context-independent descriptions of levels of pedagogical content knowledge, thus enabling initial steps towards an empirically grounded, conceptually rich assessment that goes beyond solely preparing a quantitative score.

Keywords: Pedagogical content knowledge · Competency level model · Cross-project analysis · Physics

4.1. Einleitung

Die professionelle Kompetenz (angehender) Lehrkräfte steht seit langem im Fokus der fachdidaktischen Forschung zur Professionalisierung von Lehrkräften (Baumert & Kunter, 2006; Gess-Newsome, 1999; Shulman, 1986; Terhart, 2012). Die professionelle Kompetenz wird dabei in unterschiedlichen Konzeptualisierungen als wesentlich für die Handlungsqualität im Unterricht oder für den Unterrichtserfolg aufgefasst (Ball et al., 2001; Harms & Riese, 2018; Terhart, 2012). Eine zunehmende Anzahl an Studien belegt diese Annahme (z. B. Blömeke et al., 2022; Keller et al., 2017; Kunter et al., 2013). Speziell in den Naturwissenschaften wurden in den vergangenen Jahren insbesondere die innere Struktur und die globale Entwicklung des Professionswissens sowie die Abhängigkeit dieser Entwicklung von anderen Konstrukten untersucht (Neumann et al., 2019; Riese et al., 2017; Schiering et al., 2019; Sorge et al., 2018). Darüber hinaus liegen Ergebnisse zur Bedeutung des Professionswissens für die Performanz in prototypischen Handlungssituationen vor (z. B. Förtsch et al., 2016; Kulgemeyer et al., 2020; Kulgemeyer & Riese, 2018; Riese et al., 2022a).

Im Rahmen von Projekten wie den genannten werden üblicherweise ausgehend von gängigen Operationalisierungen des Professionswissens Testinstrumente erstellt, die häufig konkrete Aspekte, wie das thematisierte Fachwissen oder spezielle Professionswissensdimensionen fokussieren. Dadurch wird ein direkter Vergleich der vorliegenden Ergebnisse erschwert, da unklar ist, inwieweit die durch diese Testinstrumente abgebildeten Konstrukte deckungsgleich sind. Gleichzeitig stellt die möglichst allgemeingültige, theoretisch begründete und empirisch fundierte Beschreibung von Ausprägungen oder sogar Entwicklungsstufen des Professionswissens und der Professionswissensdimensionen bereits länger ein Forschungsdesiderat dar (z. B. Kaiser et al.), denn die Möglichkeit zur Einordnung von Personen oder Lerngruppen in ein entsprechendes Niveaumodell ist für eine inhaltlich nützliche Diagnose und die Identifikation von Entwicklungspotenzialen notwendig.

Das fachdidaktische Wissen (FDW) stellt in den meisten theoretischen Modellen eine Kerndomäne des Professionswissens von Lehrkräften dar und eine Vielzahl empirischer Ergebnisse belegt seine praktische Relevanz (z. B. Kulgemeyer & Riese, 2018). Gerade für das FDW als „special amalgam“ (Shulman, 1987 siehe auch Neumann et al., 2019), d. h. als spezielle, für die Lehrprofession einzigartige „Mischung“ von fachlichem und pädagogischem Wissen, gestaltet sich jedoch eine projektunabhängige Beschreibung von Ausprägungen als herausfordernd, denn auch aufgrund dieses Mischungscharakters fokussieren die in unterschiedlichen Studien verwendeten Testinstrumente häufig einzelne Aspekte wie z. B. konkretes Fachwissen und Subskalen (siehe z. B. Hume et al.)[30]. Daher können bisherige Untersuchungen des FDW und deren Ergebnisse bisher meist nur eingeschränkt miteinander verglichen werden.

[30] Den Autoren ist bewusst, dass gewisse Unterschiede zwischen den international üblichen, auf Shulman (1986, 1987) zurückgehenden Konzeptualisierungen des „Pedagogical Content Knowledge“ (PCK) und dem im deutschsprachigen Raum verwendeten Konstrukt des FDW gibt (z. B. Gramzow et al., 2013; Vollmer & Klette, 2023). Da sich die Analyse auf empirisch-inhaltliche Ergebnisse stützt, wird auf eine genaue Beschreibung der hier zugrundeliegenden theoretischen Modellierungen verzichtet. Ergebnisse zum Forschungsstand werden hier unter dem FDW gelabelt, auch wenn teilweise eher PCK untersucht wurde.

Aussagen über das FDW, die auf Analysen mithilfe quantitativer Globalscores von Bearbeitungen der Testinstrumente basieren, bleiben also inhaltlich recht allgemein und die Gültigkeit über die konkreten Projektkontexte hinaus ist trotz gemeinsamer theoretischer Fundierung ungeklärt, was zusammenfassende Betrachtungen und Implikationen über mehrere Projekte hinweg schwierig macht. Dass Operationalisierung des FDW entsprechend der Natur des Konstrukts in der Regel in (unterschiedliche) fachliche Kontexte / Inhaltsbereiche eingebettet sind[31] erschwert eine Analyse zusätzlich. Die vorliegende Arbeit macht sich daher ein regressionsanalytisches Verfahren (z. B. Woitkowski & Riese, 2017) sowie das Scale-Anchroing-Verfahren (Beaton & Allen, 1992; OECD, 2018) zur Bildung von Niveaumodellen zunutze, um die nicht unmittelbar vergleichbaren quantitativen Aussagen unter Nutzung des vorhandenen Datenschatzes in inhaltlich-kriterienorientierte Beschreibungen zu überführen. Einerseits kann mithilfe solcher Beschreibungen die Vergleichbarkeit der tatsächlich abgebildeten Konstrukte, die durch die in den Projekten jeweils verwendeten Testinstrumente erfasst werden, durch eine Gegenüberstellung eingeschätzt werden. Andererseits können mithilfe der inhaltich-kriterienorientierten Beschreibungen auch inhaltliche Aussagen über Ausprägungen oder sogar Entwicklungsstufen des FDW empirisch fundiert abgeleitet werden, die wiederum differenziertere Einschätzungen der Kenntnisstände von Proband:innen oder Lerngruppen über die bloße Angabe eines Scores hinaus ermöglichen. Solche Einschätzungen würden beispielsweise in einem (Self-) Assessment für Studierende eine Möglichkeit bieten, neben quantitativen Einordnungen auch inhaltliche Lücken wie beispielweise Nachholbedarfe bezüglich konkreter fachdidaktischer Inhalte oder im Kontext konkreter Anforderungs-situationen zu ermitteln. Sowohl die Gültigkeit empirischer Ergebnisse über die konkreten Projektkontexte hinaus als auch eine inhaltliche Einschätzung von Proband:innen sind grundlegend für einen effektiven und nützlichen Transfer der wissenschaftlichen Ergebnisse in die Praxis der Lehramtsausbildung.

Im Kontext des Professionswissens von Lehramtsstudierenden wurden entsprechende Verfahren zur Niveaubildung bereits mit Erfolg angewendet (König, 2009; Schiering et al., 2023; Woitkowski, 2020; Zeller et al., 2022). Hier werden erstmals im deutschsprachigen Raum solche Niveaumodelle genutzt, um die Ergebnisse zur empirisch-inhaltlichen Beschreibung des FDW zweier Projekte vergleichend zu analysieren. Dazu werden hier die Projekte ProfiLe-P+[2] (z. B. Vogelsang et al., 2019) und KiL[25] (z. B. Kleickmann et al., 2014) bzw. dessen Folgeprojekt KeiLa[26] (z. B. Schiering et al., 2023) gemeinschaftlich in den Blick genommen. In beiden Projekten waren Physik-Lehramtsstudierende die primäre Zielpopulation der Untersuchung. Insgesamt werden 1206 Testbearbeitungen (779 aus dem ProfiLe-P+ - Projekt und 427 aus den Projekten KiL / KeiLa) von Physik-Lehramtsstudierenden zum FDW genutzt, um Niveaumodelle mithilfe des Scale-Anchoring-Verfahrens (z. B. Mullis & Fishbein, 2020) und eines regressionsanalytischen Ansatzes (z. B. Nold et al., 2008; Woitkowski & Riese, 2017) auf Basis hierarchischer Komplexität (Commons et al., 1998) entwickeln, welche anschließend zu projektübergreifenden, vergleichenden

[31] In der hier vorliegenden Analyse wurde dabei im ProfiLe-P - Projekt der fachphysikalische Inhalt auf „Mechanik“ fokussiert, während in den Projekten KiL / KeiLa mehrere Fachinhalte (Mechanik, Elektrizitätslehre, Optik, Thermodynamik, Atom- und Kernphysik, spezielle Relativitätstheorie, Festkörperphysik & Quantenmechanik) abgedeckt wurden.

Betrachtungen auf inhaltlicher Ebene genutzt werden.

Diese projektübergreifende Betrachtung soll, wie oben bereits angedeutet, die Verallgemeinerbarkeit bzw. Allgemeingültigkeit etwaiger inhaltlicher Beschreibungen untersuchen. Durch die bisher isoliert stehenden Modellierungen können beispielsweise Untersuchungen der Entwicklung des FDW mithilfe der projektspezifischen Testinstrumente, wie etwa zur Evaluation einer Lehrveranstaltung, keine allgemeingültigen inhaltlichen Aussagen über den Wissenszuwachs der Proband:innen treffen. Es bleibt unklar, ob oder inwieweit ein über beide Projekte äquivalenter Wissenszuwachs auf Basis quantitativer Scores auch ähnliche Zuwächse in der Fähigkeit konkrete Anforderungen zu bewältigen beschreibt. Unter Umständen kann auch aus methodischer Sicht die Vorgehensweise selbst als Vorlage für projektübergreifende Analysen in Fällen dienen, in denen eine direkte gemeinsame quantitative Analyse nicht möglich ist, da sich Testinstrumente und Stichproben unterscheiden bzw. sogar beide disjunkt sind.

Abschließend werden Limitationen und Anwendungsmöglichkeiten der erhaltenen inhaltlichen Beschreibungen von Ausprägungen des FDW diskutiert. Darüber hinaus werden Optionen für weiterführende Forschung erörtert.

4.2. Theoretischer Hintergrund

Das Professionswissen von Lehrkräften wird in der Tradition Shulmans (1986, 1987) üblicherweise in Fachwissen (FW), Pädagogisches Wissen (PW) und FDW gegliedert (Baumert & Kunter, 2006; speziell für das Fach Physik vgl. Riese, 2009). Das FDW wird demnach als dasjenige Wissen aufgefasst, welches zur adressatengerechten Aufbereitung des FW notwendig ist und stellt somit eine zentrale Komponente des Professionswissens dar (Shulman, 1987). Nachfolgend wird das in diesem Beitrag fokussiert betrachtete Konstrukt des FDW aus der Perspektive der Naturwissenschaftsdidaktik präzisiert und in relevante theoretische Rahmungen eingebettet.

4.2.1 Fachdidaktisches Wissen

Die Modellierungen des FDW (im englischsprachigen und internationalen Raum auch „Pedagogical Content Knowledge“, kurz PCK, genannt[30]) unterscheiden sich zwar häufig im Detail (Gess-Newsome, 1999; Hume et al., 2019), gemein ist jedoch allen theoretischen Grundmodellen die o. g. Auffassung von FDW als spezifisches Wissen von Lehrkräften, welches zur adressatengerechten Aufbereitung von Fachwissen notwendig ist und mit den anderen Domänen des Professionswissens (FW & PW) in Beziehung steht (Baumert & Kunter, 2006; Riese, 2009; Shulman, 1986). Dabei gibt es unterschiedliche strukturelle Ansätze, das FDW in der Bandbreite von eher deklarativem Wissen bis hin zu gezeigten Handlungen zu positionieren.

Einen prominenten Ansatz stellt hier das häufig als „Kontinuumsmodell“ bezeichnete Konzept von Blömeke et al. (2015) dar, das Kompetenz als Kontinuum zwischen latenten kognitiven Dispositionen und gezeigter Performanz in für die Profession spezifischen Handlungssituationen beschreibt. Das in Testinstrumenten abrufbare FDW im hier

beschriebenen Sinne lässt sich in diesem Modell eher auf Seite der kognitiven Dispositionen verorten, die wiederum eine Grundlage für situationsspezifische Fähigkeiten und Fertigkeiten darstellen (Blömeke et al., 2015). International speziell im Bereich der Naturwissenschaftsdidaktik etabliert ist darüber hinaus auch das sog. „Refined Consensus Model of PCK" (kurz RCM, Carlson et al. 2019), welches das FDW in die Bereiche *collective* PCK (cPCK), *personal* PCK (pPCK) und *enacted* PCK (ePCK) gliedert (siehe auch Alonzo et al., 2019). Dabei stellt cPCK die kollektive Wissensbasis der fachdidaktischen Community dar, pPCK das explizite Wissen einzelner Akteur:innen und ePCK das internalisierte Wissen, welches sich durch Performanz in spezifischen Situationen äußert. Eine knappe Gegenüberstellung der beiden theoretischen Ansätze des Kontinuumsmodells und des RCMs ist z. B. bei Kulgemeyer et al. (2020, S. 4–7) zu finden. Beide Modelle nehmen dabei an, dass das FDW bzw. PCK eine wichtige Voraussetzung für späteres professionelles Handeln im Klassenzimmer ist.

Hierzulande ist eine Gliederung des FDW in drei Dimensionen üblich (z. B. Gramzow, 2015; Kröger, 2019; Tepner et al., 2012). Dabei wird das FDW grundsätzlich als abhängig vom konkret betrachteten Fachinhalt (Dimension 1) aufgefasst. Im Falle der Physik sind dabei konkrete Inhaltsgebiete wie beispielsweise „Mechanik", „Optik" oder „Elektrizitätslehre" und nicht übergeordnete fachliche Dimensionen wie „Erkenntnisgewinnung" gemeint. Weiterhin umfassen die Modellierungen meist eine Dimension, die unterschiedliche fachdidaktische Inhalte / Facetten (Dimension 2) wie beispielsweise Schülerkognition oder Instruktionsstrategien abbildet. Es existieren zahlreiche Kataloge relevanter Facetten, die u. a. Kirschner (2013) in einer Übersicht gegenübergestellt hat. Dabei ist auffällig, dass die Facetten *Schüler und Schülerkognition*[32] sowie *Instruktions- und Vermittlungsstrategien* fast allen Modellierungen gemein ist. Diese und die weiteren genutzten Facetten werden primär aus den ursprünglichen theoretischen Modellierungen des FDW (z. B. Carlson et al., 2019; Shulman, 1986), Analysen der Curricula der Lehrerbildung bzw. Literatur-Reviews (z. B. Kröger 2019; Gramzow et al. 2013) sowie Expertenbefragungen zu Sicherstellung der curricularen Validität entsprechender Items (z. B. Gramzow 2015) abgeleitet. Auch die Items zu den o. g. Facetten *Schüler und Schülerkognition* und *Instruktions- und Vermittlungsstrategien* wurden in den entsprechenden Befragungen als curricular passend eingeschätzt (Gramzow 2015, S. 166-168). Aus Gründen der Testökonomie und Zumutbarkeit wird bei der Entwicklung konkreter Testinstrumente meist eine Auswahl entsprechender Facetten getroffen. Die dritte Dimension der Itementwicklungsmodelle dient üblicherweise zur Anreicherung der Anforderungsbereiche der Testinstrumente (Klieme et al., 2003). So findet sich bei Tepner et al. (2012) sowie Kröger (2019) eine Dimension „Wissensarten" (S. 19 bzw. 50) und bei Gramzow (2015) eine Dimension „Kognitive Aktivität" (S. 104).

[32] Die Facette wird hier wie im Original benannt und daher nicht geschlechtsneutral umformuliert.

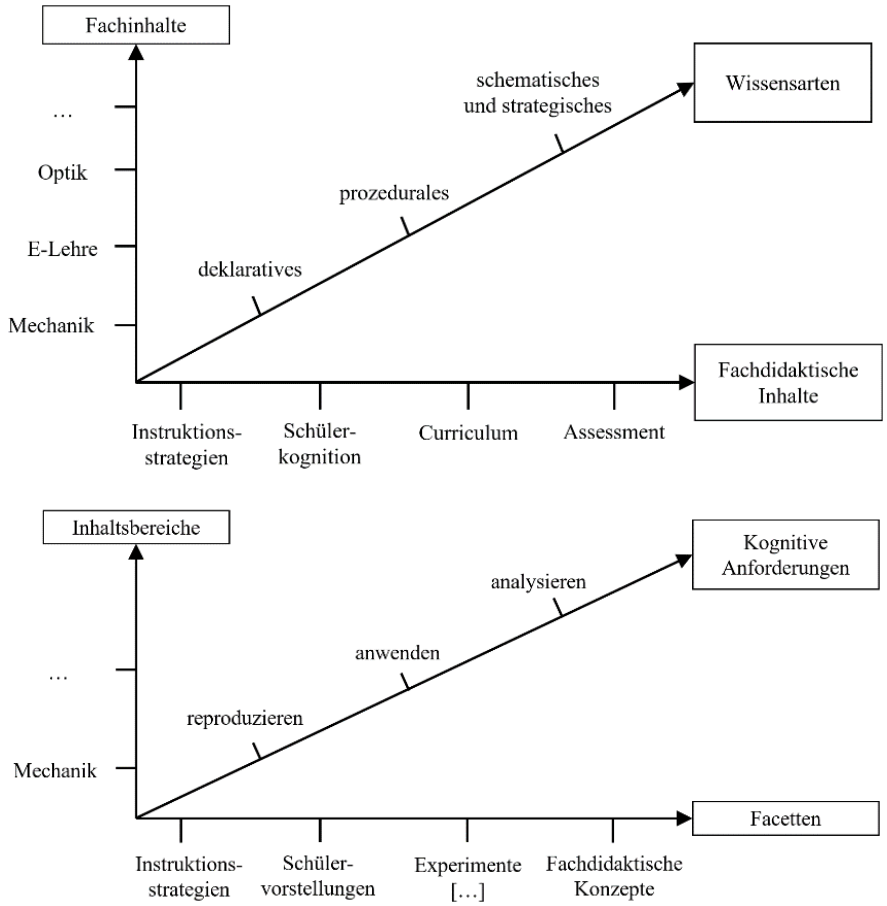


Abbildung 4.1 Itementwicklungsmodelle zu den Testinstrumenten nach Kröger (2019, S. 50) oben und Gramzow (2015, S. 104) unten.

Für die Physik sind hier die Modelle des FDW, die den Testinstrumenten von Kröger (2019) und Gramzow (2015) (zur Itementwicklung) zugrunde liegen, exemplarisch dargestellt (Abbildung 4.1). Auffällig ist auch hier, dass in beiden Modellen jeweils eine Facette zu Schülerkognition und eine Facette zu Instruktionsstrategien enthalten ist. Auch Tepner et al. (2012) schließen in ihrer Dimensionierung, die weitgehend Analog zu der von Kröger (2019) aufgebaut ist, die Facette der Schülervorstellungen explizit mit ein. Die anderen beiden Facetten weichen jedoch voneinander ab. Für die Begründung der Auswahl der entsprechenden Facetten sei auf die Originalquellen (Gramzow, 2015, S. 96–105; Kröger, 2019, S. 46–47; Tepner et al., 2012, S. 13–16) verwiesen.

Speziell für das Fach Physik belegen konkrete Forschungsergebnisse aus Quer- und Längsschnitten signifikante Zuwächse des FDW im Studium und Vorbereitungsdienst (Kirschner, 2013; Kröger, 2019; Riese & Reinhold, 2012). Weiterhin zeigen sich im naturwissenschaftlichen Bereich signifikante Zusammenhänge zwischen FDW und FW bzw. PW (Riese & Reinhold, 2012; Schiering et al., 2019) und Zusammenhänge zwischen FDW und Performanz in prototypischen Anforderungssituationen, wie beispielsweise (1) der Unterrichtsplanung (Behling et al., 2022b; Riese et al., 2022b; Schröder et al., 2020), (2) dem Erklären physikalischer Phänomene (Kulgemeyer et al., 2020; Kulgemeyer & Riese, 2018), (3) dem Reflektieren über Unterricht (Kulgemeyer et al., 2021), (4) der kognitiven Aktivierung

(Förtsch et al., 2016), (5) der Nutzung von physischen Modellen (Förtsch et al., 2018) sowie (6) diagnostischen Handlungen (Kramer et al., 2021). Für den MINT-Bereich wurden zudem (häufig mediative) Einflüsse des FDW auf Aspekte der Unterrichtsqualität bzw. des Unterrichtserfolgs (Behling et al., 2022a, 2022b; Blömeke et al., 2022; Keller et al., 2017) festgestellt. Diese Ergebnisse sind konform zu den theoretischen Annahmen, beispielsweise der angenommenen Notwendigkeit von FDW zur Aufbereitung fachlicher Inhalte bei Shulman (1986). Auch die angenommene Wirkkette der schulischen Bildung nach Terhart (2012) macht diese Ergebnisse plausibel. Somit ist das besondere Augenmerk auf das FDW als wichtige Dimension des Professionswissens sowohl empirisch als auch theoretisch zu rechtfertigen.

Statistische Zusammenhangs- und Mediationsanalysen in der Art der genannten Studien zielen dabei naturgemäß im Wesentlichen auf Schlussfolgerungen auf Basis quantitativer Ausprägungen ab (Reinhold et al., 2017) und treffen dabei keine Aussagen über die (inhaltliche) Art dieser Ausprägungen. In der Folge stellen Mientus et al. (2022) im Rahmen eines systematischen Reviews fest, dass in bisheriger internationaler Forschung zur inhaltlichen Charakterisierung des FDW im MINT-Bereich primär qualitative Untersuchungsmethoden genutzt wurden. Weiterhin beobachten sie, dass quantitative empirische Analysen, wenn auch zur Beantwortung unterschiedlicher Forschungsfragen und Untersuchung unterschiedlicher Zusammenhänge, weitestgehend auf Globaleinschätzungen abzielen.

4.2.2 Kompetenzniveaumodelle

Kompetenzniveaumodelle werden allgemein als geeignetes Mittel zur inhaltlichen Beschreibung von hierarchischen Ausprägungen unterschiedlicher Konstrukte aufgefasst (Beaton & Allen, 1992; Lok et al., 2016) und wurden beispielsweise in den Large-Scale Schulleistungsstudien wie PISA und TIMSS zur inhaltlichen Beschreibung von Fähigkeitsniveaus verwendet (Mullis et al., 2016; OECD, 2018). Die inhaltliche Beschreibung entsprechender Ausprägungen auf Basis quantitativer Daten bietet dabei die Chance, quantitative Ergebnisse und qualitative Beschreibungen zu verbinden. Die Nutzung der Testdaten validierter Testinstrumente stellt hierbei auch ein Validitätsargument für die erhaltenen Niveaumodelle dar. Es existieren unterschiedliche Möglichkeiten, aus Testscores inhaltliche Niveaumodelle abzuleiten, die sich deutlich unterscheiden. Woitkowski (2020) stellt im Rahmen seiner Adaption eines dieser Verfahren eine Übersicht u. a. des Scale-Anchoring-Verfahrens und regressionsanalytischer Ansätze vor. Beide Verfahren nutzen ein IRT[33]-Modell als Ausgangspunkt, mit dem eine gemeinsame Abbildung von Personenfähigkeiten und Aufgabenschwierigkeiten auf eine Skala mit inhärenter Hierarchie ausgenutzt wird, so dass Aufgaben und Personen direkt miteinander in Beziehung gesetzt werden können (siehe z. B. Moosbrugger & Kelava, 2020; Neumann, 2014).

Im Scale-Anchoring-Verfahren wird über mehrere Schritte aus einem IRT-Modell ein inhaltliches Niveaumodell gebildet (Mullis & Fishbein, 2020; OECD, 2018). Dabei werden zunächst Personengruppen mithilfe der Fähigkeits-Verteilungen gebildet (beispielsweise eine

[33] IRT wird als Abkürzung für Item Response Theorie verwendet.

Gruppe mit niedriger, eine mit mittlerer und eine mit hoher Fähigkeit). Anschließend werden die Aufgaben gemäß ihrer Lösungshäufigkeit in den unterschiedlichen Personengruppen wiederum in Gruppen eingeteilt. Die mittleren Schwierigkeitsparameter der Aufgabengruppen dienen dann zur Bildung der Niveaugrenzen, da sie sich durch die Nutzung des IRT-Modells direkt auf die Personenfähigkeiten beziehen lassen. Die inhaltlichen Beschreibungen der Niveaus werden anschließend durch die Aufgaben, deren Schwierigkeitsparameter sich nahe an den Niveaugrenzen befinden, erstellt. Der genaue Ablauf des Verfahrens wird auch in Abschnitt 4.4 noch einmal bei der konkreten Anwendung deutlich. Die Niveaustruktur und die inhaltlichen Niveaucharakterisierungen werden somit vollständig induktiv aus dem Modell abgeleitet, wodurch der qualitative Aufwand sich auf die inhaltliche (Re-)Analyse weniger Aufgaben reduziert. Das Verfahren zeichnet sich dadurch durch vergleichsweise hohe Objektivität und Effizienz aus. Allerdings ist eine möglichst große Anzahl an Aufgaben an den jeweiligen Niveaugrenzen für eine reliable Niveaucharakterisierungen hier optimal. Das Scale-Anchoring-Verfahren wurde bereits mehrfach zur Analyse des FDW im deutschsprachigen Raum eingesetzt (Schiering et al., 2023; Schiering et al., 2019; Zeller et al., 2022). In Niveauanalysen im Kontext anderer Domänen des Professionswissens werden anstelle des Scale-Anchoring-Verfahrens meist stärker theoriegeleitete Ansätze genutzt.

Eine Alternative zum Scale-Anchoring-Verfahren bietet beispielsweise ein regressionsanalytischer Ansatz (Blömeke et al., 2008b; Nold et al., 2008; Woitkowski, 2020). Dazu werden schwierigkeitserzeugende Merkmale aus theoretischen Überlegungen abgeleitet (z. B. sprachliche Terminologie und Komplexität kognitiver Bearbeitungsprozesse bei König, 2009) und die Aufgaben bezüglich dieser Merkmale gruppiert. Anschließend wird mithilfe einer linearen Regression die Varianzaufklärung dieser Gruppierung bzgl. der Aufgabenschwierigkeit bestimmt und somit die Eignung des Modells geprüft. Zeigt das Modell eine ausreichende Passung, können wiederum die mittleren Aufgabenschwierigkeiten durch das IRT-Modell als Niveaugrenzen aufgefasst werden (analog zu den Aufgabengruppen aus dem Scale-Anchoring-Verfahren). Die Niveaucharakterisierungen ergeben sich dann implizit durch die Beschreibung der schwierigkeitserzeugenden Merkmale. Da der regressions-analytische Ansatz die Entwicklung eines Modells für schwierigkeitserzeugende Merkmale und eine (Re-)Analyse aller Aufgaben bzgl. dieser Merkmale erfordert, ist er aufwändiger als das Scale-Anchoring-Verfahren. Auf der anderen Seite können mithilfe des regressionsanalytischen Ansatzes (nach entsprechender theoretischer Vorarbeit) Informationen aus allen Aufgaben und Expertenwissen bzgl. aller Aufgaben zur inhaltlichen Charakterisierung mit herangezogen werden, weshalb dieser Ansatz gerade bei einer geringen Anzahl verfügbarer Aufgaben attraktiv ist. Besonders für eine projektübergreifende Analyse sollte das theoretisch zugrunde gelegte Modell schwierigkeitserzeugender Merkmale unabhängig vom konkreten Testinstrument sein. Im naturwissenschaftsdidaktischen Kontext wurde der regressionsanalytische Ansatz bereits mehrfach bei Fachwissenstests eingesetzt (Bernholt, 2010; Woitkowski, 2019; Woitkowski & Riese, 2017) .

4.2.3 Hierarchische Komplexität des FDW

Bei den in Abschnitt 4.2.2 genannten regressionsanalytischen Ansätzen zur Kompetenzniveau-ermittlung wurde als „schwierigkeitserzeugendes Merkmal“ mehrfach ein Modell

hierarchischer Komplexität der Aufgabenanforderungen angelehnt an das „Model of hierarchical Complexity“ nach Commons et al. (1998) (siehe auch Commons et al., 2014) entwickelt bzw. für das jeweils fokussierte Konstrukt adaptiert. Die hierarchische Komplexität stellt dabei ein Schema dar, nach dem die Qualität von Wissen als propositionales Netzwerk im lernpsychologischen Sinne (z. B. Schnotz, 1994) eingeschätzt werden kann. Der grundlegende Ansatz ist, dass höhere Qualität des Wissens nicht durch bloße Breite, sondern durch den Grad der Vernetzung des Wissensnetzwerks entsteht. Höhere Komplexitätsstufen bauen dabei auf niedrigeren auf, indem sie die Wissensstrukturen dieser niedrigeren Stufen reorganisieren. Es stellt somit einen etablierten, vereinheitlichten Ansatz dar, um die Qualität von Wissensstrukturen in unterschiedlichen Bereichen zu beschreiben (siehe Woitkowski & Riese, 2017).

Das Modell hierarchischer Komplexität wurde also bereits in unterschiedlichen Kontexten erfolgreich genutzt. Es umfasst allgemeine kognitive Prozesse und ist insofern auch für das FDW ein aussichtsreicher Kandidat zur vereinheitlichten Beschreibung schwierigkeitserzeugender Merkmale. Da für das physikalische Fachwissen bereits ein Komplexitätsmodell existiert, welches mit Erfolg zur Modellierung von Niveaustufen genutzt wurde (Woitkowski & Riese, 2017) wäre es zudem wünschenswert die Adaptierbarkeit dieses Modells für das FDW zu überprüfen (siehe Abschnitt 4.4).

4.3. Ziele der Analyse

Die empirisch fundierte inhaltliche Beschreibung von Ausprägungen des FDW z. B. in Form von Niveaumodellen stellt nach wie vor ein Desiderat fachdidaktischer Forschung dar. Eine Möglichkeit der Beschreibung solcher Ausprägungen von Studierenden und Lerngruppen, ist sowohl für individual- als auch systemdiagnostische Zwecke und die Entwicklung oder Auswahl passender Fördermöglichkeiten notwendig. Bisher liegen jedoch von empirischer Seite im deutschsprachigen Raum hauptsächlich quantitative, globale Analysen und Ergebnisse zum FDW vor, in welchen die inhaltliche Komponente weniger fokussiert wurde. Erste entsprechend inhaltlich angereicherte, kriterienorientierte Ergebnisse sind Projekt- bzw. Testinstrument-spezifisch und stehen dadurch zunächst isoliert. Prinzipiell bieten IRT-Modellierungen die Möglichkeit, auch Datensätze zu unterschiedlichen Testinstrumenten zu verbinden, indem Stichproben von Proband:innen die mehrere Testinstrumente bearbeiten haben, gebildet werden oder indem identische Ankeritems in beiden Tests verwendet werden (siehe z. B. Lee & Lee, 2018). Die nachträgliche Erhebung von entsprechenden Normstichproben gestaltet sich aber in der Fachdidaktik aufgrund kleiner Populationsgrößen und schwierigem Zugriff auf geeignete Stichproben meist nicht praktikabel. Eine projektübergreifende inhaltliche Beschreibung von Ausprägungen des FDW ist aber sowohl zur Vergleichbarkeit von gefundenen quantitativen Ausprägungen des FDW unter der Nutzung unterschiedlicher Testinstrumente als auch zur Validierung von Einordnungen von Proband:innen vor dem Hintergrund einzelner Modellierungen notwendig.

Erst seit kurzem wird auch die inhaltliche Beschreibung von Ausprägungen des FDW auf Basis quantitativer empirischer Ergebnisse in den Blick genommen. Dazu wurden erste datenbasierte kriterienorientierte / inhaltliche Beschreibungen von Ausprägungen des FDW im

Rahmen von IRT-Modellierungen entwickelt. Dabei wurde das Scale-Anchoring-Verfahren (Mullis et al., 2016) auf die Daten aus dem KiL - Projekt (Schiering et al., 2019) sowie vorläufigen Daten ($N < 150$) zu einer geschlossenen Version des in ProfiLe-P konzipierten und verwendeten Testinstruments (Kulgemeyer et al., 2023) angewandt (Zeller et al., 2022). Die Ergebnisse dieser Analysen deuteten in beiden Projekten auf übergeordnete Parallelen bzgl. der erhaltenen Niveaustufen hin: In niedrigen Ausprägungen schien sich das FDW vor allem auf reproduktive Aspekte zu beschränken, während in höheren Ausprägungen auch kreative und evaluierende Elemente hinzukamen (Schiering et al., 2019, S. 224; Zeller et al., 2022, S. 770). Um diese Beobachtung weiter zu systematisieren und ggf. zu bestätigen, soll in diesem Beitrag eine erweiterte Niveauanalyse der Daten aus den KiL / KeiLa - Projekten von Schiering et al. (2023) mit einer Re-Analyse des ProfiLe-P+ - Datensatzes im Rahmen von Niveaumodellierungen inhaltlich verglichen werden. Dieses Vorgehen kann sich unter Umständen als Vorlage für ähnliche projektübergreifende Betrachtungen in anderen verwandten Felder erweisen.

Ziel dieses Beitrags ist also erstens die datengestützte kriterienorientiert-inhaltliche Beschreibung von Ausprägungen des FDW, um damit zweitens die Verknüpfung der Ergebnisse zweier unabhängiger Large-Scale Studien (für fachdidaktische Größenordnungen) auf Basis entsprechender inhaltlicher Ergebnisse zu ermöglichen. Dazu werden die folgenden Forschungsfragen formuliert:

> ***FF1****:* Inwieweit lassen sich mithilfe des Scale-Anchoring-Verfahrens projektübergreifend inhaltliche Strukturen des FDW identifizieren und inhaltlich charakterisieren?
>
> ***FF2****:* Inwieweit lassen sich Stufen hierarchischer Komplexität des FDW projekt-übergreifend identifizieren und inhaltlich charakterisieren?

Zunächst wird dazu analog zum Vorgehen von Schiering et al. (2023) das Scale-Anchoring-Verfahren auf den ProfiLe-P+ - Datensatz angewendet. Der inhaltliche Vergleich der Ergebnisse findet dann durch eine Gegenüberstellung der erhaltenen Niveaubeschreibungen statt. Anschließend wird ein Modell hierarchischer Komplexität für das FDW zur Niveaubildung mithilfe eines regressionsanalytischen Ansatzes ausgehend vom ProfiLe-P+ - Datensatz vorgeschlagen und die Übertragbarkeit auf die KiL / KeiLa - Daten untersucht. Es wird dabei in den Blick genommen, ob mit den Scale-Anchoring-Analysen erhaltene inhaltliche Parallelen sich durch ein solches Modell hierarchischer Komplexität unterstützen, erweitern oder erklären lassen. Etwaige projektübergreifende Strukturen bieten einerseits Potentiale für die Nutzung als Grundlage für Feedback im Rahmen der Lehrpraxis, andererseits erweitern sie den Forschungsstand um allgemein zutreffende Aussagen über Ausprägungen des FDW.

4.4. Methoden

Zur Beantwortung der Forschungsfragen werden das Scale-Anchoring Verfahren und ein regressionsanalytischer Ansatz zur Niveaubildung synchron auf die Daten der beiden Projekte angewandt. Im Falle des Scale-Anchroing Verfahrens findet die projektübergreifende Analyse

durch die gemeinsame vergleichende Betrachtung der erhaltenen Niveauformulierungen statt. Die regressionsanalytische Betrachtung fußt auf einem zu diesem Zweck entwickelten Modell hierarchischer Komplexität für das FDW. Die projektübergreifende Analyse findet hierbei durch die Überprüfung der Anwendbarkeit des Komplexitätsmodells auf beide Datensätze statt. Beide in dieser Analyse verwendete Operationalisierungen lassen sich vor dem Hintergrund des RCM im Rahmen des pPCK, d. h. dem „testbaren" persönlichen FDW der Proband:innen, interpretieren (siehe Riese et al., 2022b für ProfiLe-P sowie Schiering et al., 2023 für KiL / KeiLa).

Sowohl das Scale-Anchoring-Verfahren als auch der regressionsanalytische Ansatz basieren auf einem IRT-Modell des jeweiligen Datensatzes. Für die KiL / KeiLa - Daten wurde dasselbe IRT-Modell wie bei Schiering et al. (2023) verwendet. Für die ProfiLe-P+ - Daten wurde nach einer Bereinigung des Datensatzes ein neues IRT-Modell erstellt. In beiden Fällen wurde dabei das Paket „Test Analysis Modules" (Robitzsch et al., 2024) auf Basis der Statistik-Software R (R Core Team, 2024) verwendet. Der Workflow der Analysen ist in Abbildung 4.2 dargestellt.

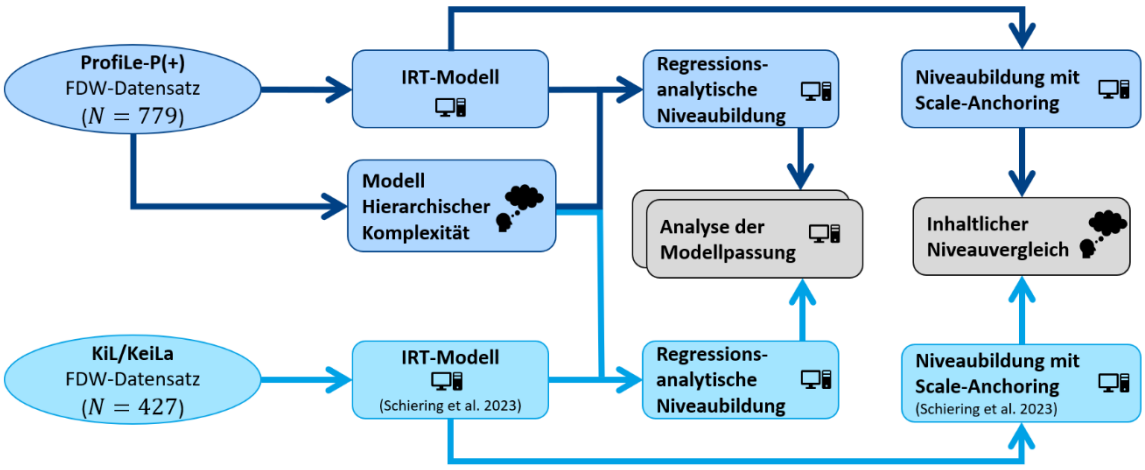


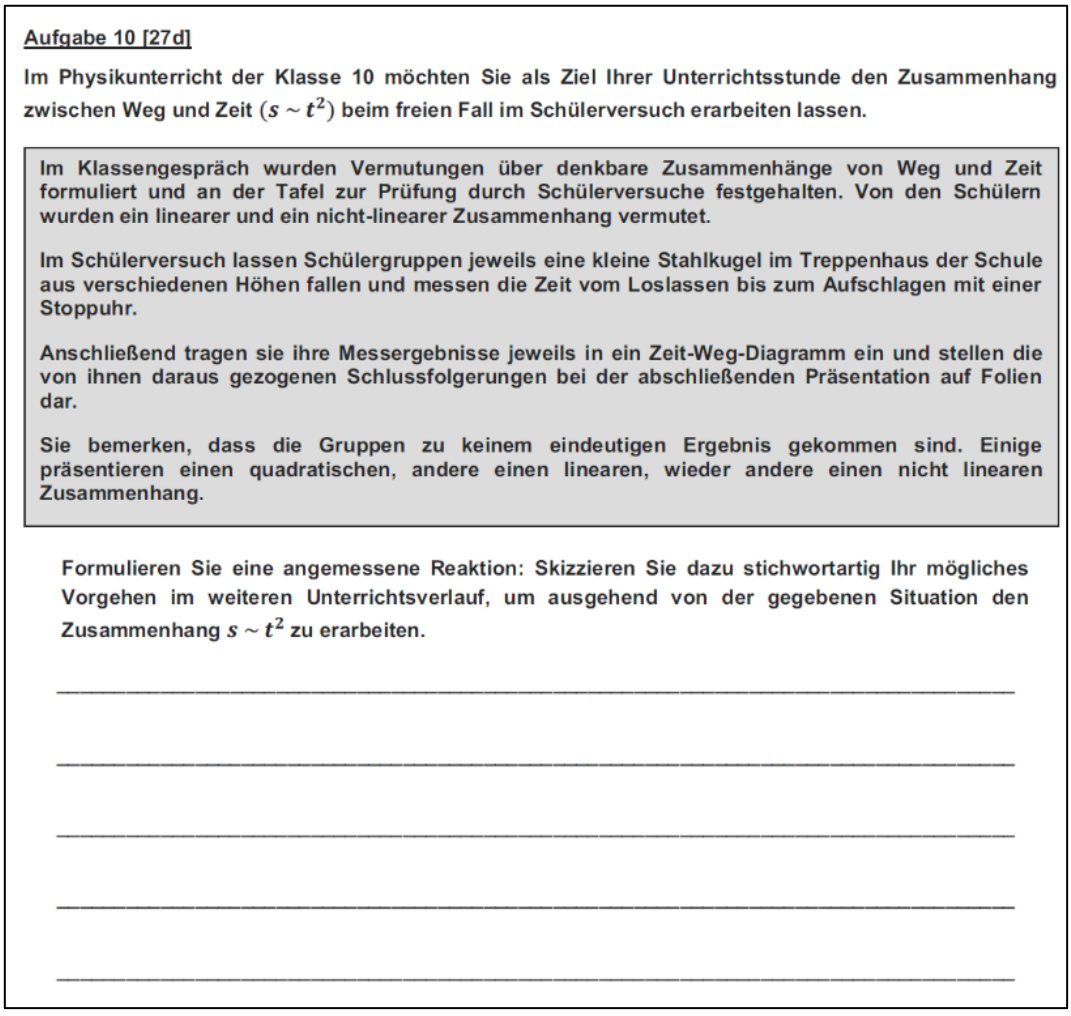
Abbildung 4.2 Analyse-Workflow der vorgestellten Untersuchung.

4.4.1 Testinstrumente und Stichproben

Der Datensatz des ProfiLe-P+ - Projekts (Vogelsang et al., 2019) beinhaltet 846 Bearbeitungen des FDW-Testinstruments nach Gramzow (2015), das FDW in den Facetten *Schülervorstellungen*, *Fachdidaktische Konzepte*, *Experimente und Vermittlung eines angemessenen Wissenschaftsbegriffs* sowie *Instruktionsstrategien* abbildet. Beschreibungen des inhaltlichen Verständnisses dieser Facetten haben Riese et al. (2017, S. 103–104) knapp zusammengefasst. Bezüglich des fachphysikalischen Inhalts wurde sich im ProfiLe-P - Projektverbund übergreifend auf die Mechanik festgelegt, um zu diesem Bereich empirisch trennbare Teilskalen auf Facettenebene erfassen zu können (Riese et al., 2015). Insgesamt besteht das Testinstrument aus 20 offenen und 4 geschlossenen (Multiple-Choice) Aufgaben und wurde im Rahmen des ProfiLe-P+ - Projekts in den Jahren 2016 bis 2019 von Bachelor- und Masterstudierenden des Physik-Lehramts aus 12 deutschsprachigen Universitäten

bearbeitet. Ein Beispielitem aus diesem Testinstrument ist in Abbildung 4.3 dargestellt. Aus diesen Erhebungen blieben nach einer intensiven Bereinigung der Daten und dem Ausschluss von unvollständigen Bearbeitungen 779 Bearbeitungen (34 % weiblich, Studienjahr $M = 2{,}11$, $SD = 1{,}75$) für die hier verwendete Modellierung.

In den Projekten KiL und KeiLa wurde ein FDW-Testinstrument (Kröger, 2019; Sorge et al., 2019) eingesetzt, welches FDW im Rahmen der fachdidaktischen Inhalte (analoge Dimension zu den „Facetten“ in ProfiLe-P+) *Schülerkognition*, *Instruktionsstrategien*, *Curriculum* und *Assessment* abbildet. Das inhaltliche Verständnis dieser Aspekte führt Kröger (2019, S. 46–47) genauer aus. Es wurde darauf abgezielt, das FDW bzgl. der fachlichen Inhalte breit zu untersuchen und somit die fachphysikalischen Inhalte Mechanik, Elektrizitätslehre, Optik, Thermodynamik, Atom- und Kernphysik, spezielle Relativitätstheorie, Festkörperphysik sowie Quantenmechanik eingeschlossen.



Aufgabe 10 [27d]

Im Physikunterricht der Klasse 10 möchten Sie als Ziel Ihrer Unterrichtsstunde den Zusammenhang zwischen Weg und Zeit ($s \sim t^2$) beim freien Fall im Schülerversuch erarbeiten lassen.

Im Klassengespräch wurden Vermutungen über denkbare Zusammenhänge von Weg und Zeit formuliert und an der Tafel zur Prüfung durch Schülerversuche festgehalten. Von den Schülern wurden ein linearer und ein nicht-linearer Zusammenhang vermutet.

Im Schülerversuch lassen Schülergruppen jeweils eine kleine Stahlkugel im Treppenhaus der Schule aus verschiedenen Höhen fallen und messen die Zeit vom Loslassen bis zum Aufschlagen mit einer Stoppuhr.

Anschließend tragen sie ihre Messergebnisse jeweils in ein Zeit-Weg-Diagramm ein und stellen die von ihnen daraus gezogenen Schlussfolgerungen bei der abschließenden Präsentation auf Folien dar.

Sie bemerken, dass die Gruppen zu keinem eindeutigen Ergebnis gekommen sind. Einige präsentieren einen quadratischen, andere einen linearen, wieder andere einen nicht linearen Zusammenhang.

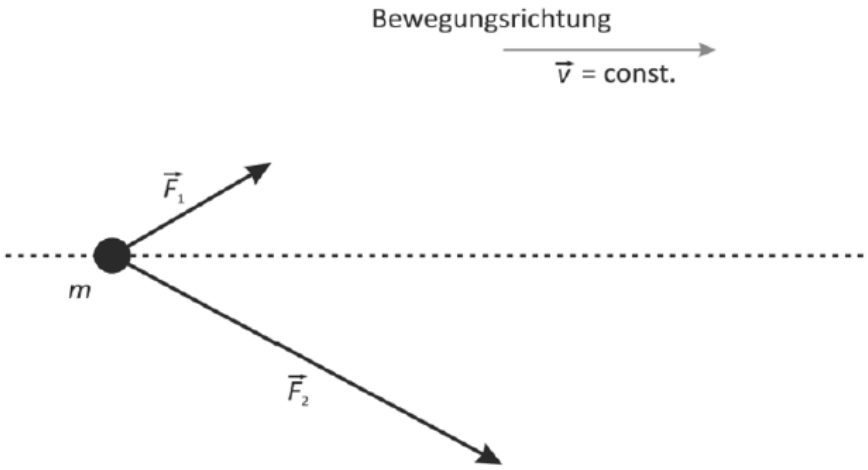
Formulieren Sie eine angemessene Reaktion: Skizzieren Sie dazu stichwortartig Ihr mögliches Vorgehen im weiteren Unterrichtsverlauf, um ausgehend von der gegebenen Situation den Zusammenhang $s \sim t^2$ zu erarbeiten.

Abbildung 4.3 Beispielitem aus dem FDW-Testinstrument des ProfiLe-P+ - Projekts (Gramzow, 2015, S. 235).

Das Testinstrument besteht insgesamt aus 18 offenen und 21 geschlossenen Aufgaben. Ein Beispielitem aus diesem Testinstrument ist in Abbildung 4.4 dargestellt. Der Datensatz des KiL / KeiLa - IRT-Modells besteht insgesamt aus 200 Bearbeitungen dieses Testinstruments aus der Querschnitterhebung des KiL - Projekts (2013, 12 Universitäten) und 227

Bearbeitungen aus den Längsschnitterhebungen des KeiLa - Projekts (2014 bis 2017, 20 Universitäten)[34].

Schülerinnen und Schülern fällt es oft schwer, die Newtonschen Axiome zur Lösung konkreter Aufgaben anzuwenden.
Betrachten Sie die folgende Situation: Ein kleiner Körper der Masse *m* bewegt sich reibungsfrei und mit konstanter Geschwindigkeit $\vec{v}$ nach rechts. Auf den Körper wirken dabei drei Kräfte. Zwei davon sind eingezeichnet. Sie bitten die Schülerinnen und Schüler, die dritte Kraft einzuzeichnen.



Welche physikalisch falsche Antwort würden Sie von den Schülerinnen und Schülern erwarten?

Mögliche korrekte Antworten:
[Kraftpfeil in Bewegungsrichtung] – Es muss eine Kraft für die Bewegung verantwortlich sein.
[Kraftpfeil als Summe von F_1 und F_2] – Reflexartiges Zeichnen eines Kräfteparallelogramms.
[Kraftpeil als Verlängerung von F_1] – Summe aller Kräfte muss in Bewegungsrichtung zeigen.

Abbildung 4.4 Beispielitem aus dem FDW-Testinstrument des KiL – Projekts (Schiering et al., 2019, S. 225).

4.4.2 Item-Response-Modellierungen

Um möglichst vergleichbare Niveaumodelle zu konstruieren, wurde bereits bei der IRT-Modellierung ein analoges Vorgehen zu der bereits bestehenden Analyse von Schiering et al. (2023) gewählt. Aufgrund der für die Anwendung von Niveaubildungsverfahren vergleichsweise geringen Aufgabenanzahl wurde ein eindimensionales Partial-Credit-Modell (Masters, 1982) verwendet, wobei Thurstone-Thresholds zur Schätzung der Itemschwierigkeiten bei polytomen Aufgaben verwendet wurden (Linacre, 1998). Zur gemeinsamen Modellierung wurden Datensätze, die derselben Person sind, im Rahmen der Methode virtueller Proband:innen (Davier et al., 2008) als unabhängige Datensätze modelliert, d. h. jede Bearbeitung fließt in die Modellierung als eigene „Datenzeile“ ein, ohne dass weiter beachtet wird, dass es sich um dieselbe Person handelt. Das erhaltene Modell für die ProfiLe-

[34] Eine ausführlichere Beschreibung der Stichproben der Projekte KiL und KeiLa kann in Schiering et al. (2023, S. 8) gefunden werden.

P+ - Daten wies mit einer EAP-Reliabilität von 0.71 und Item-Outfits im Bereich von 0.8 bis 1.2 hinreichende Fit-Qualität für die weitere Analyse auf.

Für die Daten der KiL / KeiLa - Projekte wurde das bereits bestehende IRT-Modell von Schiering et al. (2023) basierend auf 427 Bearbeitungen herangezogen. Auch hier waren die Fit-Gütekriterien mit einer EAP-Reliabilität von 0.72 und Item-Outfits ebenfalls im Bereich von 0.8 bis 1.2 zufriedenstellend.

4.4.3 Scale-Anchoring-Verfahren

Zur Beantwortung der ersten Forschungsfrage wurde das Scale-Anchoring-Verfahren (z. B. Mullis et al., 2016) auf das IRT-Modell der ProfiLe-P+ - Daten angewendet. Im ersten Schritt wurden dazu die Item- und Personenparameter gemeinsam auf eine praktikablere Skala mit Mittelwert 500 und Standardabweichung 100 transformiert. Anschließend wurden drei Probandengruppen durch eine äquidistante Zerlegung der Fähigkeitsskala gebildet (Abbildung 4.5). Zur absichernden Kontrolle, dass die so gefundenen Gruppen ausreichend unterschiedlich (Woitkowski & Riese, 2017) waren, wurden inferenzstatistische Betrachtung mithilfe verteilungsfreier Tests (Kruskal-Wallis und Mann-Whitney U Tests) nach dem Vorbild von (Schiering et al., 2023) durchgeführt, die eine ausreichende Differenzierung der Gruppen bestätigten (Tabelle 4.1).

Auf Basis dieser Probandengruppen wurden die Aufgaben analog zum von Schiering et al. (2023 adaptiert nach Mullis & Fishbein, 2020) genutzten Schema in Aufgabengruppen eingeteilt:

1. Aufgabengruppe 1: Mehr als 55 % der Personen aus Personengruppe 1 haben die Aufgabe gelöst.
2. Aufgabengruppe 2: Mehr als 55 % der Personen aus Personengruppe 2 und weniger als 50 % der Personen aus Personengruppe 1 haben die Aufgabe gelöst.
3. Aufgabengruppe 3: Mehr als 55 % der Personen aus Personengruppe 3 und weniger als 50 % der Personen aus Personengruppe 2 haben die Aufgabe gelöst.
4. Aufgabengruppe 3+: Weniger als 50 % der Personen aus Personengruppe 3 haben die Aufgabe gelöst.

Die Mittelwerte der Schwierigkeitsparameter der Aufgabengruppen dienten dann als Schätzungen für die empirischen Niveaugrenzen. Auch hier wurden, um eine Vergleichbarkeit zu Schiering et al. (2023) beizubehalten, anschließend an die Zuordnung der Aufgaben verteilungsfreie statistische Tests zur Überprüfung der Unterscheidbarkeit der Aufgabengruppen durchgeführt (Tabelle 4.2). Dabei wurde zudem das Abstandskriterium überprüft, d. h. es wurde getestet, ob eine Person mit einem Fähigkeitsparameter, der der Niveaugrenze des Niveaus n entspricht, einer Aufgabe an der Niveaugrenze des Niveaus $n + 1$ mit einer Wahrscheinlichkeit von maximal 30 % (Beaton & Allen, 1992) löst. Zur inhaltlichen Charakterisierung der Niveaus wurden diejenigen Aufgaben herangezogen, die sich nahe bei den Niveaugrenzen befinden.

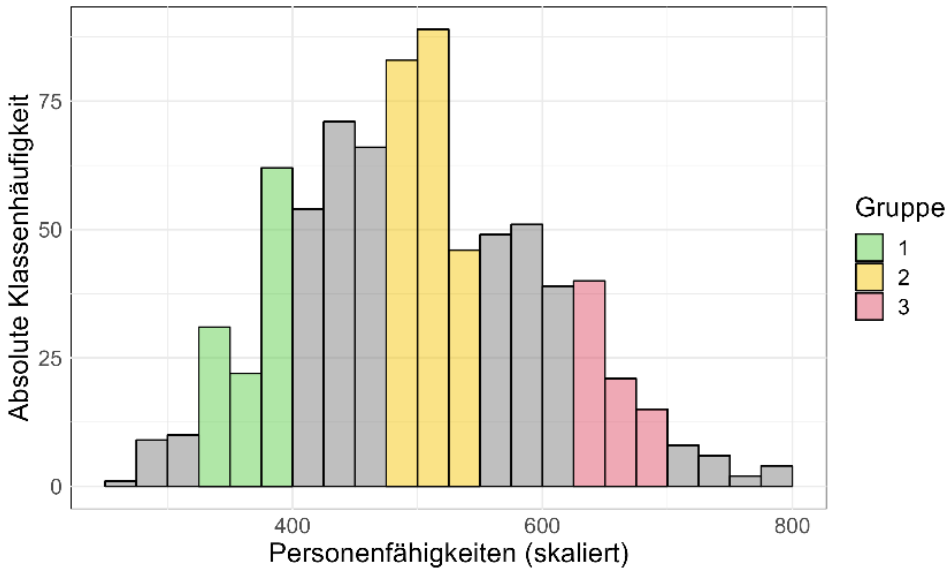


Abbildung 4.5 Personengruppen aus dem ersten Schritt des Scale-Anchoring-Verfahrens (ProfiLe-P+ - Daten). Die Personengruppen wurden als äquidistante Abschnitte der (skalierten) Fähigkeitsparameter gewählt. Das Scale-Anchoring Verfahren erwies sich als robust gegenüber leichter Verschiebungen dieser Abschnitte.

Tabelle 4.1 Beschreibung der Personengruppen aus dem ersten Schritt des Scale-Anchroing-Verfahrens (Profile-P+ - Daten). Ein Kruskal-Wallis Test bestätigte signifikante Gruppenunterschiede ($\chi^2\ (2) = 335, p < 0.001$). In der Tabelle sind anschließend paarweise Post-Hoc Mann-Whitney U Tests berichtet.

Gruppe	**Fähigkeitsspanne**	***N***	***M***	**SD**	**Differenz und *p*-Wert**
1	325 – 400	115	370	23	140 ($W = 0, p < 0.001$)
2	475 – 550	218	510	19	143 ($W = 0, p < 0.001$)
3	625 – 700	76	653	22	

Tabelle 4.2 Beschreibung der Aufgabengruppen aus dem zweiten Schritt des Scale-Anchroing-Verfahrens (Profile-P+ - Daten). Ein Kruskal-Wallis Test bestätigte signifikante Gruppenunterschiede ($\chi 2\ (3) = 29, p < 0.001$). In der Tabelle sind anschließend paarweise Post-Hoc Mann-Whitney U Tests berichtet. Dabei ist der Vergleichstest für die Aufgabengruppen 1 und 2 hier nur der Vollständigkeit halber angegeben, da er aufgrund der geringen Aufgabenanzahl in Aufgabengruppe 1 nicht sinnvoll interpretierbar ist - hier ist $p = 0.096$ bereits der „minimal erreichbare“ p-Wert beim Vergleich zweier Gruppen mit 2 und 5 Elementen.

Aufgabengruppe	***N***	***M***	**SD**	**Differenz und *p*-Wert**	***P* Abstandskriterium**
1	2	-1.57	0.24	1.06 ($W = 0, p = 0.096$)	0.26
2	5	-0.51	0.24	0.84 ($W = 2, p < 0.001$)	0.30
3	13	0.32	0.41	1.52 ($W = 2, p < 0.001$)	0.18
3+	14	1.85	0.78		

Die Ergebnisse der Anwendung des Scale-Anchoring-Verfahrens beider Projekte sind in Abbildung 4.6 und Abbildung 4.7 und dargestellt. Die sich aus diesen Ergebnissen ergebenden inhaltlichen Niveaubeschreibungen und deren Gegenüberstellung werden in Abschnitt 4.5.1 vorgestellt.

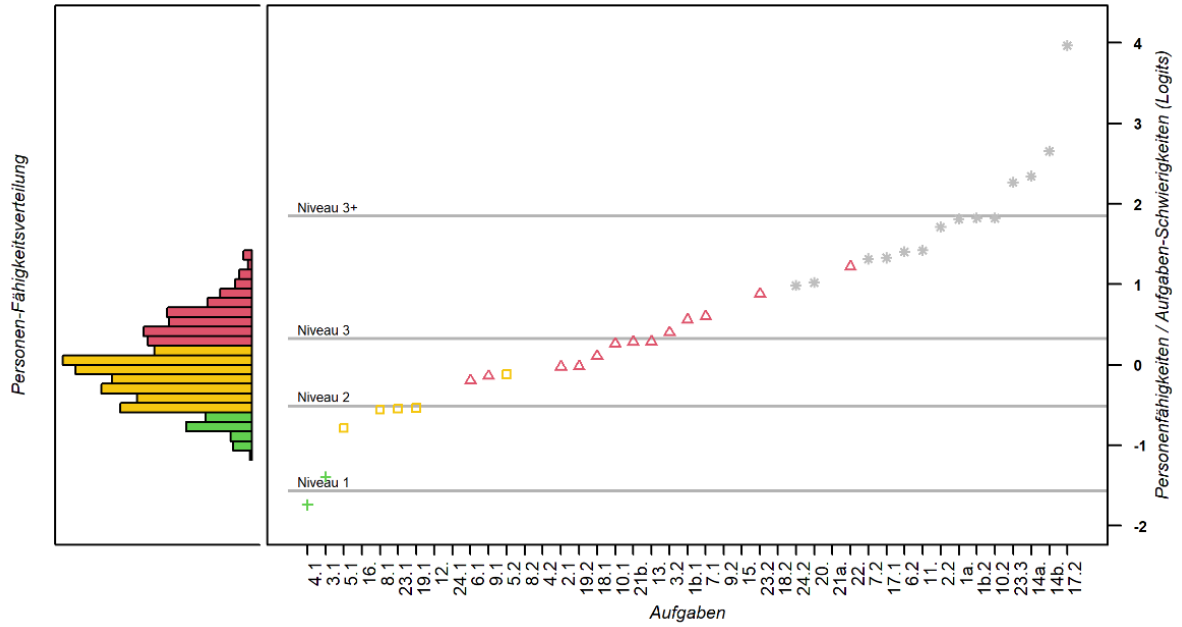


Abbildung 4.6 Finale Wright-Map mit Ergebnissen des Scale-Anchoring-Verfahrens (ProfiLe-P+ - Daten).

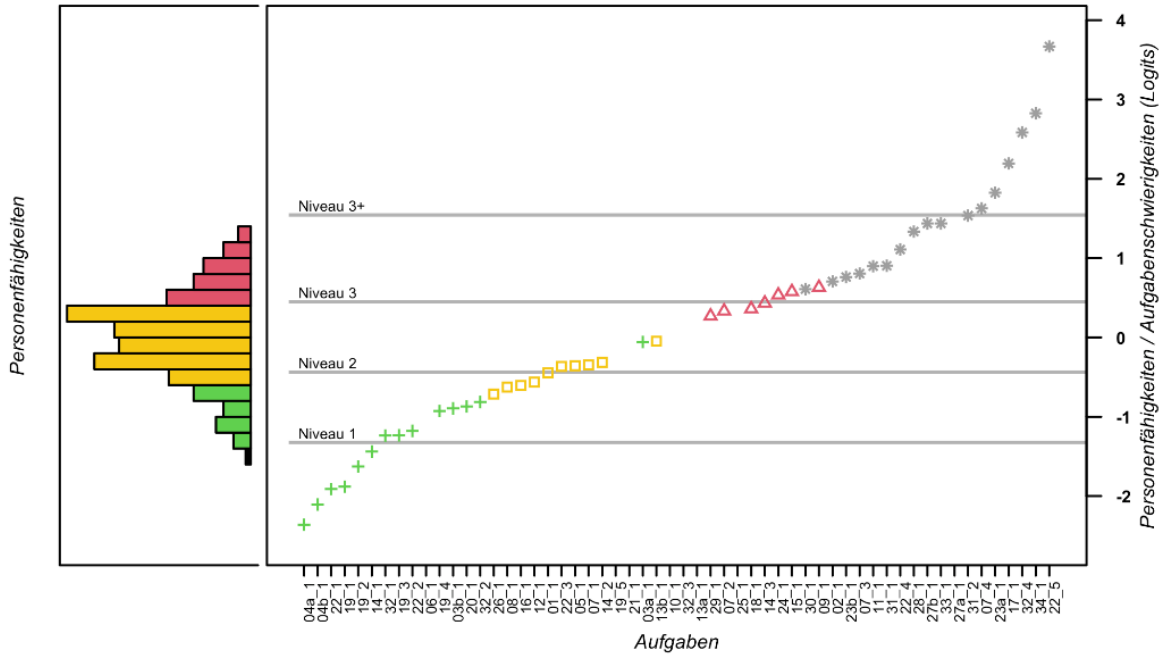


Abbildung 4.7 Finale Wright-Map mit Ergebnissen des Scale-Anchoring-Verfahrens (KiL/KeiLa) nach Schiering et al. (2023, S. 15).

4.4.4 Regressionsanalytisches Verfahren auf Basis eines Modells hierarchischer Komplexität des FDW

In der Naturwissenschaftsdidaktik zeigen Ansätze wie die bereits genannten Analysen von Bernholt (2010) sowie Woitkowski und Riese (2017), dass das Modell der hierarchischen Komplexität nach Commons et al. (1998) geeignet sein kann, Niveaustufen im Fachwissen auf Basis theoretischer Überlegungen zu definieren und erklären. In einem weiteren Analyseschritt wurde daher überprüft, ob und inwieweit sich die gefundenen Gemeinsamkeiten in den Niveaumodellen des FDW mithilfe eines Modells hierarchischer Komplexität untermauern, erklären und ggf. erweitern lassen.

Zu diesem Zweck wurde zunächst ein Modell hierarchischer Komplexität für das FDW entwickelt. Dazu wurden die bereits genannten Arbeiten zur Entwicklung von hierarchischen Komplexitätsmodellen für das Fachwissen von Woitkowski (2015) bzw. Woitkowski und Riese (2017) auf das FDW übertragen. Über mehrere Iterationen hinweg wurde das in Tabelle 4.3 beschriebene 3-stufige Modell ausgearbeitet. Die Stufen *„(I) Fakten“* und *„(II) Einstufige Kausalität“* (Tabelle 4.3) umfassen die bloße Reproduktion sowie die Verknüpfung einzelner Wissenselemente und sind weitgehend analog zu den Stufen *„(I) Fakten“* und *„(III) Lineare Kausalität“* des Komplexitätsmodells nach Woitkowski und Riese (2017, S. 41) angelegt. Die Stufe *„(II) Prozessbeschreibungen“* von Woitkowski und Riese (2017) ließ sich auf das FDW in der operationalisierten Form nicht übertragen, da für das FDW weniger „Prozesse“ im Sinne eines zeitlichen Ablaufs als vielmehr Ursache-Wirkungs-Argumentationen im Zentrum stehen. Daher wird die Stufe der Prozessbeschreibungen in die Einstufige Kausalität integriert (siehe Tabelle 4.3). Die höchste hier betrachtete Komplexitätsstufe stellt somit die Stufe *„(III) Mehrstufige Kausalität“* dar. Sie tritt an die Stelle der Stufe *„(IV) Multivariate Interdependenz“* des Fachwissensmodells und umfasst mehrstufige Argumentationsstränge. Wir argumentieren, dass es sich bei mehrstufigen Argumentationen um eine substanziell höhere Anforderungsstufe im Sinne des Modells hierarchischer Komplexität handelt, als bei einstufigen Argumentationen, da hier mehrere mentale Schemata miteinander in Beziehung gesetzt werden müssen und diese Beziehungen wiederum voneinander abhängig sind.

Um die Passung dieses Komplexitätsmodells zu den empirischen Daten zu testen, wurden die Aufgaben der jeweiligen Testinstrumente zunächst disjunkt zu den Komplexitätsstufen zugeordnet. Dies geschah durch die Analyse der jeweiligen Aufgabe vor dem Hintergrund der in Tabelle 4.3 beschriebenen Komplexitätsstufen. Leitfragen der Zuordnung waren:

1. Erfordert die Aufgabe lediglich die Reproduktion von Fakten? (→ Fakten)
2. Erfordert die Aufgabe die Analyse eines komplexeren Elements (z. B. beschriebene Unterrichtssituation, Dialog, Zeichnung)? (→ einstufige Kausalität)
3. Erfordert die Aufgabe die Kreation eines komplexeren Elements (z. B. Beschreibung eines Experiments, Beschreibung einer Handlungsoption)? (→ einstufige Kausalität)
4. Erfordert die Aufgabe mehrere Schritte im Sinne der Frage 2 und / oder Frage 3? (→ mehrstufige Kausalität)

Beide dargestellten Beispielaufgaben (Abbildung 4.3 und Abbildung 4.4) werden somit der

mehrstufigen Kausalität zugeordnet. In der ProfiLe-P - Aufgabe muss zunächst eine beschriebene Unterrichtssituation analysiert werden, um auftretende Problemstellen zu identifizieren und anschließend müssen darauf aufbauend geeignete Handlungsoptionen generiert werden, um diese Probleme zu bewältigen[35]. In der KiL / KeiLa - Aufgabe muss im ersten Schritt eine komplexe Schüleraufgabe analysiert (und dabei mutmaßlich auch selbst gedanklich korrekt gelöst) werden und im zweiten Schritt davon ausgehend eine typische falsche Lösung mithilfe des Wissens über Schülervorstellungen generiert werden[36].

Tabelle 4.3 Dreistufiges Modell hierarchischer Komplexität für das FDW. Die Charakterisierung diente als Grundlage für die Einordnung der Testaufgaben in das Komplexitätsmodell und wurde an die jeweiligen Rater gegeben.

(I) Fakten

- Reproduktion einzelner, unverbundener Informationen
- Keine oder kaum Bezugnahme auf Situation oder sonstige Beschreibung
- Keine oder kaum Verknüpfung der genannten Informationen
- *Beispiel:* Nennen von Fakten zu einem Fachdidaktischen Konzept

(II) Einstufige Kausalität

- Verknüpfung von zwei oder mehr Fakten, Informationen oder Äußerungen zu einem Produkt (z. B. Schlussfolgerungen, Argumentationen)
- Begründungen, Analysen und Argumentationen mir nur einer Argumentations- / Analysestufe
- *Beispiel:* (einstufige) Analyse oder Evaluation einer Situation

(III) Mehrstufige Kausalität

- Begründungen, Argumentationen, Evaluationen mit mehr als einer Argumentations- / Analysestufe
- Alle Anforderungen, die komplexere Analysen / Argumentation verlangen als II
- *Beispiel:* Analyse und Evaluation einer Situation

Diese Zuordnung wurde pro Testinstrument durch zwei Personen durchgeführt. Die Beurteilerübereinstimmung betrug beim ProfiLe-P - Testinstrument $\kappa = 0.86$ und beim KiL / KeiLa - Testinstrument $\kappa = 0.82$. Uneinigkeiten wurden durch eine kommunikative Validierung (Steinke, 1999) geklärt, sodass für beide Testinstrumente eine Konsens-Aufgabenzuordnung vorlag. Tabelle 4.4 zeigt die Anzahl an Aufgaben pro Komplexitätsstufe nach Projekt getrennt. Diese Zuordnung wurde anschließend genutzt, um mithilfe einer

[35] Eine „analoge“ Aufgabe in der einstufigen Kausalität wäre beispielsweise die reine Kreation eines Unterrichtsverlaufs zum Fallgesetz.

[36] Eine „analoge“ Aufgabe in der einstufigen Kausalität wäre dies beispielsweise dann, wenn eine typisch falsche Lösung aufgrund von Schülervorstellungen bereits eingezeichnet wäre und lediglich die zugehörige Schülervorstellung identifiziert werden müsste.

linearen Regression der Aufgaben-Schwierigkeitsparameter gegen die Aufgabenzuordnung zum Komplexitätsmodell die Passung auf die jeweiligen Datensätze und somit die „Gültigkeit" des Komplexitätsmodells für die jeweils abgebildeten Konstrukte einzuschätzen (Abschnitt 4.5.2).

Tabelle 4.4 Anzahl an Aufgaben in den Komplexitätsstufen nach Projekt getrennt. Die Gesamtaufgabenanzahl weicht hier für beide Testinstrumente von den in Abschnitt 4.4.1 ab, da Punkteschwellen (z. B. 1 vs. 2 Punkte) im Rahmen der Partial-Credit Modellierung getrennt wurden.

Komplexitätsstufe	***N* Profile-P**	***N* KiL/KeiLa**
I – Fakten	13	12
II – Einstufige Kausalität	23	34
III – Mehrstufige Kausalität	7	10

4.5. Ergebnisse

4.5.1 Scale-Anchoring-Verfahren: Niveauformulierungen und Vergleich

Der zentrale Gegenstand des Scale-Anchoring-Verfahrens ist die erhaltene Wright-Map mit den entsprechenden Zuordnungen und Werten (Abbildung 4.6 und Abbildung 4.7) Für beide Datensätze zeigt sich hier ein vergleichsweise homogenes Bild, d. h. die Aufgabengruppen zerfasern nicht stark über die Schwierigkeitsspanne hinweg. Gleichzeitig zeigen die statistischen Betrachtungen (Tabelle 4.1 und Tabelle 4.2 sowie Schiering et al., 2023, S. 14–15) die empirische Trennbarkeit der Stufen. Im Falle des ProfiLe-P+ - Modells erkennt man, dass das Testinstrument vergleichsweise schwierig für die Zielgruppe ist. Dementsprechend stehen für die Charakterisierung der unteren Niveaus nur wenige Aufgaben zur Verfügung, was die spätere Interpretation erschwert. Die Niveauformulierungen auf Basis der Aufgaben nahe der entsprechenden Niveaugrenzen sind in Tabelle 4.5 zusammengefasst, wobei eine Loslösung vom fachlichen Inhalt der jeweiligen Aufgabe hier vorerst nicht forciert wurde, da allgemein eine Abhängigkeit des FDW vom jeweils nötigen FW angenommen wird.

Für die projektübergreifende Analyse werden die erhaltenen Niveaustufen aus beiden Datensätzen verglichen. Es zeigen sich keine auffälligen Parallelen in den fachlichen und fachdidaktischen Inhalten. Demgegenüber sind allerdings Gemeinsamkeiten der Niveaubeschreibungen bzgl. der auftretenden lernpsychologisch interpretierbaren Operatoren (Tabelle 4.6) auffällig. In den niedrigen Niveaus 1 und 2 treten primär Operatoren, welche reproduktive Aspekte beschreiben (grün in Tabelle 4.6), auf. In den höheren Niveaus kommen Operatoren, die kreative (gelb in Tabelle 4.6) und bewertende (rot in Tabelle 4.6) Aspekte beschreiben, hinzu. Es zeigt sich eine deutliche Parallele bezüglich des Auftretens dieser Operatoren auf den jeweiligen Niveaus.

Tabelle 4.5 Gegenüberstellung der Scale-Anchoring Niveauformulierungen der Profile-P+ - und KiL/KeiLa - Modelle. Die jeweiligen Aufgaben, auf die sich der Aspekt bezieht, sind in Klammern mit angegeben.

	Profile-P+	**KiL/KeiLa** (Übers. nach Schiering et al. 2023, S. 15)
Niveau 1:	*Schülervorstellungen:* Studierende können einzelne Ursachen für die Entstehung von Schülervorstellungen nennen. (A4.1) *Experimente:* Studierende können einzelne Ziele des Experimentierens im Physikunterricht nennen. (A3)	*Schülervorstellungen:* Studierende unterscheiden in ihrer Charakterisierung wissenschaftliche Modelle von der gängigen Schülervorstellung, weil sie ein wissenschaftliches Modell nicht als richtig oder falsch, sondern als geeignet für die Erklärung eines Phänomens charakterisieren. (A32.1) *Instruktionsstrategien:* Studierende kennen typische Merkmale des entdeckenden Physikunterrichts. (A14.1) *Curriculum:* Studierende kennen Bedeutungsdimensionen der Wissenschaftsgeschichte für den Physikunterricht. (A19.2, A19.3) *Curriculum:* Studierende können zwischen zwei der drei Leistungsniveaus von Aufgaben unterscheiden. (A22.2)
Niveau 2:	*Schülervorstellungen:* Studierende können einzelne problematische Äußerungen, die durch Schülervorstellungen zum Thema Kraft und Reibung entstehen, erkennen. (A8.1) *Fachdidaktische Konzepte:* Studierende können einzelne Aspekte Didaktischer Rekonstruktion er- kennen und nennen. (A19.1, A23.1)	*Schülervorstellungen:* Studierende kennen typische und untypische Schülervorstellungen im Bereich des Elektromagnetismus. (A1.1) *Schülervorstellungen:* Studierende können einfache Experimente planen, um zu demonstrieren, dass die menschliche Haut keine Temperatur misst. (A5.1) *Schülervorstellungen:* Studierende können das Verständnis der Schüler:innen für wissenschaftliche Methoden durch Experimente fördern. (A7.1) *Assessment:* Studierende können zwischen allen drei Leistungsniveaus für Aufgaben unterscheiden. (A22.3)
Niveau 3:	*Experimente:* Studierende können erste Planungselemente in Bezug auf eine situationsspezifische Unterrichtssituation zum Thema gleichmäßig beschleunigte Bewegung entwickeln. (A10.1) Studierende können mehrere Ziele des Experimentierens im Physikunterricht nennen. (A3.1) *Schülervorstellungen:* Studierende können manche Schülervorstellungen aus Schüleräußerungen zum Thema Kraft und Reibung rekonstruieren. (A21b) *Instruktionsstrategien:* Studierende können die Missverständlichkeit eines Diagramms im Kontext der Kinematik evaluieren. (A13)	*Instruktionsstrategien:* Studierende kennen typische Merkmale verschiedener Unterrichtsmethoden. (A14.3) *Curriculum:* Studierende können Themen (z. B. zur Elektrizität) gemäß dem Spiralansatz anordnen. (A18.1) *Assessment:* Studierende können, Multiple-Choice-Aufgaben hinsichtlich des Stammes und der Distraktoren bewerten. (A24.1)
Niveau 3+:	*Experimente:* Studierende können vollständige Reaktionen in Bezug auf eine situationsspezifische Unterrichtssituation zum Thema gleichmäßig beschleunigte Bewegung entwickeln. (A10.2) *Schülervorstellungen:* Studierende können mehrere Schülervorstellungen aus einem Schülerdialog zum 3. Newtonsches Axiom rekonstruieren. (A1b.2) *Instruktionsstrategien:* Studierende können das Vorgehen einer Lehrkraft zum Erklären des 3. Newtonschen Axiom evaluieren. (A1a.)	*Schülervorstellungen:* Studierende können mögliche Quellen von Missverständnissen in wissenschaftlichen Darstellungen identifizieren. (A31.1) *Schülervorstellungen:* Studierende können die Vorstellungen der Schüler:innen zu wissenschaftlichen Experimenten (z. B. zum Verständnis der Natur der Wissenschaft) durch Experimente zu fördern. (A7.4) *Instruktionsstrategien:* Studierende können Anweisungen auf der Grundlage des Verständnisses der Schüler erstellen, die ihnen helfen, ihre wissenschaftlichen Konzepte zu ändern. (A33.1) *Curriculum:* Studierende können außerschulische Aktivitäten im Hinblick auf das Lernen der Schüler zu begründen. (A23a.1) *Assessment:* Studierende können Validität hinsichtlich eines Physiktests definieren. (A27b.1) *Assessment:* Studierende können Aspekte der Kompetenz der Schüler zu identifizieren, die durch Aufgaben bewertet werden können. (A28.1)

Tabelle 4.6 Gegenüberstellung der Scale-Anchoring Niveauformulierungen der Projekte. Die Operatoren der KiL/KeiLa - Ergebnisse wurden aus Schiering et al. (2023) übersetzt.

Niveau	ProfiLe-P+	KiL/KeiLa
1	nennen, erkennen	unterscheiden (×2), kennen (×2), charakterisieren
2	nennen, erkennen (x2)	unterscheiden, kennen, planen, fördern
3	nennen, entwickeln, rekonstruieren, evaluieren	kennen anordnen bewerten
3+	 entwickeln, rekonstruieren, evaluieren	definieren identifizieren (×2), erstellen, fördern, begründen

4.5.2 Passung eines Modells hierarchischer Komplexität des FDW zu den Testdaten

Zur Einschätzung der Passung des Modells hierarchischer Komplexität bzw. der Nutzbarkeit von Stufen hierarchischer Komplexität als schwierigkeitserzeugendes Merkmal des FDW wurden Regressionsanalysen für beide Testinstrumente bzw. beide Datensätze durchgeführt. Die Zuordnungen zu den Komplexitätsniveaus werden dabei als 3 Dummy-Variablen kodiert (Woitkowski & Riese, 2017). Die Ergebnisse der Regressionsanalysen sind in Tabelle 4.7 zusammengefasst und Abbildung 4.8 illustriert diese mithilfe von Violinplots.

Sowohl Abbildung 4.8 als auch die Varianzaufklärung von $R^2 = 0.39$ (multiples R^2) im Regressionsmodell ($F\ (2, 40) = 12.77$, $p < 0.001$) zeigen, dass das Komplexitätsmodell für den Datensatz aus ProfiLe-P+ einen substanziellen Anteil der Varianz der Aufgabenschwierigkeit aufklärt. Hier wäre es durchaus geeignet, als Niveaustufenmodell für das FDW herangezogen zu werden. Allerdings ist dies für den Datensatz aus KiL / KeiLa nicht in gleicher Form möglich. In Abbildung 4.8 zeigt sich nur ein leichter tendenzieller Anstieg der Aufgabenschwierigkeiten mit zunehmendem Komplexitätsniveau. Das Regressionsmodell selbst wird nicht signifikant ($F\ (2, 53) = 1.13, p = 0.33$) und klärt weniger als 5 % ($R^2 = 0.041$) der Varianz der Aufgabenschwierigkeit auf.

Die Komplexitätsstufen scheinen also nicht geeignet, um eine vom Testinstrument unabhängige Beschreibung von inhaltlichen Ausprägungen des FDW liefern zu können. Es wird daher hier darauf verzichtet, mögliche Wright-Maps mit Personenzuordnungen in die Niveaus abzubilden.

Tabelle 4.7 Ergebnisse der Regressionsanalysen zur Passung des Komplexitätsmodells an die Daten. Signifikanzniveaus p<0.05: *, p<0.001: ***. Das Regressionsmodell ist so konfiguriert, dass die Regressionskonstante den Mittelwert der Schwierigkeiten der Komplexitätsstufe I - Aufgaben beschreibt. Die Mittelwerte der anderen Stufen ergeben sich durch Addition ihrer jeweiligen Regressionsparameter zur Konstanten. Die Signifikanzniveaus geben an, ob die jeweiligen Schätzer signifikant von 0 verschieden sind. Auch wenn diese Frage hier zweitrangig ist, sind die Signifikanzniveaus der Vollständigkeit halber hier mit angegeben.

Komplexitätsstufe	**Regr. – Parameter b_i ProfiLe-P+**	**Regr. – Parameter b_i KiL/KeiLa**
Konstante (≃ I - Fakten)	−0.11 (n. s.)	−0.11 (n. s.)
II - Einstufige Kausalität	0.71*	0.18 (n. s.)
III - Mehrstufige Kausalität	2.15***	0.77 (n. s.)

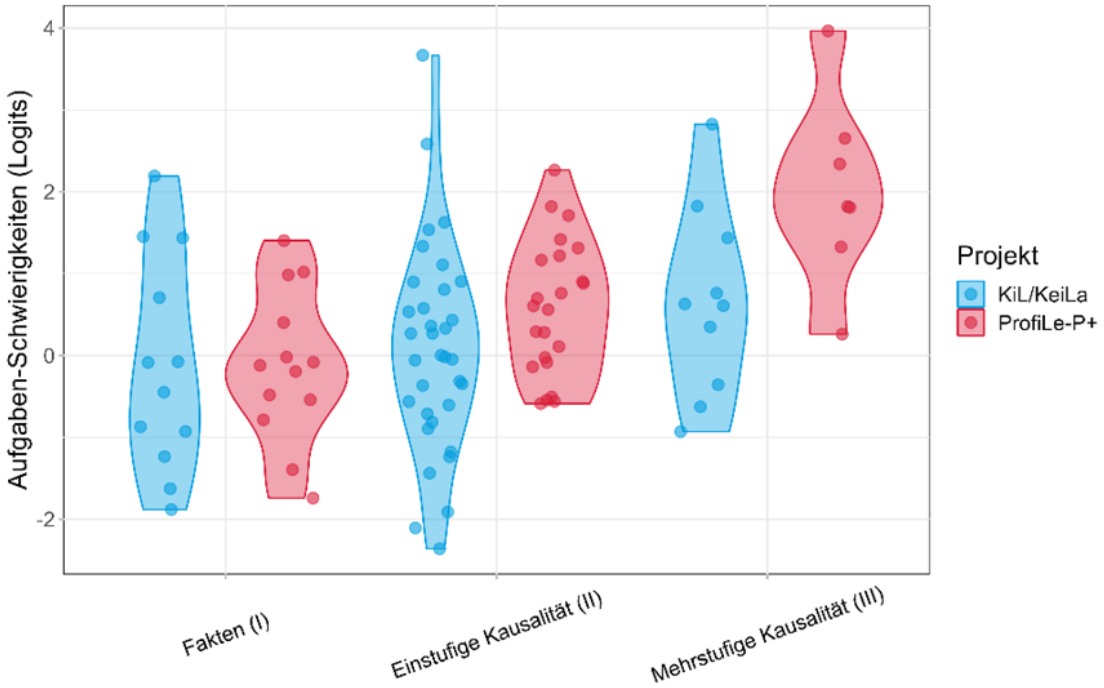


Abbildung 4.8 Violinplots der Item-Schwierigkeiten beider Projekte mit Einordnung in die Stufen hierarchischer Komplexität. Die Formen stellen die Wahrscheinlichkeitsverteilung der Datenpunkte dar; die Punkte sind die tatsächlichen Schwierigkeiten der Aufgaben.

4.6. Diskussion

Ziel dieses Beitrags war es, zu überprüfen, inwieweit sich projektübergreifend inhaltliche Ausprägungen des FDW mithilfe des Scale-Anchoring-Verfahrens sowie eines regressionsanalytischen Ansatzes zur Bildung von Niveaumodellen finden lassen. Solche inhaltlichen Beschreibungen von Ausprägungen stellen eine notwendige Voraussetzung für die gewinnbringende Übertragung der Forschungsergebnisse in die Lehrpraxis dar und sind darüber hinaus von übergeordnetem Interesse für das Forschungsfeld. Die projektübergreifende Analyse stellt zudem einen Forschungsansatz in Richtung einer

vereinheitlichten Beschreibung des FDW nicht nur auf theoretischer, sondern auch auf empirischer Ebene dar. Die verwendete Methode der Überführung quantitativer Ergebnisse in Niveaumodelle mithilfe von IRT-Analysen kann ggf. als Vorlage für andere verwandte Felder dienen.

Zunächst wurden die durch das Scale-Anchoring-Verfahren erhaltenen Niveaubeschreibungen der Projekte gegenübergestellt. Es zeigten sich dabei keine Ähnlichkeiten bzgl. fachlicher oder fachdidaktischer Inhalte, aber bzgl. des Auftretens von Handlungsoperatoren, die sich auf einer lernpsychologischen Ebene interpretieren lassen. Dabei fällt die Limitation der beschränkten Anzahl an Aufgaben für die Beschreibung des ersten Niveaus in ProfiLe-P(+) - Daten weniger ins Gewicht, da die beobachtete Systematik bzgl. des Auftretens der Operatoren hier für Niveau 1 *und* Niveau 2 gilt. Die so erhaltenen Abstufungen sind insgesamt konform mit Ergebnissen der Kognitionspsychologie zum Wissenserwerbsprozess (z. B. Gagné & White 1978) und lassen sich mit Standard-Taxonomien, wie beispielsweise der auf Lehr-Lernprozesse angepassten Bloom'schen Taxonomie nach Anderson und Krathwohl (2001; Erinnern, Verstehen, Anwenden, Analysieren, Bewerten, Kreieren) in Verbindung setzen. Insgesamt lässt sich somit auch die unsystematische Beobachtung zu Ähnlichkeiten in den Niveaumodellen der beiden Projekte (Abschnitt 4.3) im Sinne der FF1 bekräftigen:

> *FDW beschränkt sich unabhängig von der konkret zugrundeliegenden Operationalisierung in niedrigen Ausprägungen auf reproduktive Aspekte und erweitert sich in höheren Ausprägungen hin zu evaluierenden und kreierenden Elementen.*

Bemerkenswert ist hierbei, dass sich diese Parallele trotz einem deutlich größeren Anteil an Anfängerstudierenden im ProfiLe-P+ - Datensatz (vgl. Abschnitt 4.4.1 und Schiering et al., 2023, S. 8) zeigt.

Für den Transfer der Niveaumodelle in die Lehrpraxis zeigt sich, dass die durch das Scale-Anchoring-Verfahren erhaltenen Niveaus für die Einordnung von Lernenden in Niveaus und damit als Grundlage für das Erstellen entsprechenden Feedbacks geeignet sind. Die Niveaus und somit entsprechendes Feedback sind aber bzgl. des fachdidaktischen Inhalts abhängig vom jeweils verwendeten Testinstrument bzw. zugrundeliegender Modellierung. Das ist nicht direkt überraschend, da die beiden Testinstrumente nur in zwei von vier fachdidaktischen Facetten übereinstimmen und zudem im KiL / KeiLa - Testinstrument zusätzliche physikalisch-fachliche Inhalte thematisiert werden.

Es konnte gezeigt werden, dass die projektunabhängigen Systematiken entsprechender Niveaus primär eher allgemeine lernpsychologische Abstufungen darstellen, bzgl. derer dann auch projektunabhängige Aussagen unter Verwendung eines einzelnen Testinstruments getroffen werden können. Eine Einordnung von einzelnen Lernenden oder Lerngruppen in die Scale-Anchoring-Niveaus würde projektunabhängig bislang also beispielsweise eine Entscheidungshilfe für Lehrende bzgl. des Wechsels von eher theoretischen Lerninhalten (z. B. Vermittlung von Elementen entdeckenden Unterrichts) hin zu praxisorientierteren Elementen (z. B. Evaluation von Unterrichtsbeobachtungen) bieten. Auch bezüglich dieser lernpsychologischen Stufung kann eine Niveau-Einordnung allerdings noch keine differenziertere Empfehlung für eher kreative oder eher evaluierende Lerninhalte für Lernende

auf den höheren Niveaus unterstützen.

Aus theoriebildender Perspektive zeigen die Ergebnisse des Scale-Anchoring-Verfahrens, dass bei Austausch des fachlichen Inhalts sowie der fachdidaktischen Facetten bei ansonsten nahezu identischen theoretischen Annahmen in der Operationalisierung im Wesentlichen allgemeine kognitive Anforderungen als gemeinsame Systematiken einer hierarchischen Modellierung des FDW verbleiben. Es stellt sich also die Frage, ob aus Datenanalysen der Erhebungen mit entsprechenden Testinstrumenten abgeleitete Aussagen nicht grundsätzlich enger an die einbezogenen fachlichen (hier: physikalischen) Inhalte und fachdidaktischen Facetten gekoppelt sein müssten. Andererseits kann man die Ergebnisse des Scale-Anchoring-Verfahrens in folgendem Sinne auch als (Konstrukt-) Validitätsargument für die verwendeten Testinstrumente auffassen: In den beiden Testinstrumenten weichen die fokussierten Inhalte bzgl. der ersten zwei Dimensionen (1. fachphysikalische Inhalte und 2. fachdidaktische Inhalte / Facetten) der äußerst ähnlichen Itementwicklungsmodelle voneinander ab. Die sich zeigende übergeordnete Niveaustruktur lässt sich anschließend gerade durch die vergleichbare übrige Facette der „kognitiven Aktivierung“ (Gramzow, 2015) bzw. „Wissensarten“ (Kröger, 2019; Tepner et al., 2012) interpretieren. Dadurch werden die Annahmen der Operationalisierungen bzgl. einer entsprechenden Dimensionierbarkeit des FDW unterstützt.

Um die Vergleichbarkeit unterschiedlicher Operationalisierungen darüber hinaus weiter zu untersuchen, wären Studien wünschenswert, in welchen Proband:innen Testinstrumente aus unterschiedlichen Projekten bearbeiten. Korrelations- und Faktorenanalysen entsprechender Datensätze können ggf. weitere Aufschlüsse über Gemeinsamkeiten und Unterschiede der entsprechenden abgebildeten Konstrukte liefern. Für die Anwendung des Scale-Anchoring Verfahrens wären solche Datensätze auch interessant, da dann mehr Aufgaben in einem gemeinsamen Datensatz vorliegen würden, sodass die Niveaus detaillierter beschrieben werden und ggf. bisher unerkannte Systematiken zu Tage treten können.

Um die Ergebnisse der durch das Scale-Anchoring-Verfahren erhaltenen Stufen weiter auszuschärfen, wurde anschließend versucht, mithilfe der projektunabhängigen, lernpsychologisch begründeten Stufen hierarchischer Komplexität die Varianz der Aufgabenschwierigkeiten im FDW zu erklären. Während das entwickelte Modell hierarchischer Komplexität sich als sehr passend für die Daten aus ProfiLe-P+ erwiesen, zeigten sich trotz guter Übereinstimmung der Aufgabeneinordnung in das Komplexitätsmodell für beide Testinstrumente deutliche Limitationen in Bezug auf die Übertragbarkeit auf die Daten der KiL / KeiLa - Projekte. Da das Komplexitätsmodell aus dem ProfiLe-P+ - Team heraus vorgeschlagen wurde, ist nicht auszuschließen, dass es sich bei der mangelnden Übertragbarkeit auf KiL / KeiLa - Daten um ein Artefakt der Modellentwicklung handelt. Eine Konfundierung des Komplexitätsmodells durch bestimmte Überzeugungen und Blickwinkel auf das Konstrukt des FDW oder durch die Art der verwendeten Aufgabentypen des ProfiLe-P+ - Testinstruments konnte hier eventuell nicht vollständig vermieden werden. Das FDW scheint als „amalgam“ (Shulman, 1987) im Vergleich zum FW eine weniger stark kumulative Struktur aufzuweisen, was die Konstruktion eines projektunabhängigen theoretischen Modells schwierigkeitserzeugender Merkmale erschwert. (Physikalisches) FW ist auch aufgrund der starken Mathematisierung und damit verbundenen sehr klaren Beschreibbarkeit von Begriffen

und Konzepten stark hierarchisch geprägt. Begriffe und Konzepte aus der Fachdidaktik sind oft schwieriger exakt zu beschreiben und werden erst durch die gegenseitigen Beziehungen greifbar (z. B. „Didaktische Rekonstruktion“, „Elementarisierung“ und „Schülervorstellungen“).

Das hier vorgeschlagene Modell hierarchischer Komplexität allein stellt somit kein geeignetes Modell zur projektübergreifenden Aufklärung der Aufgabenschwierigkeit dar. Weitere mögliche Einflussfaktoren im Sinne eines „amalgams“ sind z. B. der thematisierte Fachinhalt, der sich in den beiden Projekten unterschied, das auftretende Fachvokabular oder auch die theoretische Thematisierung unterschiedlicher didaktischer Inhalte zu unterschiedlichen Zeitpunkten im Studium, d. h. die vorhandene Studienstruktur (Schiering, 2021). Letzteres kann auch einen Ansatzpunkt bieten, um zu erklären, weshalb auch auf hohen Niveaustufen offenbar teilweise noch neue deklarative Aspekte hinzukommen (siehe Tabelle 4.5 & Tabelle 4.6). Die Interaktion der genannten und weiterer möglicher Einflussfaktoren, scheint die hierarchische Struktur des FDW deutlich komplexer werden zu lassen, als mit einem stark verdichteten Modell hierarchischer Komplexität fassbar ist. Für eine umfassendere regressionsanalytische Niveaubildung mit einer größeren Anzahl an möglichen schwierigkeitserzeugenden Merkmalen wären allerdings Testinstrumente mit einer deutlich größeren Anzahl an Testitems notwendig, damit entsprechenden multivariaten Regressionsmodellen eine ausreichende Datengrundlage geboten wird.

Insgesamt konnten in diesem Beitrag vor allem mithilfe des Scale-Anchoring-Verfahrens trotz Unterschieden in der Testinstrument-Konzeption besonders hinsichtlich fachlicher und fachdidaktischer Inhalte projektübergreifende kriterienorientierte Systematiken von Ausprägungen des FDW ermittelt werden. Limitiert werden diese Beschreibungen vor allem durch die aus Gründen der Testökonomie und Zumutbarkeit vergleichsweise kleinen Aufgabenanzahl der FDW-Testinstrumente. So kann etwa in den höheren Niveaustufen keine Hierarchie zwischen kreierenden und evaluieren Elementen festgestellt werden. Es ist also noch weitere Forschung zu Vergleichen und zur Vereinheitlichung der empirischen Ergebnisse notwendig.

Da für die oben vorgeschlagene Erhebung neuer Datensätze mit Proband:innen, die mehrere Testinstrumente bearbeiten, große organisatorische Hürden überwunden werden müssten, wäre es dafür auch denkbar, ein gemeinsames IRT-Modell durch eine Normierung über die mittlere Personenfähigkeit einer hinsichtlich relevanter demographischer Merkmale ununterscheidbaren jeweiligen Unterstichprobe und anschließender konditionierter Schätzung der Item-Schwierigkeiten aufzustellen. In einer neuerlichen Anwendung des Scale-Anchoring Verfahrens könnten dann die Aufgabenschwierigkeiten auf Basis der fixen gemeinsam normierten Personenparameter geschätzt werden und es stände unmittelbar ein deutlich vergrößerter Aufgabenpool für die Charakterisierung der Niveaustufen zur Verfügung. Dafür müssten sowohl die Stichproben noch einmal im Detail auf eine Vergleichbarkeit geprüft werden als auch eine andere Software genutzt oder selbst entwickelt werden, da das hier genutzte R-Paket TAM (Robitzsch et al., 2024) keine direkte Schätzung von Aufgabenschwierigkeiten unter fixierten Personenfähigkeiten ermöglicht.

Die Betrachtung der Systematiken bzgl. lernpsychologisch interpretierbarer Operatoren als

Teil der inhaltlich kriterienorientierten Niveaubeschreibungen weisen auf eine praktikable Anwendbarkeit von lernpsychologischen Taxonomien auf das FDW hin. Gleichzeitig scheinen hierarchische Modelle evaluierende und kreative Elemente, die ab einer mittleren FDW-Ausprägung auftreten, nicht trennen zu können. Eine Alternative zu hierarchischen Modellen bieten Clusteranalysen (z. B. Duda et al., 2001) oder auch eng verwandte Latente Profil- oder Klassenanalysen (z. B. Spurk et al., 2020), die im naturwissenschaftsdidaktischen Kontext bisher nur wenig eingesetzt wurden (Zhai et al., 2020a; Zhai et al., 2020b). Daher bestehen in diesem Kontext noch keine prototypischen Vorgehensweisen, die synchron auf Datensätze unterschiedlicher Projekte angewendet werden könnten; die Entwicklung entsprechender Vorgehensweisen ist hier also zunächst das Ziel weiterer Forschung. Für die Daten aus dem ProfiLe-P+ - Projekt werden in diesem Kontext aktuell Vorgehensweisen erprobt, welche Clusteranalysen der Scores (Zeller & Riese, 2023) mit Methoden zur Machine-Learning-basierten Sprachanalyse der Sprachproduktionen der Proband:innen verbinden. Im Gegensatz zu IRT-Modellen können solche Ansätze auch nicht-hierarchische Strukturen aufdecken und hier womöglich zur Unterscheidung der Einflüsse von kreativen und evaluierenden Aspekten dienen.

Danksagung

Der Erstautor bedankt sich bei seinen Koautoren, diese Veröffentlichung im Rahmen seiner kumulativen Promotion nutzen zu können und daher besonders für die Zusammenarbeit und das erhaltene Feedback.

Beiträge der Autoren

- *Konzeptualisierung:* Jannis Zeller, Dustin Schiering, Josef Riese
- *Datenerhebung:* Christoph Kulgemeyer, Knut Neumann, Stefan Sorge, Josef Riese (sowie andere an den Projekten beteiligte Personen, die hier nicht alle als Autor:innen fungieren)
- *Datenpflege:* Jannis Zeller (ProfiLe-P+ - Daten), Dustin Schiering & Stefan Sorge (KiL / KeiLa - Daten)
- *Methodik und Formale Analyse:* Jannis Zeller (Scale-Anchoring-Verfahren: ProfiLe-P+ - Daten, Regressionsanalytischer Ansatz: beide Datensätze), Dustin Schiering (Scale-Anchoring-Verfahren: KiL / KeiLa - Daten)
- *Ergebnisinterpretation:* Jannis Zeller, Dustin Schiering, Christoph Kulgemeyer, Knut Neumann, Stefan Sorge, Josef Riese
- *Ursprünglicher Entwurf:* Jannis Zeller, Josef Riese
- *Review und Überarbeitung:* Jannis Zeller, Dustin Schiering, Christoph Kulgemeyer, Knut Neumann, Josef Riese, Stefan Sorge
- *Fördermittelbeschaffung:* Jannis Zeller, Christoph Kulgemeyer, Knut Neumann, Josef Riese

Alle Autoren haben der veröffentlichten Version des Manuskripts zugestimmt.

Förderung

Das Projekt ”Messung professioneller Kompetenzen in mathematischen und naturwissenschaftlichen Lehramtsstudiengängen“ (Akronym *KiL*) wurde durch die Leibniz Gemeinschaft unter dem Kennzeichen SAW-2011-IPN-2 gefördert. Das Projekt „Kompetenzentwicklung in mathematischen und naturwissenschaftlichen Lehramtsstudien-gängen“ (Akronym *KeiLa*) wurde durch die Leibniz Gemeinschaft unter dem Kennzeichen SAW-2014-IPN-1 gefördert. Das Projekt ”Professionskompetenz im Lehramtsstudium Physik“ (Akronym *ProfiLe-P+*) wurde vom Bundesministerium für Bildung und Forschung im Rahmen des BMBF-Rahmenprogramms ”Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor - Validierungen und methodische Innovationen“ (Akronym *KoKoHs*) unter dem Kennzeichen 01PK15005A-D gefördert. Die hier verwendeten Daten stammen aus den o. g. Projekten. Das Manuskript ist im Rahmen einer kumulativen Promotion entstanden, die mit einem Promotionsstipendium der Studienstiftung des deutschen Volkes gefördert wurde.

4.7. Kommentare und Ergänzungen

Das zentrale Ergebnis, d. h. die beobachteten Parallelen bezüglich der lernpsychologischen Operatoren in den Scale-Anchoring-Niveaubeschreibungen (Tabelle 4.5 & Tabelle 4.6) hat für das zweite Zielpaket der Arbeit den Fokus insbesondere auf die FDW-Dimension der kognitiven Anforderungen (z. B. Abbildung 4.1) gelenkt. Auch, wenn in anderen FDW-Modellen stattdessen teilweise die Dimension „Wissensarten" genutzt wird (Kröger, 2019; Tepner et al., 2012), so deutet das Ergebnis des ersten Artikels darauf hin, dass sich die Testaufgaben auch anderer Projekte im Rahmen einer gemeinsamen zugrundeliegenden Struktur bezüglich kognitiver Prozesse interpretieren lassen. Für das zweite Zielpaket lag daher der Fokus auf diese Dimension und die Nutzung einer entsprechenden lernpsychologischen Taxonomie (z. B. Anderson & Krathwohl, 2001) nahe.

Auch wenn die regressionsanalytische Niveaubildung hier projektübergreifend nicht genutzt werden konnte, ist aus Gründen der Transparenz und Dokumentation die Handreichung, die als „Manual" zur Zuordnung von Testaufgaben zu den Niveaus hierarchischer Komplexität des FDW erstellt wurde, in Anhang B dieser Arbeit enthalten. Sie stellt das Ergebnis eines iterativen Prozesses mit dem Ziel der Erreichung hoher Interrater-Übereinstimmung bei großer Expressivität dar. Diese Handreichung beschreibt implizit auch das Verständnis der einzelnen Stufen noch einmal deutlich.

Die Ergebnisse zum ersten Zielpaket haben den großen Vorteil, dass sie projektübergreifend sind und somit einem besonders hohen Anspruch an Generalisierbarkeit und projektübergreifende Bedeutsamkeit genügen. Dieser Ansatz wurde auch dadurch ermöglicht, dass von Seiten des KiL / KeiLa Projekts bereits ein Ergebnis sowie ein etablierter Workflow für Teile der Analyse vorlag.

5. Competency Profiles of PCK Using Unsupervised Learning (*Artikel 2*)

Einordnung in das Gesamtprojekt

In den Analysen zum ersten Zielpaket zeigten sich projektübergreifende hierarchische Strukturen, die inhaltlich im Kontext kognitiver Anforderungen interpretiert und beschrieben werden konnten. Diese erhaltenen Kompetenzniveaus sind allerdings (a) inhaltlich recht grob und (b) auf hierarchische Abstufungen beschränkt. Insbesondere kann mit ihrer Hilfe keine Unterscheidung zwischen Personengruppen mit Stärken oder Schwächen in den interessanten „oberen" kognitiven Anforderungen wie dem Evaluieren oder Kreieren getätigt werden. Dementsprechend wurde anschließend eine nicht-hierarchische Analyse mithilfe von Unsupervised-Learning-Methoden angestrebt, die primär auf den kognitiven Anforderungen basiert. Zu diesem Zweck wurde das Testinstrument von drei Expert:innen bezüglich der Taxonomie von Anderson und Krathwohl (2001) re-analysiert. Auch in diesem iterativen Prozess wurde eine Handreichung für diese Zuordnungen erstellt, die in Anhang B zu finden ist. Diese Handreichung beschreibt implizit auch das Verständnis der einzelnen kognitiven Anforderungskategorien noch einmal deutlich.

Die zur Vorbereitung der Cluster-Analyse somit angestrebte Zusammenfassung der Scores zu den kognitiven Anforderungen kann übergeordnet im Sinne der CGT als Pattern Refinement Schritt aufbauend auf den Ergebnissen der Pattern Detection in Zielpaket 1 aufgefasst werden. Die Cluster-Analyse selbst ist in dieser Betrachtung dann eine erneute Pattern Detection. Der Fokus auf die kognitiven Anforderungen (gegenüber beispielsweise den fachdidaktischen Facetten) erhöht vor dem Hintergrund der Ergebnisse der Niveauanalysen die Wahrscheinlichkeit einer projektübergreifenden Bedeutsamkeit und Anwendbarkeit der Ergebnisse der nicht-hierarchischen Analysen.

Ein direkt projektübergreifendes Vorgehen wie in Artikel 1 war für die Analysen zum zweiten (und dritten) Zielpaket im Rahmen dieses Projekts noch nicht möglich, da die nicht-hierarchischen Analysen hier in dieser Form erstmalig eingesetzt wurden und der Workflow dabei erst entstanden ist. Mithilfe der entwickelten Python- und R-Tools (siehe auch Kapitel 6 & 7 sowie Anhang G) ist die Übertragung des Analyseworkflows auf andere Projekte mit ähnlichen (nicht nur FDW-) Datensätzen aber ohne großen Aufwand möglich.

Bibliographische Angabe

Zeller, J. & Riese, J. (2025). Competency Profiles of PCK Using Unsupervised Learning: What Implications for the Structures of pPCK Emerge From Non-hierarchical Analyses? [Kompetenzprofile des FDW mithilfe von Unsupervised Learning: Welche Implikationen zu den Strukturen des FDW werden durch nicht-hierarchische Analysen sichtbar?]. *Journal of Research in Science Teaching.* https://doi.org/10.1002/tea.70001. *Vorab-Print (Stand 13.01.2025)*

Preprint-Statement

A newer version of this article has been published in the Journal of Research in Science Teaching. In this work, the article is presented as a pre-print as of January 13, 2025. The Version of Record is available online at: https://doi.org/10.1002/tea.70001.

Abstract

There have been several attempts to conceptualize and operationalize pedagogical content knowledge (PCK) in the context of teachers' professional competencies. A recent and popular model is the Refined Consensus Model (RCM), which proposes a framework of dispositional competencies (personal PCK - pPCK) that influence more action-related competencies (enacted PCK - ePCK) and vice versa. However, descriptions of the internal structure of pPCK and possible knowledge domains that might develop independently are still limited, being either primarily theoretically motivated or strictly hierarchical and therefore of limited use, e.g., for formative feedback and further development of the RCM. Meanwhile, a non-hierarchical differentiation for the ePCK regarding the plan-teach-reflect cycle has emerged. In this study, we present an exploratory computational approach to investigate pre-service teachers' pPCK for a similar non-hierarchical structure using a large dataset of responses to a pPCK questionnaire (N=846). We drew on theoretical foundations and previous empirical findings to achieve interpretability by integrating this external knowledge into our analyses using the Computational Grounded Theory (CGT) framework. The results of a cluster analysis of the pPCK scores indicate the emergence of prototypical groups, which we refer to as competency profiles: (1) a group with low performance, (2) a group with relatively advanced competency in using pPCK to create instructional elements, (3) a group with relatively advanced competency in using pPCK to assess and analyze described instructional elements, and (4) a group with high performance. These groups show tendencies for certain language usage, which we analyze using a structural topic model in a CGT-inspired pattern refinement step. We verify these patterns by demonstrating the ability of a machine learning model to predict the competency profile assignments. Finally, we discuss some implications of the results for the further development of the RCM and their potential usability for an automated formative assessment.

Keywords: Pedagogical Content Knowledge · Machine Learning · Unsupervised Learning · Language Analysis · Computational Grounded Theory

5.1. Introduction

Since the early descriptions of teachers' professional knowledge (e.g., Shulman, 1986, 1987), extensive research has been conducted on its structure and development (e.g., Sorge et al., 2019). Furthermore, research has explored its indirect impact on action-related skills among teachers (e.g., Kulgemeyer et al., 2020) and its direct impact within classrooms (e.g., Ball et al., 2001; Blömeke et al., 2022; Keller et al., 2017; Kunter et al., 2013). Given that studies have repeatedly demonstrated the significant impact of teachers on student achievement (e.g., Hattie, 2003, 2012), high-quality teacher knowledge and training are essential.

The central component of teachers' professional knowledge is the pedagogical content knowledge (PCK, Shulman, 1986, 1987) and considerable research has been conducted regarding its conceptualization and operationalization (Berry et al., 2015; Gess-Newsome & Lederman, 1999; Hume et al., 2019; Park & Oliver, 2008). PCK can be summarized as the knowledge that is necessary to teach a specific subject matter (e.g., the concept of energy in physics or the redox reaction in chemistry) to specific students (Baumert & Kunter, 2006; Shulman, 1987). Despite its significance, it remains challenging to assess the PCK's inner structure and typical competency levels on an empirical basis. Some hierarchical level models have been developed using approaches based on item-response-modeling, which yielded promising results (Schiering et al., 2023; Zeller et al., 2024). Nevertheless, these models are methodically limited because they generate primarily hierarchical, relatively rough statements. Non-hierarchical descriptions of PCK are usually not as empirically grounded. Such approaches primarily aim at characterizing different content aspects from a theoretical normative perspective. On the other hand, empirical studies are carried out, assessing PCK-related performance in action, e.g., in the context of the "plan-teach-reflect cycle" (PTR cycle, Alonzo et al., 2019; Behling et al., 2022b).

Therefore, more nuanced, potentially non-hierarchical, and empirically grounded descriptions of the PCK's fine structure are still in demand. Such descriptions could further improve the current state of the internationally widely used Refined Consensus Model (RCM) of PCK (Carlson et al., 2019) and thereby opening new avenues for research (e.g., learning process studies). Furthermore, empirically grounded knowledge about the PCK's fine structure including typical levels and knowledge areas that can potentially be developed independently from each other as well as the ability to assess such knowledge would be useful for improving PCK learning opportunities, especially through formative assessment (e.g., Hattie & Timperley, 2007).

To meet this demand, the present study offers a comprehensive examination of a dataset ($N = 846$) that includes scores and textual responses to a well-established (e.g., Kulgemeyer & Riese, 2018; Vogelsang et al., 2022) PCK questionnaire (cf. Gramzow et al., 2013). The sample is composed of pre-service physics teachers from 12 German-speaking universities. Through categorization of the questionnaire's tasks into requirement categories and cluster analyses of the scores, non-hierarchical "competency profiles" are derived. These get further refined and supported by a computer-based probabilistic language analysis of the authentic open-ended student responses to the questionnaire tasks. The findings indicate the existence of distinct PCK competency profiles with tendencies for specific language use that can be

interpreted through the lens of the aforementioned PTR cycle. We finally discuss the implications of these findings from a theoretical perspective as well as the possibilities for their use in an automated end-to-end assessment tool that can be used to provide content-rich feedback to future pre-service teachers.

5.2. Theoretical Background

5.2.1 Conceptualization of Pedagogical Content Knowledge

Over the years, various conceptualizations and operationalizations of PCK have emerged. There were several attempts to establish an international consensus model for PCK (Berry et al., 2015; Gess-Newsome & Lederman, 1999) with the most recent model being the RCM of PCK (Carlson et al., 2019; also see Hume et al., 2019). Following the RCM, PCK consists of three main realms, the *collective PCK* (cPCK), the *personal PCK* (pPCK), and the *enacted PCK* (ePCK). The cPCK describes the explicable, declarative knowledge base of the didactical community ("bookish knowledge"). The pPCK describes the internalized yet still mainly explicable knowledge of an individual (pre-service) teacher. Lastly, ePCK comprises the situational knowledge that emerges in specific teaching situations. The latter is therefore highly contextual, closely linked to the actions displayed in the particular situation, and thus, not explicable anymore. The RCM posits that the three PCK-realms impact each other via filters and transformation mechanisms, such as prior knowledge or professional beliefs (Carlson et al., 2019). It has shown to be challenging to empirically assess such filters explicitly (e.g., Behling et al., 2022a).

For the ePCK, an additional differentiation in the form of the PTR cycle as a mechanism through which ePCK is developed has been proposed by Alonzo et al. (2019). This mechanism describes the development of ePCK by iterating through planning, teaching, and reflection phases, both on a macroscopic (≥ whole lessons) and microscopic (specific teaching situations) level. It is therefore assumed that specific ePCK components for each step of the PTR cycle exist, i.e., $ePCK_{plan}$, $ePCK_{teach}$, and $ePCK_{reflect}$.

Another prominent model of professional competence, and PCK in particular, is Blömeke et al.'s (2015) "Model of Competence" (MoC). This model postulates a continuum of competence ranging from dispositions to performance. In the MoC PCK as a cognitive resource is positioned closer to the dispositional side of the model (Kulgemeyer et al., 2020) while the situational knowledge that emerges in specific teaching situations is positioned closer to the performance side of the MoC. Although both models refer to similar cognitive resources, the MoC and the RCM differ in their assumption of the relationship between PCK and other main domains of professional competence, namely content knowledge (CK) and (general) pedagogical knowledge (PK). The MoC places PCK, CK, and PK side-by-side on the same level to elicit situation-specific performance. In contrast, the RCM views CK and PK as foundational to PCK and PCK itself comprises situation-specific performance in the form of ePCK. Amongst other reasons, these differences arise due to cultural differences in their respective regions of origin, with the MoC being closer to a Central European transformative model of PCK and the RCM being closer to an Anglo-American integrative model of PCK

(Gess-Newsome, 1999; Mientus et al., 2022; Vollmer & Klette, 2023). As a consequence, the "PCK" construct defined in the MoC corresponds primarily to the pPCK realm of the RCM and the MoC contains the RCM's ePCK within the situation-specific performance.

Roughly summarized, apart from the conceptual differences in the relationship between the domains of professional knowledge (PCK, CK, & PK), the RCM can be viewed as a discretization of the MoC's continuum concerning PCK (for a more detailed discussion, see, e.g., Kulgemeyer et al., 2020; Vollmer & Klette, 2023). Questionnaires that measure (pre-service) teachers' PCK can be interpreted as primarily assessing pPCK in the context of the RCM or as focusing more on the dispositional edge of the MoC (e.g., Kulgemeyer et al., 2020; Schiering et al., 2023). Figure 5.1 summarizes the described framework models of PCK and professional competence (RCM and MoC) and shows the differences between the two models. The positioning of the construct measured by our test instrument (see Methods section) in the frameworks is highlighted in green.

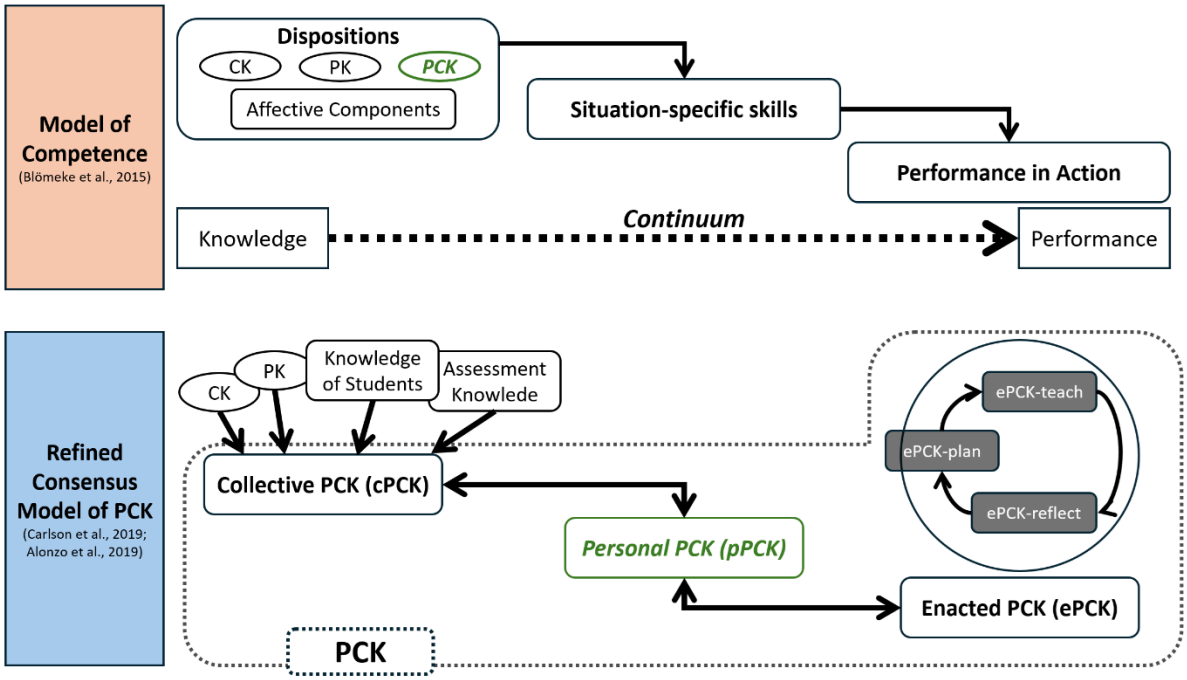


Figure 5.1 Framework models for PCK. The figure is inspired by Kulgemeyer et al. (2020) and comprises the basics of the Model of Competence (Blömeke et al., 2015) and the Refined Consensus Model of PCK (Carlson et al., 2019). We included the single ePCK components introduced in the context of the plan-teach-reflect cycle (Alonzo et al., 2019). The positioning of the construct measured by the test instrument used in this study is highlighted in green.

5.2.2 Structure and Development of Personal Pedagogical Content Knowledge (pPCK)

While the theoretical conceptualization of the ePCK is focused on actual actions occurring in the context of teaching and learning (e.g., planning, teaching, reflecting), theoretical conceptualizations of the internal structure of pPCK typically focus primarily on two dimensions: the associated CK and pPCK-subscales (Magnusson et al., 1999; Park & Oliver, 2008). The dependency on the associated CK stems from early descriptions by Shulman (1986,

1987) and has been a widely accepted assumption ever since (e.g., Hume et al., 2019). The pPCK-subscales describe different subsets of knowledge that are related to knowledge transfer. Most models include the two main subscales that were already described by Shulman (1986), namely *instructional strategies* and *student cognition*, but this collection is typically enriched by additional components relevant to the particular study context (e.g., Kulgemeyer et al., 2020; Magnusson et al., 1999; Park & Oliver, 2008; Schiering et al., 2023). The subscales are typically identified through argumentative means, expert interviews, and curricular evaluations of teacher education programs (ibid.). Some models that operationalize pPCK for an assessment include an additional dimension representing different levels of cognitive activity for the development of questionnaire tasks. This also holds for the development model of the test instrument that was used to generate the dataset analyzed in this study (Gramzow, 2015).

Although conceptualizations such as pPCK-subscales or the separation of cognitive activities provide an overview of the pPCK's presumed inner structure, it remains mostly unclear how empirically supported these distinctions are. Thus, it is uncertain whether these knowledge domains represent discrete components that can be developed independently. However, analyses of such potentially independent components would pose the potential for (a) further development of the conceptualization of pPCK as well as (b) formative assessment of PCK and construction of useful feedback for pre-service teachers.

To further investigate the internal structure of pPCK on an empirical basis, level analyses using item-response models have recently been conducted (Schiering et al., 2023, Zeller et al., 2024). Comparable analyses have also been carried out for CK (Woitkowski & Riese, 2017) and PK (König, 2009) in the German-speaking region and have yielded promising results. For pPCK, the results of Schiering et al. (2023) and Zeller et al. (2024), independently of the concrete context of the studies, found that the pPCK is limited to more reproductive, declarative knowledge at lower levels and extends to more analytical, creative, and evaluative aspects at higher levels.

Another line of research focuses on the relationship between pPCK and teaching practices (e.g., Großmann & Krüger, 2022; Kulgemeyer et al., 2020; for a comprehensive review, see Mientus et al., 2022). Behling et al. (2022b) were able to show that a learning opportunity focused on cultivating the ePCK components assumed in the PTR cycle (*plan*, *teach*, *reflect*) can also significantly increase the pPCK.

In summary, research on the empirical foundation of the internal structure of PCK is still ongoing. For pPCK, level models in particular have been discussed, while for ePCK, more action-related competencies have been investigated, e.g., based on the PTR cycle proposed by Alonzo et al. (2019). The level models for pPCK inductively showed promising results in describing different prototypes of pPCK in terms of operators that can be interpreted through the lens of cognitive psychology, e.g., analyzing, evaluating, and creating (Schiering et al., 2023; Zeller et al., 2024). However, these descriptions are rather general and methodically limited to hierarchical views. Specifically, while these analyses showed that cognitive activities can be used for project-independent descriptions of pPCK, the models by design are unable to distinguish non-hierarchical groups of students. With the term "non-hierarchical" we refer to groups of students that share the same overall pPCK level but differ in their competence w. r.

t. specific knowledge areas. Analyses of such potential non-hierarchical groups, which we call "pPCK *competency profiles*", would (a) be beneficial to further develop the description of pPCK beyond "lower" and "higher" levels and (b) enable the assessment and potential feedback that can focus on specific strengths and weaknesses of individual students. Given the ePCK's non-hierarchical distinction in the context of the PTR cycle, it seems promising to empirically study non-hierarchical structures of the pPCK as well. Imaginable is, e.g., a distinction between "$pPCK_{plan}$", describing a declarative knowledgebase necessary for lesson-planning, and "$pPCK_{reflect}$", describing a declarative knowledgebase necessary for effective reflection of teaching situations. Such a distinction would provide valuable insights into how pPCK is typically structured and could be efficiently fostered in teacher education programs. Therefore, our objective is to explore the emergence of such non-hierarchical structures for the pPCK. We conduct cluster analyses that are capable of detecting such patterns (e.g., Duda et al., 2001).

Cluster analyses, as a form of unsupervised Machine Learning, have been used only sparingly in science education research, in part due to challenges regarding the interpretability of the resulting structures (Zhai et al., 2020b). However, we argue that such approaches, when embedded in appropriate methodological frameworks, offer opportunities for the discovery of novel structures and information about non-hierarchical pPCK structures. Therefore, to improve the interpretability of our exploratory unsupervised analyses, we use ideas and concepts from the methodological framework developed by Nelson (2020), namely the Computational Grounded Theory (CGT). To enable the reader to follow the analysis and arguments, we explain some basic terminology and the CGT in more detail in the following section.

5.2.3 Unsupervised Learning in the framework of Computational Grounded Theory

The growing capabilities and increasing accessibility of Machine Learning (ML) methods have stimulated research on frameworks for categorizing and directing the use of these methods in science education research. Roughly speaking, ML can be described as the field of research that aims to automate human tasks using computer-based methods by "learning" from data (e.g., Géron, 2019). This learning process takes place through the application of various algorithms, such as (linear/logistic) regression models and clustering models.

Zhai et al. (2020b) showed that the majority of ML applications in science education research aim to automate assessment in supervised analysis settings. Supervised ML settings involve the prediction of (typically) manually generated labels, also referred to as "target", given some so-called "feature"-variables by an automated model (e.g., Géron, 2019). This can be the prediction of a specific class, e.g., the allocation to a certain group of people (target) from the responses to a questionnaire (features). Zhai et al. (2020a) proposed a classification framework for ML-based assessment in science education. They emphasized the potential of these methods to evaluate more complex constructs, potentially leading to a fundamental shift from simply replacing basic tasks to fundamentally redefining the assessment process and generating new opportunities.

Expanding ML applications beyond supervised settings is challenging. In unsupervised settings, there is no pre-existing target variable to predict, unlike in supervised settings (e.g., Géron, 2019). Unsupervised methods, such as cluster analyses, aim to uncover new patterns in data that can reveal previously unnoticed structures and generate fresh perspectives (Duda et al., 2001). They are typically employed when the amount of data exceeds a human-processable amount. To make new patterns and structures detected by unsupervised ML methods interpretable, they need to be linked to human expert knowledge (e.g., Nelson, 2020; Sherin, 2013).

Sherin (2013) suggested that using algorithmic ML methods "in tandem" with human expert knowledge and interpretive power can effectively leverage the potential of unsupervised analysis and increase confidence in the results generated at the same time (Sherin, 2013, p. 602; cf. Rosenberg & Krist, 2021). Nelson (2020) proposed the CGT framework to effectively guide such an in-tandem analysis. The CGT consists of three main steps:

(1) *Pattern detection*: Unsupervised techniques are used to identify new patterns and structures in the data. In the case of questionnaire data, this might be a cluster analysis of the available scores.

(2) *Pattern refinement*: The identified patterns are refined through in-depth analysis, i.e., human expert knowledge and interpretation power are introduced into the analysis. In the case of questionnaire data, this may be the aggregation of scores in the form of subscales or a language analysis of open-ended responses belonging to the found clusters.

(3) *Pattern confirmation*: To provide an argument for the stability and therefore validity[37] of the identified patterns and structures, the predictive power of algorithmic models for classifying the previously found categories is evaluated. In the case of questionnaire data, various models can be used to predict the previously identified clusters.

These steps from the original description of the CGT are strongly tailored to text analysis, where the first step is to find patterns in text data. However, it should be noted that these steps may/must be adjusted for specific projects depending on the data sources, research questions, and applicable methods at hand (Nelson, 2020, p. 10).

In the case of the present study, the data sources are the manually assigned scores for the tasks of a pPCK questionnaire and the digitalized text responses of the participants. By including both components of this rich data set, we aim to fully exploit its potential. To achieve this, additional theoretically and methodologically motivated preparation steps were introduced between the CGT's steps. The full workflow is discussed in detail in the Methods section.
The CGT has proven effective in science education settings, e.g., for elaborating students' ideas about the generality of their model-based explanations (Rosenberg & Krist, 2021) and for discovering argumentation patterns in students' problem-solving processes (Tschisgale et al.,

[37] The validity is assessed in the following sense: If ML-models are able to classify instances into the categories found during the pattern detection (and refinement), this is evidence for the existence of latent structures in the data, which correspond to the respective constructs (Nelson, 2020). The use of (potentially elaborated) models in this step includes non-linear structures which would often be overlooked when sticking to "classical" models like factor analyses.

2023). In particular the latter application by Tschisgale et al. (2023) is specifically aimed at presenting a prototypical CGT-oriented analysis and therefore also serves as a methodological guideline for structuring the results of the present study. Recently, Kubsch et al. (2022) proposed the *Distributing Epistemic Functions and Tasks* (DEFT) framework which can be seen as highlighting the untapped potential of unsupervised analyses in science education research. They explicitly name the CGT as a promising approach for unsupervised analyses of complex constructs. We therefore use extensive guidance from these theoretical foundations and previous empirical results to thoroughly interpret the patterns found.

The existence of non-hierarchical structures of pPCK is suggested by theoretical considerations in the context of the RCM. The amount of available data makes an exploratory analysis by human effort (e.g., qualitative manual analysis) infeasible. Using qualitative (manual) methods, it would be unlikely to capture or even consider all potential structural components of the data and it would also be challenging to link the different data sources available (scores and response texts). A computational, non-hierarchical analysis using unsupervised ML techniques is not only more efficient and perhaps more objective (in terms of reproducibility) but also facilitates the linkage between the different data sources. However, ML-based methods are limited by their inability to account for nuances and finer details, for example in the analysis of textual responses.

In summary, we conducted an exploratory analysis with a cluster model of the questionnaire scores at its core to uncover potential non-hierarchical structures of the pPCK. Knowledge of such structures would provide potential for the further development of the RCM. Furthermore, a potential assessment based on such results would guide the selection and evaluation of learning opportunities offered during teacher education programs. To structure and guide the incorporation of human expert knowledge and theoretical foundations, we use ideas and concepts from the CGT and DEFT frameworks. In the following sections, we discuss our goals and applied methods in more detail. Some additional technical details of the methods and algorithms used are presented in the Methods section, along with their application in the analysis, instead of discussing them as part of the Theoretical Background. We have found that this option facilitates the understanding of the methods and enables a shorter description.

5.3. Goal and Research Questions

As presented, the analysis of the fine structure of PCK is essential for advancing and consolidating the RCM and improving learning opportunities in teacher education programs. Regarding pPCK, which comprises PCK components that are developed during more theoretically focused learning opportunities, (primarily) theoretical descriptions of content subscales and strictly hierarchical level models are primarily available. For ePCK, the focus has been on (non-hierarchical) empirical analyses, particularly in the context of the PTR cycle. The hierarchical item-response-theory-based models for pPCK showed the potential for applying psychological learning operators and taxonomies (e.g., Anderson & Krathwohl, 2001) to pPCK independent of the specific study context. Such operators can also be loosely mapped onto the ePCK's PTR cycle, e.g., with evaluative and analytical aspects potentially being more

closely related to the reflect component, and applicative and creative aspects being potentially more closely related to the plan component.

Therefore, this study aims to investigate which structures can be empirically detected in a relatively large pPCK dataset using non-hierarchical cluster analyses. Yet, we do not aim at simply replicating the PTR cycle's structure for pPCK via some kind of confirmatory analysis. Instead, we conduct an exploratory analysis to allow for the discovery of new, previously undetected structures. However, previous findings related to the PTR cycle of ePCK and pPCK levels suggest that there might be a relationship between such structures of the pPCK and the PTR cycle, which will be part of the discussion. We call these (for now hypothetical) non-hierarchical pPCK structures "competency profiles", which should consist of content-oriented descriptions of strengths and weaknesses of prototypical physics (pre-service) teachers w. r. t. inductively analyzed criteria. The term "competency" emphasizes our focus on pPCK in the context of the RCM, as opposed to "performance" in action included in the ePCK, or at the dispositional edge of competency in terms of the MoC. The term "profiles" emphasizes our focus on non-hierarchical structures, as opposed to "levels", that have been analyzed using item-response models. The primary difficulty lies in empirically deriving such content- and criterion-oriented descriptions from the two data sources at hand, namely the pPCK scores and the authentic open-ended responses to the pPCK questionnaire tasks. To link these data sources, we assume that membership in a particular competency profile should reflect prototypical response behavior to the questionnaire tasks and vice versa. Therefore, we aim to carry out non-hierarchical cluster analyses using a distinctive blend of quantitative data (the manually generated scores) and qualitative data (the genuine open-ended responses of the participants).

To address the challenges discussed regarding the interpretability of the results of such exploratory cluster analyses, we extensively refer to the current state of research on pPCK-level models and ePCK conceptualizations within the PTR cycle. To structure and guide this combination of theoretical descriptions and exploratory analyses, we draw on ideas and concepts from the CGT. We therefore formulate the following three research questions, with each of them specifically focusing on one of the three steps of the CGT. The first research question describes our exploratory efforts for the analysis of non-hierarchical pPCK structures:

> ***RQ1*** *(~ pattern detection)*: Which competency profiles of pPCK emerge from the score dataset of a pPCK questionnaire using cluster analyses?

We do not yet narrow this research question down to the analysis of connectionist relations to the PTR cycle, to also allow for the discovery of previously unnoticed structures. To augment and elaborate the (for now hypothetical) non-hierarchical structures found in the scores, we carry out a language analysis of the test persons' authentic responses to the questionnaire tasks. This should provide valuable insights into the central thoughts and concepts on which each test person focuses:

> ***RQ2*** *(~ pattern refinement)*: Do test persons belonging to a specific competency profile show tendencies for specific language use in the open-ended responses to the pPCK questionnaire's tasks?

To consolidate the findings regarding RQ1 and RQ2, we additionally analyze the predictive

power of ML models to recover the found structures from the data in a CGT pattern confirmation step:

> ***RQ3** (~ pattern confirmation)*: How well can an (automated) ML model predict the competency profiles for unseen data?

This step is rather methodologically motivated than it is necessary from a theoretical perspective. However, given the difficulties in interpreting and replicating exploratory results, we argue in line with the CGT framework that RQ3 is still a valuable and necessary step in our study.

5.4. Methods

In the upcoming sections, we will initially provide details about the dataset we used and the corresponding studies. Subsequently, we will discuss our analyses in more detail.

5.4.1 Data Collection and Dataset

The data used in the present study was collected in the ProfiLe-P[38] project (Vogelsang et al., 2022) which took place from 2016 to 2019. This project aimed to assess the longitudinal development of teachers' professional competence, as well as relationships between professional knowledge and action-related skills. As part of this study, a (p)PCK questionnaire (Gramzow et al., 2013; Kulgemeyer & Riese, 2018), which included 20 open-ended tasks and 4 multiple-choice (MC) tasks[39], was a central part of the assessments. The study and the pPCK test instrument focused on precisely describing the relationships between the domains of professional knowledge for the specific CK-area of classical mechanics. During the piloting of the pPCK test instrument, multiple methods were used to assess and improve its validity and reliability. These included a validation against typical university curricula, a think-aloud analysis, and an evaluation of inter-rater reliability (Gramzow, 2015). The final version of the test instrument demonstrated satisfactory to excellent statistical properties, with an EAP reliability of $.84$ and a Cohen's of $\kappa = .87$ (cf. Kulgemeyer et al., 2020). This inter-rater Cohen's κ was estimated using a double coding of 267 full test edits. The test instrument covers the pPCK subscales *students' misconceptions and how to deal with them*, *instructional strategies*, *experiments and teaching of an adequate understanding of science*, as well as *PCK-related theoretical concepts* (Gramzow, 2015 translated by Kulgemeyer et al., 2020). Following the guidelines proposed by Klieme et al. (2003) a further dimension representing different levels of cognitive activities has been incorporated into the task development model, which comprises the cognitive activities *reproduce, apply,* and *analyze*. The complete model for task development is presented in Figure 5.2. An example of one of the questionnaire tasks

[38] German acronym "**Prof**essionskompetenz **i**m **Le**hramtsstudium **P**hysik" (professional competence in physics' teacher training). The project was funded by the German federal ministry of education and research.

[39] Depending on which elements are considered as the codable units, a total of 43 "items" (smaller codable units than "tasks") can be identified (e. g., Kulgemeyer et al., 2020). Therefore, we stick to the term "task" to denote the codable units we consider for this analysis.

belonging to the *student's misconceptions* subscale, including a response given by one of the tested prospective teachers, is presented in Figure 5.3.

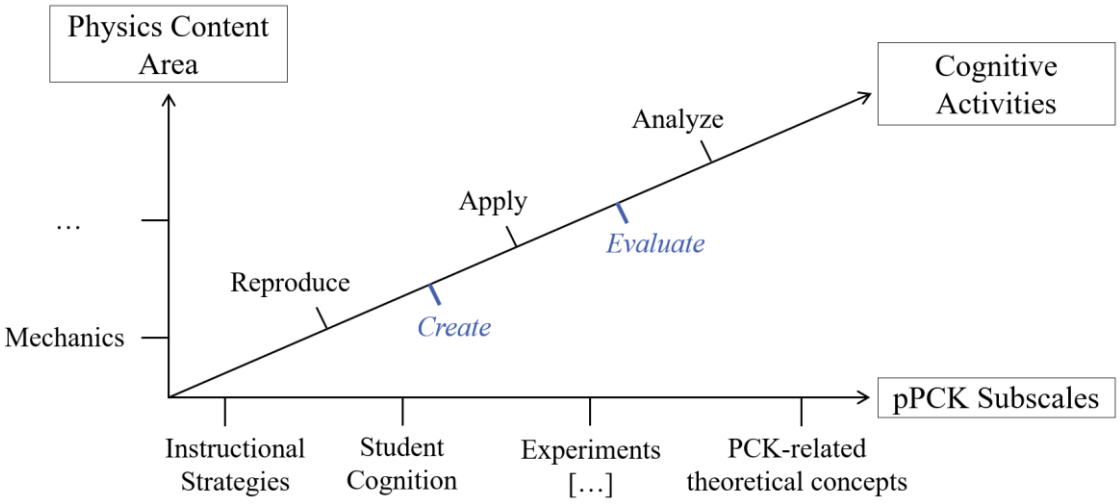


Figure 5.2 Model for task-development of the used test instrument. The original categories in the three considered dimensions used during task development are presented in black. The two cognitive activities added in blue are added for a more fine-grained differentiation of cognitive activities based on the findings of Schiering et al. (2023) and Zeller et al. (2024).

The final dataset used in the present study contains 846 edits of this questionnaire by pre-service physics teachers from 12 German-speaking universities. Teacher education in Germany takes the form of a bachelor's and master's degree program at the university level. The corresponding curricula offer distinct courses for the different domains of professional knowledge. For more details on the German teacher education system, we refer to van Dusen et al. (2021). Since the ProfiLe-P project had a longitudinal design, some participants took part in assessments up to three times. The individual edits are treated independently for this analysis according to the method of virtual subjects (Davier et al., 2008). Participants were on average in their second year of study ($M = 2.05$, $SD = 1.73$) and 34 % identified as female. The test instrument and all collected responses are written in German. The open-ended responses were coded by a trained German-native coder using detailed scoring rubrics (Gramzow, 2015; example in Table A1) and the MC tasks were scored using thresholds (cf. Krebs, 1997). In addition, the open-ended responses were digitized to allow for computational language analyses.

5.4.2 RQ1: Exploring Possible Competency Profiles with Score-Cluster Analyses

In the pattern detection step (RQ1), we aim to apply non-hierarchical cluster analysis methods to the pPCK score dataset. Due to its high dimensionality (> 20 tasks), the direct application of clustering algorithms to the "raw" score data does not yield sufficiently interpretable aggregations. Therefore, we first performed a theoretically guided step. Based on the findings

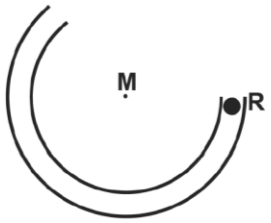
in the context of hierarchical item-response models (Schiering et al., 2023; Zeller et al., 2024) we suspected that relevant structures might emerge when focusing on common learning psychological operations. We therefore analyzed the questionnaire tasks using Anderson and Krathwohl's (2001) taxonomy and categorized them accordingly. A similar distinction was already included as the dimension of cognitive activities during the task development phase of the questionnaire (Figure 5.2), i.e., this preparation step is primarily an augmentation of this dimension with a re-evaluation of the tasks in these terms. An alternative approach would have been to utilize the pPCK subscales as categories for our analysis. However, this would have resulted in a lack of generalizability of the results, as different studies often target different selections from a wide variety of possible subscales (e.g., Hume et al., 2019; Park & Oliver, 2008). Conversely, the study by Zeller et al. (2024) demonstrated that cognitive requirement dimensions may be a more generalizable approach for the categorization of pPCK tasks and for assessing pPCK content-wise regardless of the concrete operationalization used and the physics content areas covered.

Although the taxonomy by Anderson and Krathwohl (2001) is intended to reflect a hierarchical ordering, the level analyses of the pPCK (Schiering et al., 2023; Zeller et al., 2024) showed that a hierarchical approach is not sufficient to distinguish between certain operations at the group level. Therefore, the application of non-hierarchical methods using these categorizations is a promising approach. In terms of the CGT, this categorization corresponds to the inclusion of human expert knowledge in the analysis (Nelson, 2020). We argue that this approach in combination with RQ2 retains enough openness to detect previously unnoticed novel structures compared to a direct categorization of tasks within the PTR cycle. Moreover, due to the more abstract nature (compared to the fundamental idea of ePCK) of the pPCK questionnaire tasks at hand using the PTR cycle directly would yield questionable categorizations anyway.

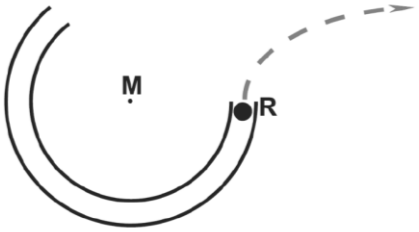
We focused on the operations *remember*, *understand*, *apply*, *analyze*, *evaluate*, and *create* (Anderson & Krathwohl, 2001). While Anderson and Krathwohl (2001) suggest that a learning objective should focus on a single operation in the taxonomy, we argue that it is valid and may even be necessary for more sophisticated and complex tasks to be able to focus on multiple operations. Therefore, we have allowed a single task to be categorized into more than one of the six operations. In an iterative process, a guideline for this task classification was established and refined several times to account for comprehension difficulties. A clear distinction between *remember* and *understand* still was difficult to make. We decided to collapse the two dimensions into a combined category called *reproduce*, i.e., tasks that primarily require the reproduction of explicit facts. We argue that this is still a valid step in the taxonomy and that this operation can be assessed from an outside perspective in a much more reliable and valid way. Figure 5.2 also contains the integration of the resulting five cognitive requirement dimensions in the task development model yielding the re-evaluation of the tasks accordingly.

Exercise 15

Students should consider the following situation: A ball rolls in the channel shown (top view) and leaves it at point *R*.



Student *A* draws the following path that the ball should follow after leaving the channel:



Solution of student A

Assuming that the student correctly understands the drawing as a top view: What wrong conception of student *A* causes the drawn trajectory?

Test person's answer:

"The student thinks that the centripetal force acts outwards and not towards the center of the circle."

Figure 5.3 Example task of the questionnaire used for generating the dataset analyzed in this study. The task belongs to the student cognition subscale. The task and response were translated by the authors. The scoring rubric as well as additional responses to this task from the dataset are appended as supplementary material (Table A1 & Table A2).

The final categorization was done three times by experts and resulted in the categorization agreement shown in Table 5.1. As shown there, an additional dimension "*teaching situation"* was added to describe whether a reference to a teaching situation, e.g., in the form of a vignette is part of the task. This dimension was found through an inductive categorization along with two other dimensions that were later found to be irrelevant. While it conceptually differs from the five cognitive activities, it is still usable as an argument for consistency because tasks belonging to the *create* and *analyze* dimension often refer to teaching situations. The five levels of the taxonomy and the additional *teaching situation* dimension will be interpreted together and referred to as "requirement dimensions" in the following. For the subsequent analysis, a consensus categorization was agreed upon by the three experts. In this categorization, task 15, which is displayed in Figure 5.3, was categorized into the *analyze* and *teaching situation* requirement dimensions. Table 5.2 shows the number of tasks and the maximum score that can be achieved in each of the requirement dimensions. It can also be interpreted as a measure of the granularity of each dimension.

Table 5.1 Cohens' κ values of the task-categorizations to the requirement dimensions. Based on these categorizations, a consensus-categorization was set up.

Raters	Reproduce	Apply	Analyze	Evaluate	Create	Teaching-Situation
κ_{12}	.84	.62	.76	.62	.71	.77
κ_{13}	.83	.55	.52	.71	1	.84
κ_{23}	.83	.59	.62	.41	.71	.62

Table 5.2 Maximum score for the dimensions based on the consensus-categorization of tasks. The test instrument is more focused on reproductive and analytical requirements. The implications and limitations of this for the analysis and the interpretation of the results are discussed in the Discussion section. If a category contains multiple choice tasks, this is denoted in parentheses, e.g. the Reproduce category contains 12 tasks in total of which three tasks are in multiple choice format. Note that a task can be allocated to multiple of the categories.

	Reproduce	Apply	Analyze	Evaluate	Create	Teaching Situation
Task Count	12 (*3 MC*)	5	10 (*2 MC*)	4	5	12 (*1 MC*)
Max. Score	23	8	13	5	9	16

After this theoretically motivated preparatory step, the actual cluster analysis was performed using the aggregated scores in the requirement dimensions as input data. To generate the clusters, we omitted cases in which less than 50 % of the tasks were completed or in which more than 25 % of consecutive tasks were not worked on at the end of the test instrument. We interpret such cases as instances where the test instrument was either not worked on seriously or the work was stopped early for some reason. For the cluster generation using this selection 779 instances remained. The aggregated score data allows for a proper interpretation of the clustering results using the averages of the resulting groups w. r. t. the requirement dimensions. The numerical properties of the dataset proved to be insufficient for the application of sophisticated clustering methods such as density-based algorithms (e.g., Campello et al., 2013; McInnes et al., 2017) or probabilistic Gaussian Mixture Models (cf. Spurk et al., 2020). Deviations from the normal distribution as well as discretization along the requirement dimensions (cf. Table 5.2) prevented the formation of meaningful clusters or even the convergence of the algorithms when using such methods. Therefore, we reverted to the simple but reliable K-Means algorithm (MacQueen, 1967) that is more agnostic to certain data requirements. The implications of this methodological choice are discussed in the Discussion section.

Additionally, we prepared the data by scaling the subscale scores to a range between 0 and 1, which facilitates K-Means convergence and cluster localization. The K-Means algorithm does not itself inductively estimate an appropriate number of clusters. There are some methods

to guide the selection of the number of clusters, such as the Silhouette-score method (Rousseeuw, 1987) or the elbow method (e.g., Géron, 2019). These methods aim to calculate metrics that represent the internal consistency of the clusters and the differentiation of the clusters from each other in a cluster model. In the case of the elbow plot, the sum of the distances of the data points from their respective cluster center is visualized. To achieve a balance between a low sum of distances and a moderately high number of clusters, an "elbow" is sought in a plot of the sum of distances against the number of clusters. This is analogous to the use of scree plots in exploratory factor analysis. The elbow plot for our data is presented in Figure 5.4. The bends in this plot at cluster numbers two and four are relatively smooth and do not provide a strong argument for selecting a particular cluster number.

A silhouette score analysis is similarly uninformative for our dataset and is therefore not discussed or presented in more detail due to space limitations. However, both procedures generally favor lower cluster numbers ($\lesssim 7$). From a theoretical perspective, a cluster number that is large enough to enable the discovery of non-hierarchical structures would be desirable. A cluster number of just two would be insufficient for our theoretical goal of finding non-hierarchical structures as it allows only for a simple differentiation between low and high-performing test persons. From a methodological perspective, a lower number of clusters would be preferable for future work on automating the allocation of test edits to such clusters. Given the slight bend at a cluster number of four in the elbow plot and the need to balance the theoretical and methodological requirements, we decided to use a cluster number of four for the subsequent analysis. It should be noted that higher cluster numbers (five to seven) also yield distinct and interpretable clusters that resemble finer-grained differentiations of the sample.

The remaining parts of the score cluster analysis lead directly to the results of RQ1 and are therefore described in the results section to enhance the comprehensibility of this article.

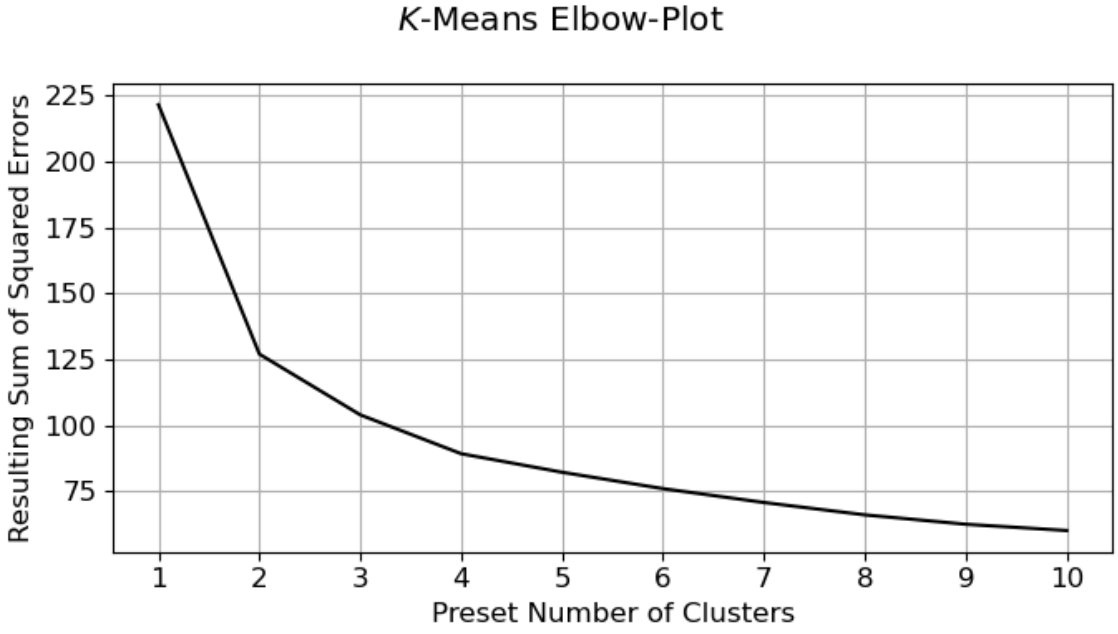


Figure 5.4 Elbow-Plot to guide the decision for a fixed cluster number for the score-cluster analysis.

5.4.3 RQ2: Refining the Score-Clusters to Competency Profiles via Topic Analysis

For the second research question, which is aimed at the pattern refinement step of the CGT, we provide insights into possible tendencies for specific language usage of the participants belonging to specific score clusters from the first analysis step. We want to gain insights into the focused concepts and ideas of these groups to (1) refine our knowledge about them and (2) provide an argument for the concurrent validity (Miller & Lovler, 2018) of the interpretation of the groups by assessing the consistency between their strengths and weaknesses in terms of the requirement dimensions and their language usage.

Although standard works on PCK in the context of the RCM do not (yet) explicitly investigate connections between PCK and specific language use (e.g., Hume et al., 2019), it can be assumed that language can be viewed as a central medium through which cPCK and especially pPCK as a cognitive construct are expressed and shared. This is also reflected in the coding rubrics of the test instrument used in this study that specifically pay attention to terminology in the context of teaching and learning physics (Table A1) Therefore, investigating the language use of test persons belonging to the clusters discovered in the RQ1-analysis should yield information on central constructs and ideas that the test persons consider relevant when tackling the questionnaire tasks. This information can be used to gain insight into the test persons' personal understanding of PCK, i.e., their pPCK and thereby extend the description and interpretation of the clusters beyond the score aggregations.

To assess potential prototypical language use of groups or to identify groups of prototypical language use in a dataset of texts (called a *corpus* of individual *documents*) typically so-called topic models are used (Chen & Liu, 2017). The term topic model describes methods that aim at characterizing topics in a corpus. A topic is characterized by a set of words. The co-occurrence of words in different documents determines the topics and the topic prevalence of the documents, i.e., how much a document addresses a particular topic. The original topic modeling algorithm, called Latent Dirichlet Allocation (Blei et al., 2003), was specifically tailored for topic modeling in the famous paper by Blei (2012). By using the basic or extended version of the topic modeling algorithm one can infer the topic-word and document-topic relations given only the words of the documents. A thorough description of the inner workings of the model requires considerable prior knowledge of probability theory and is beyond the scope of this article. Modern open-source software packages provide easy-to-use interfaces for applying topic models without having to delve deeply into the mathematical foundations (e.g., Roberts et al., 2019).

Without claiming completeness, two lines of research can be identified for extending the basic topic model. The most recent iterations use deep learning-based language models (e.g., BERT by Devlin et al., 2019) to transform the documents into numerical representations that are subsequently used in cluster analyses. The documents belonging to specific clusters are then used to characterize the topics by extracting characteristic words (Grootendorst, 2022). Such approaches are already being used in science education research and are yielding promising results and insights into short text elements (sentences) extracted from longer documents (Tschisgale et al., 2023; Wulff et al., 2022). However, applying a similar approach

to the dataset of the present study is not practical for our data structure, because we cannot use short elements like sentences in a meaningful way.

Therefore, we refer to the second line of research on the extension of the basic topic model mentioned above. Models emerging from this second line of research do not use deep learning methods but instead, directly extend the mathematical model of the basic topic model. In doing so they aim to directly incorporate additional covariates into the model and thus guide the formation of topics by these covariates (e.g., Blei & Lafferty, 2005). A relatively new model that has emerged from these approaches is the so-called structural topic model (STM, Roberts et al., 2016; Roberts et al., 2019). It allows the use of covariates that influence the topical prevalence of particular documents as well as the use of covariates that influence the content of particular topics, i.e., the topic-word-relation. The probabilistic model and the inference algorithm become even more sophisticated (Roberts et al., 2016). In our case, this model is particularly interesting because we can focus on the most relevant words by applying certain preprocessing steps to our documents, and we can guide the topic prevalence by our results from the first analysis steps, i.e., the score clusters to which the documents are assigned.

We applied the following preprocessing steps guided by the R-software used (Roberts et al., 2019; Roberts et al., 2023):

- *Punctuation removal*: Removing punctuation, i.e., omitting characters such as periods, hyphens, etc., is a common preprocessing step when applying models that are agnostic of the order of the occurring words.
- *Lower casing*: All words are converted to lowercase to remove unnecessary variance in the corpus.
- *Removing stopwords*: "Stopwords" are words that occur so frequently that they do not provide interesting insights into the documents like "and", "I", "the", etc. Removing such stopwords can be interpreted as removing uninteresting variance from the corpus and reducing the document lengths for more efficient computation.
- *Removing words based on frequency*: Similar to stopword removal it is common to remove words that are either too frequent or too rare. Removing too frequent words serves the same purpose as stopword removal. Removing too rare words can be seen as removing variance from the corpus that cannot be explained or interpreted anyway because there is not enough data available. We decided to categorize words appearing in more than 60 % of the documents as too frequent and words appearing in two or less documents as too rare to be further used[40].
- *Stemming*: Stemming refers to the reduction of words to their core component or stem, e.g., reducing "programming", "programmer", and "programs" all to "program". Again this step is aimed at reducing the variance of the corpus and is a common technique when

[40] Using such additional thresholds is suggested by the software used Roberts et al. (2023). We found that when these thresholds were used, the resulting topics were much less dominated by the same very frequent words and therefore much more expressive. Moreover, the metrics used to determine the most characteristic words were less prone to "collapse" into words used only once.

using word-order-agnostic models. We argue that stemming specifically does not hinder us to understand and interpret the meaning and content of a topic.

For the actual topic modeling, we introduced the participants' score cluster assignments[41] from the results of the first research questions as covariates for the topic prevalence of the corresponding documents. Most topic models including the STM require the number of topics to be preset by the analyst to a fixed value. There are some ways to estimate the appropriate number of topics for a dataset in a data-driven way, but the corresponding metrics often yield inconsistent results (see e.g. Figure 3 of Gan & Qi, 2021). The choice of the topic number is generally not considered an exact science (Roberts et al., 2023). For our dataset, the available metrics of the software used roughly favored low (< 10) topic numbers. Therefore, we gradually estimated models with an increasing, but still comparatively low, preset number of topics. We reached saturation within a six-topic model, i.e., we found additional topics to be primarily unspecific or repetitive in content and therefore kept a topic model with six topics for further analysis.

In the subsequent analysis, we first interpreted the topics using the most characteristic words based on the metrics provided by the software (cf. Roberts et al., 2019; Roberts et al., 2023 for more details). In addition, we estimated the effect of belonging to a particular score cluster from RQ1 on the proportion of a document focused on a particular topic. The score clusters and their refinement through the topic analysis together form the groups that we refer to as "competency profiles" in the following. They are the lens through which we aim to describe non-hierarchical structures of the pPCK. Analogous to the previous section, the remaining parts of the topic model analysis lead directly to the results of RQ2 and are therefore described in the Results section.

5.4.4 RQ3: Confirming Competency Profiles by Automatized Prediction

The pattern refinement step of the CGT aims to assess the stability and robustness of the explored patterns and structures from the unsupervised analyses (Nelson, 2020). In the case of the present study this is primarily reflected in the stability and robustness of the identified competency profiles, since the workflow for assigning a participant or a questionnaire edit to a competency profile depends only on the scores. The topic modeling step (RQ2) is primarily intended to provide additional insight into the competency profiles and arguments for their validity; it does not influence the assignment of a person to a competency profile. To confirm the explored patterns, Nelson (2020) suggests assessing the predictive power of ML models that assign some appropriate input data ("features") to the labels generated during the pattern detection and refinement steps. Taken together, this means that in our case the goal of the pattern confirmation is to automatically predict the score clusters from RQ1. The predictive power is evaluated by splitting the data into a training set and a test set (e.g., Géron, 2019). The ML model is then trained on the training data to predict the labels. The performance on the prediction task is estimated using the "unseen" test set. A high performance on the test set

[41] Note that we are still using an approach of virtual cases and therefore the same person can be assigned to different score-clusters at different times.

serves as an argument for the reliability and validity of the explored patterns. If only a small to medium sized dataset is available, the so-called "k-fold cross-validation" approach (e.g., Géron, 2019) can be used to enhance the results. This approach is based on dividing the full dataset into k segments. The model is then trained k times on the data, with each iteration involving the omission of one of the k splits for training purposes and its exclusive use for evaluation. Furthermore, the cross-validation procedure can additionally be repeated multiple times, with different so-called "random seeds", i.e., different splits. This allows for the estimation reliable performance estimates.

Now the question arises, which part of the data should serve as the features in the present analysis? There are two main options: First, the actual text responses could be used to predict the cluster assignments. Second, the score data could be used to predict the cluster assignments. Predicting the (score) clusters using the text responses is a much more complex inference task than using the scores directly because the scores have already been used as the basis for the cluster analysis in the pattern detection step. We followed the lead of Tschisgale et al. (2023) who reused the features used in the cluster analysis again in the pattern confirmation step. Therefore, we present a pattern confirmation analysis using the score data for prediction. Note that this decision significantly reduces the complexity of the prediction task compared to using the textual responses. Furthermore, such a model is primarily intended to be used for pattern confirmation to complete the CGT methodology and thereby provide an argument for the stability and validity of the identified competency profiles. However, it is of little practical relevance as the main work of assigning scores to the open-ended responses must still be done manually. Full automation, i.e., scoring the open-ended responses and assigning participants to competency profiles, requires much more sophisticated approaches that are beyond the scope of this article but will be part of our future work (also see "Perspectives and Outlook").

We evaluated a logistic regression classifier model using a 10-fold cross-validation (e.g., Géron, 2019) with 10 different random seeds, resulting in a total of 100 estimates for the performance of the model in predicting competency profiles from the scores. The dataset is imbalanced with an uneven distribution of the target variable, i.e., the cluster assignments (see Table 5.3). Therefore, the cross-validation splits were set up such that the distribution of the cluster assignments is almost equal in all splits. In addition to the predictive accuracy, we also report the weighted F_1-score as well as the (linearly weighted) Cohen's κ score for the test-set predictions. These scores account for imbalance in the data sets and Cohen's κ also account for random agreement.

The logistic regression classifier was configured and trained using the Scikit-Learn Python package[42] (Pedregosa et al., 2011). To facilitate the classifier's generalization from the training to the test data, we chose an L_2-regularization value of 1.0[43].

[42] https://scikit-learn.org; we used version 1.3.0.

[43] For additional information on regularization in general please refer to Géron (2019, pp. 28-33). For information on regularization in regression models specifically please refer to Géron (2019, pp. 135-141).

5.5. Results

5.5.1 RQ1: Score-Clusters in pPCK Data

The core of the competency profiles is formed by clusters in the scores as described in the Methods section. The resulting six-dimensional (5+1 requirement dimensions) clusters cannot be visualized directly, but dimensionality reduction techniques can be used to project the data down to two dimensions. Figure 5.5 shows such a visualization for the pPCK scores dataset used for the cluster analysis. The clusters are well distinguishable even in the dimensionality-reduced projection. Since the clusters have non-circular shapes and erratic densities, the deviations from a Gaussian distribution are visible. The visualization in Figure 5.5 shows the legitimacy of using the clusters to further describe competency profiles, but also some limitations related to the distributions and overlaps which will be discussed in the Discussion section.

However, the shapes and distribution of the clusters especially when projected to two dimensions do not directly indicate the potential strengths and weaknesses of the corresponding competency profiles. For this purpose, we take a closer look at the average scores of all instances within the clusters resulting from the K-Means algorithm. For the present analysis with only four clusters and six-dimensional data, these averages can be displayed with radar plots or line plots as shown in Figure 5.6. The radar plot (Figure 5.6, top) is scaled to the highest score achieved in each dimension. Consequently, a "*reproduce* score" of 0.6 means that $60\ \%$ of the scores of the best-performing individual have been achieved. This is due to the scaling of the data to the interval $[0, 1]$ in the cluster analysis' pre-processing. The line plot (Figure 5.6, bottom) is scaled to the overall best-achieving cluster to directly highlight the differences between the clusters. Additionally, the line plot shows the 95% confidence intervals for the means as shaded tubes. Clusters 1 and 4 are identified as clusters with generally low and high overall achievement. Clusters 2 and 3 differ significantly in their scores on the *analyze/evaluate* and *apply/create* requirement dimensions. While Cluster 2 achieves significantly higher scores in the dimensions *apply* and *create*, cluster 3 achieves significantly higher scores in the *analyze* and *evaluate* dimensions. Nevertheless, these clusters show little difference in their scores regarding the *reproduce* dimension.

Table 5.3 presents additional details on the clusters. The lower-performing cluster 1 contains by far the largest number of students. This is most likely due to the large number of first-year students in the sample. Clusters 1 and 2 as well as 3 and 4 respectively show significant differences in their average year of study and their total pPCK score. Clusters 2 and 3 do not show significant differences in their year of study ($T = 1.68,\ p = .1,\ \mathrm{df} = 320$), yet they show (barely) significant differences in total pPCK scores ($T = 2.01,\ p = .05,\ \mathrm{df} = 320$). The latter is strongly influenced by the relatively large size of the groups, which is also reflected in the small Cohen's d effect sizes of these differences of $d = 0.18$ and $d = 0.23$ respectively. The absolute differences are much smaller than the differences between clusters 1 and 2 as well as 3 and 4 respectively. Based on the average scores of the clusters in the requirement dimensions and their average total pPCK score (as presented in Figure 5.6) we have created labels for the clusters: the *Low Achievers* (cluster 1), the *Applying Creatives* (cluster 2), the

Analytic Evaluators (cluster 3), and the *High Achievers* (cluster 4). Because the score clusters form the core of the competency profiles, these labels also become the names of the competency profiles.

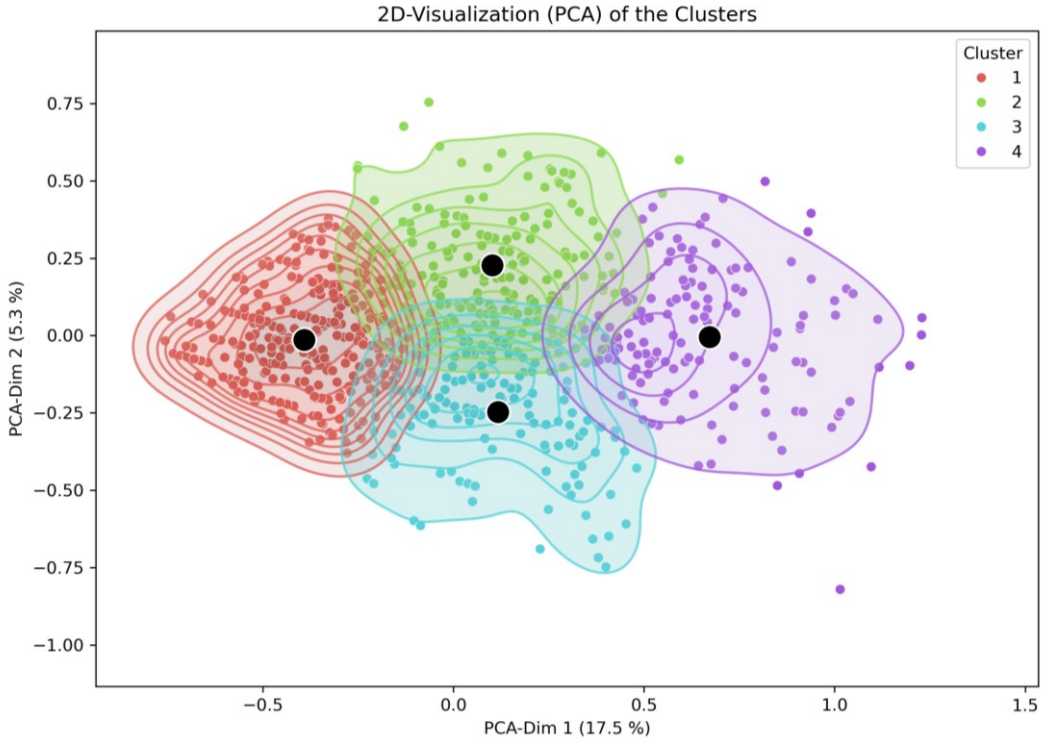


Figure 5.5 Two-dimensional visualization of the dataset and clusters. Each colored dot represents one test person. The bold black dots represent the centroids of the respective clusters. We used the Principal Component Analysis dimensionality reduction technique (for details refer to, e.g., Jolliffe, 2002) to project the six-dimensional data to two dimensions. The percentages in the axis-labels denote how much of the variance of the full six-dimensional dataset is retained by the corresponding reduced dimension. The shading and contour lines represent the density of the datapoints.

Table 5.3 Sizes (*N*), average year of study and average total pPCK scores of the clusters. The counts for the full datasets are generated by applying the fitted model to the cases previously omitted due to incompleteness. The means are calculated using the filtered dataset. The differences when the previously omitted instances are included in these aggregations are small.

Cluster	**Year of Study**		**Total pPCK Score**		***N***	***N* (full dataset)**
	M	***SD***	***M***	***SD***		
1	1.34	1.20	9.17	2.99	321	383
2	2.17	1.59	15.91	3.66	179	181
3	2.39	1.81	16.72	3.62	144	147
4	3.49	2.00	23.82	3.90	135	135

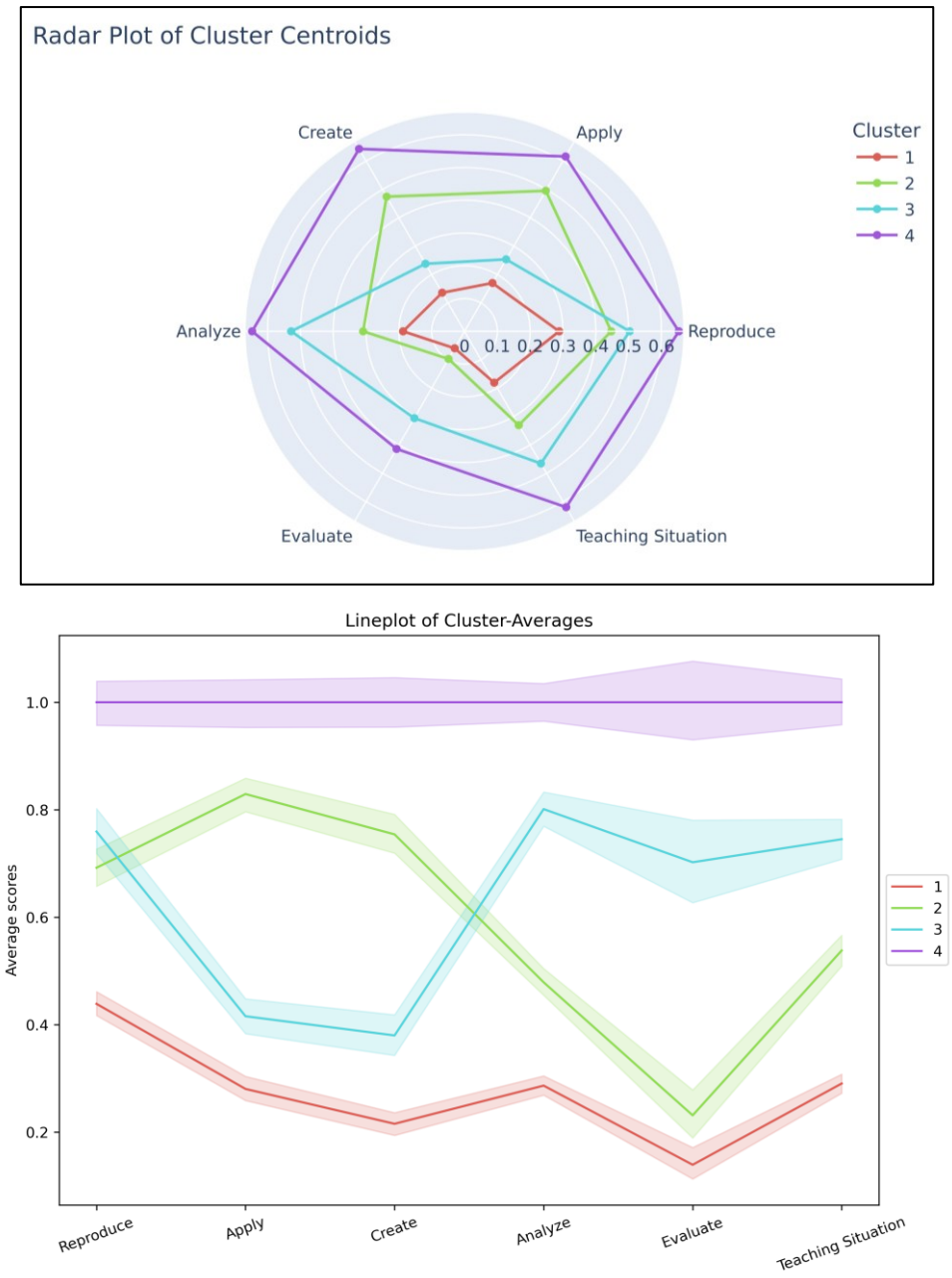


Figure 5.6 Visualizations of the cluster centroids. The cluster centroids are the means of the instances in each cluster and represent the typical scores of the competency profiles in each of the requirement dimensions. In the radar plot, the scores for each dimension are normed to the highest score a person achieved on that dimension. In the line plot, the scores for each dimension are normed to the mean scores of cluster 4 to emphasize differences between the clusters.

5.5.2 RQ2: Typical Language Use of Participants Belonging to the Score-Clusters

As described in the Methods section, we used a structural topic model to further refine the description of the competency profiles beyond their typical scores in the requirement dimensions. The score cluster assignment was used as a covariate for the topic prevalence of the documents. The documents are the person-wise concatenated responses to the open-ended tasks of the pPCK test instrument. The generation and preprocessing of the documents for this

analysis are also described in the Methods section. Note that we decided to use the full dataset including the edits that were omitted for the clustering step to make use of the full language data available. This is possible by performing the cluster assignments for the previously omitted edits using the clustering model that was only fitted using the complete edits.

The initial step involves using the characteristic words that emerge from the model to describe the content of the topic. Different metrics can be used together for this purpose (Roberts et al., 2019). Table 5.4 illustrates these word lists, excluding duplicate words belonging to the same topic and non-specific general words. Note that the structural topic model is a soft assignment model, i.e., it allows the same word to appear in different topics. The resulting wordlists were categorized using six deductive topics that were derived by interpreting the "human-interpretable" groups of words that appeared in these lists. Based on these deductive topics (columns in Table 5.4) the actual inductively found topics from the topic model (rows in Table 5.4) were characterized. We provide a brief interpretation or title for each inductive topic in Table. These inductive topics will be referred to simply as "topics" in the following. To refine the competency profiles, we assessed the relationship between the assignment of a document to a score cluster and the proportion of a document dedicated to a specific topic. Thus, we estimated the effect that the assignment of a document to a score cluster has on the topic proportion of the document by aggregating the topic proportions generated by STM and grouped by the cluster assignments[44]. A full numerical comparison of these effects (via ANOVA and post-hoc tests) is not presented due to space limitations. In summary, with the exception of *reasoning on examples*, all topics were significantly affected by the score-cluster assignment ($p < .001$ for the remaining five). To further compare these effects, we present them as a heatmap (Figure 5.7).

The effect differences between the competency profiles range from 0 to $.40$. For space reasons only important outstanding observations are reported and further refined. It is important to recognize that the proportions focused on specific topics in a document are always relative and sum to 1 when accumulated across all topics. Therefore, the proportions shown in Figure 5.7 sum to 1 row-wise. The heatmap in Figure 5.7 shows that engagement in the *student cognition* topic is increasing alongside the average total score of the competency profile; note that the competency profiles are arranged with increasing total scores along the vertical axis. Additionally, there is a simultaneous decrease in the emphasis on general concepts (topics *general concepts focusing subject* and *general concepts focusing knowledge*) and the extensive use of examples (topic *usage of examples*). A similar trend, but on a smaller (non-significant) scale, can be seen when comparing the Analytic Evaluators with the Applying Creatives, with the Analytic Evaluators more strongly prioritize the *student cognition* topic over the *usage of examples* and *reasoning on example* than the Applying Creatives. Lastly, the Low Achievers show a significantly ($p < .05$ for each post hoc comparison) reduced emphasis on the *symbolic descriptions* topic when compared to the other competency profiles.

[44] On closer inspection, we switch from a generative probabilistic modeling approach of the STM to a more frequentist approach by using the STM's initially predicted "mean" topic proportions. We decided to do this because this approach is much easier to follow and the actual numerical differences compared to using the generative utilities (Roberts et al., 2023) are negligibly small.

Table 5.4 Characteristic words and short interpretations / titles for the topics discovered in the structural topic modeling analysis. The words are translated from German to English. The words have been grouped according to some interpretative topics identified by the authors in the occurring words (columns).

Topic	Subject concepts & their transfer	(Scientific) Working methods	Student cognition	Symbolic / Math. language	Examples	Application & Legitimation of examples	*Short interpretation / Topic title*
1	teacher, movement, force, momentum, connection, result, learned, system, vacuum, faulty, rotary axis	neglect, objective, evaluation, discussion, precondition			swing, hammer throw, spinning top, ferry		*general concepts focusing subject*
2	correct, explain, teacher, children		imagine		car, drive, mountain, bend, driving	illustrative material	*usage of examples*
3	teacher, momentum, autonomous, forget, basic knowledge, subject matter		imagine, imagination	actio	ball, roll		*general concepts focusing knowledge*
4	body, teach			$v, f, 0, p, p = mv, \sim,$ actio, reactio	apple, boat, ball, mountainous		*symbolic descriptions*
5	body, movement, physical, force, teacher	physics' methods, working methods, knowledge acquisition	student cognition, everyday life experience, cognitive, concept, conflict, reference to everyday life, re-interpret		wire		*student cognition*
6			imagination	actio, reactio	car, accelerate, acceleration, driver, billiard balls, coin	illustrate, graph, thought-provoking impulse, diagram, show	*reasoning on examples*

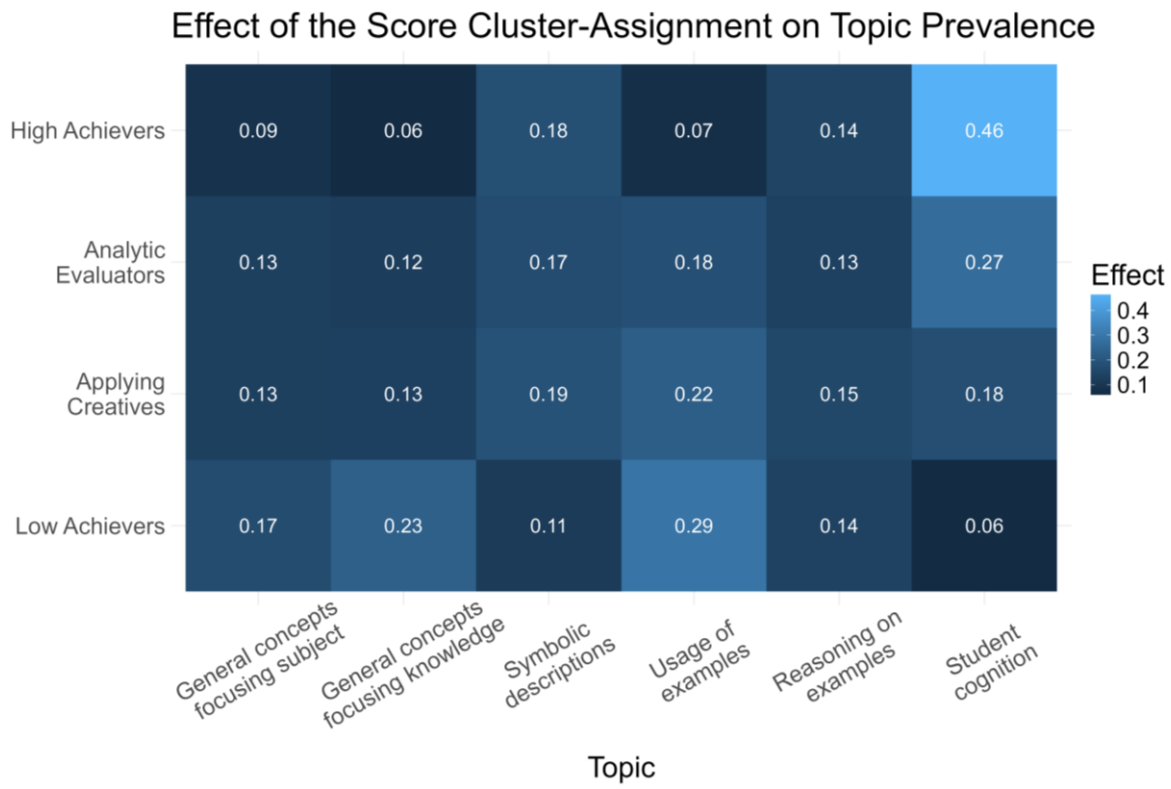


Figure 5.7 Visualizations of the effect that the assignment of a document to a cluster has on the proportion of documents dedicated to a specific topic. For example, the average effect of 0.46 that belonging to the High Achievers has on the topic of Student cognition. This indicates that, on average, 46 % of each document of a High Achiever is dedicated to the topic of Student Cognition.

5.5.3 RQ3: Prediction of Competency Profiles

To compare the logistic regression classifier to a baseline, a dummy classifier is set up that simply predicts the most frequently occurring competency profile. This dummy classifier reaches an average accuracy of $.453$ and an average Cohens κ of 0. The logistic regression classifier achieves excellent prediction accuracy, both in absolute terms and compared to the dummy classifier, as presented in Table 5.5.

Table 5.5 Pattern confirmation: Predictive power of the logistic regression classifier on the test dataset. The table contains the evaluation of the predictive power of the logistic regression classifier predicting the competency profiles from the scores are presented. All values are obtained from the test dataset. In the first row, the metric's value averaged over the 10 cross-validation splits (times 10 different random seeds) is denoted. In the second row 95 % Gaussian confidence intervals of across the 100 values in total are presented. CI=Confidence Interval, LL=Lower Limit, UL=Upper Limit.

	Accuracy	F_1	**Cohen's κ**
Average value	$.943$	$.943$	$.918$
95% CI [LL, UL]	$[.939, .948]$	$[.938, .948]$	$[.911, .925]$

5.5.4 Summary of the Competency Profiles

To further discuss the results of the cluster and language analyses below, we present the tendencies shown in our findings on RQ1 and RQ2 competency profile-wise:

- *Low Achievers*: The Low Achievers show a significant focus on general concepts (focusing knowledge) with an emphasis on using examples. It is important to note that this observation is relative to the normalized topic distribution. In other words, the Low Achievers do not necessarily use more examples than the High Achievers. However, the Low Achievers do devote a greater proportion of their produced text to the topic of using examples compared to the High Achievers. This may be because the description of some general PCK-related topics or simple examples of physical phenomena might be already available at lower levels of pPCK. It is important to note that simply addressing these topics does not necessarily imply a high level of quality in the accompanying text segments.
- *Applying Creatives*: The Applying Creatives also focus on the use of examples, but in addition, they incorporate the reasoning behind the use of examples and the student's cognition as well as symbolic descriptions more strongly into their text productions. As a result, examples are more integrated with other pPCK-related concepts and additional aspects are considered in the thinking process.
- *Analytic Evaluators*: The Analytic Evaluators place a greater emphasis on student cognition, while still dedicating a reasonable amount of their writing to examples. They appear to focus slightly less on the *using examples* and *reasoning about examples* topics than the Applying Creatives, although this difference is not statistically significant. However, this is consistent with their typical scores on the respective requirement dimensions: Analytical and evaluative tasks typically (sometimes even explicitly) require to consider student cognition, while applied and creative tasks often require this only implicitly upon closer examination.
- *High Achievers*: The High Achievers show a strong focus on student cognition. As a result, the proportion of all other topics decreases, with the exception of the *reasoning on examples* topic. However, their scores indicate that they also achieve comparatively high scores on the *create* and *apply* requirement dimensions. These observations suggest that they integrate their creative text elements (such as the description of examples) in a much more theoretically informed manner, e.g., with additional reasoning about their usefulness and the consideration of student cognition.

5.6. Discussion

High-quality teacher knowledge is a crucial prerequisite for effective teaching and learning (Hattie, 2003, 2012; Hume et al., 2019). PCK as a central component of teachers' knowledge (Shulman, 1986, 1987) has therefore been the subject of intense research (e.g., Behling et al., 2022a, 2022b; Hume et al., 2019; Kulgemeyer et al., 2020; Mientus et al., 2022; Schröder et al., 2020; She et al., 2024; Sorge et al., 2019). Recently, hierarchical competency level models

of (preservice) physics teachers' PCK have been developed using item-response models (Schiering et al., 2023; Zeller et al., 2024). Based on the used test instruments such models can be interpreted as describing pPCK through the lens of the RCM of PCK (Carlson et al., 2019). At the same time, research has been conducted on ePCK in the context of the proposed PTR cycle (Alonzo et al., 2019) to shed light on the processes behind the development of ePCK. Therefore, a non-hierarchical distinction has been made between the three components $ePCK_{plan}$, $ePCK_{teach}$, and $ePCK_{reflect}$. We conducted a theoretically guided, exploratory, in large part computational analysis to determine whether similar non-hierarchical structures could also be identified for pPCK. Such empirically based structures can be used to further develop the RCM and to provide meaningful feedback when the corresponding test instruments are used as assessment tools. Therefore, we searched for prototypical response patterns in the scores and textual responses of participants in a large pPCK assessment focusing on classical mechanics. Our findings suggest that it is possible to differentiate competency profiles that show specific strengths and weaknesses when considering the requirement dimensions *reproduce*, *apply*, *analyze*, *evaluate*, *create*, and *teaching situation*, that are inspired by the aforementioned item-response-based results. In addition, we were able to show that individuals belonging to specific competency profiles show tendencies to focus on certain topics in their language use when responding to the open-ended tasks of our pPCK questionnaire.

5.6.1 Interpretation of the Competency Profiles

The typical scores achieved in the requirement dimensions are the core of the competency profiles. These typical scores have been identified by the cluster analysis (RQ1). The corresponding score clusters also form the basis for assigning test persons to a competency profile. Two competency profiles were identified as typical low and high-achieving students with the Low Achievers representing the largest group in our sample. Additionally, it is worth noting that even the High Achievers have plenty of room for improvement (see Figure 5.6, top). This is not surprising, as the questionnaire has previously been shown to be challenging even for well-advanced pre-service teachers (Gramzow, 2015).

In addition to the Low Achievers and High Achievers, which still show a strong hierarchical characteristic (see Table 5.3), two other competency profiles could be identified. Based on their typically reached scores in the requirement dimensions the two additional competency profiles were labeled as Applying Creatives and Analytic Evaluators. The Applying Creatives show a comparatively much higher score on the requirement dimensions aimed at applying PCK to described situations or generating elements of instructional actions descriptively. The Analytic Evaluators show a comparatively much higher score in the requirement dimensions aimed at using PCK analytically to draw certain conclusions from descriptions in the tasks and at using PCK to evaluate described elements of teaching situations. Although these two competency profiles differ significantly in terms of their total pPCK score and their year of study, we argue that the differences between them should not be perceived as hierarchical, because these differences have small effect sizes and are marginal compared to the gap between them and the Low Achievers or High Achievers. It is highly unlikely that a distinction such as that between the Analytic Evaluators and Applying Creatives would be possible with hierarchical, e.g., item-response-theory-based methods.

Using the topic modeling approach, we were also able to identify patterns of typical language use in the questionnaire edits assigned to a particular competency profile. Overall, the *student cognition* topic was found to be the most significant and was predominantly present in high-performing edits of the questionnaire. Student cognition as a central aspect of PCK (Hume et al., 2019; Shulman, 1987) was a conceptual focus of the questionnaire and our observation of language use underlines the importance of this concept. It is noteworthy that Analytic Evaluators focused significantly ($p = .005$) more on the *student cognition* topic compared to the Applying Creatives.

The topic summarized *as general concepts focusing knowledge* was proportionally less focused by the Analytic Evaluators, Applying Creative, and High Achievers. This could be attributed to increased proficiency, since the higher performing questionnaire edits are generally longer, causing the proportion of the *general concepts focusing knowledge* topic to decrease, while the total amount of text devoted to this topic remained stable. On the other hand, it could be argued that certain terms that characterize this topic may suggest an antiquated transmissive understanding of teaching and learning (e.g., forget, basic knowledge, subject matter), which consequently diminishes as proficiency increases. However, the available evidence is not sufficient to thoroughly confirm such a conclusion. Given the influence of belief structures on performance in authentic teaching situations (e.g., Buehl & Beck, 2014; König, 2012; Kulgemeyer & Riese, 2018) further research in this direction is encouraged.

A similar general observation can be made about the *symbolic descriptions* topic. It seems that a certain level of knowledge regarding symbolic descriptions might be needed to succeed in higher-level activities (such as analyzing or applying pPCK). If the *symbolic descriptions* topic is interpreted as closely related to CK – and possibly also to mathematical knowledge in the field of physics – this observation is consistent with CK analyses that point to the necessity of a certain level of CK for the development of professional competence in general (e.g., Kulgemeyer & Riese, 2018; Sorge et al., 2019).

Considering the PTR cycle in the context of ePCK (Alonzo et al., 2019) our findings suggest the existence of similarly distinct pPCK knowledge domains that can be developed to some extent independently, similar to the distinction already suggested in the theoretical foundations section ("$pPCK_{plan}$" vs. "$pPCK_{reflect}$"). The most interesting areas are the pPCK domains aimed primarily at analytical, evaluative tasks, such as deriving students' ideas from a given dialogue ("$pPCK_{eval}$"?), and the pPCK domain aimed primarily at applicative, creative tasks, such as describing a suitable example experiment or real-world example to demonstrate a physical phenomenon ("$pPCK_{apply}$"?). At first glance, the former might be more closely associated with $ePCK_{reflect}$ and the latter with $ePCK_{plan}$. Given the exploratory nature of the analysis, these parallels do not arise primarily from information introduced into the analysis by the analysts, as would be the case in a confirmatory analysis. Therefore, we can frame our results as both a genuine, standalone description of pPCK as well as a link between the fine structures of ePCK and pPCK. An analogous concept for the PTR cycle in the context of ePCK might be a "reproduce apply evaluate" cycle in the context of pPCK.

5.6.2 Scope of Validity and Open Questions

In general, it is important to note that our analysis is based on data that has been collected using a questionnaire that focuses on classical mechanics in terms of the subject content. This is due to the overarching focus of the respective study (Vogelsang et al., 2019). As the data was collected between 2016 and 2019, it may also be slightly outdated. Nevertheless, given the challenges of recruiting participants, it is unlikely that a comparably large dataset for PCK will be generated soon in the subject of physics. We argue that it remains valuable to use this rich dataset to gain further insights, especially given that the general framework of teacher education in Germany has not changed significantly in recent years.

The analysis presented is an interplay between theoretically guided preparation and interpretation steps and computer-based automated analyses. This is the case due to our adoption of the CGT as a methodological framework (Nelson, 2020) and the data mixture of manually assigned scores and real text responses. First, the preparatory step of assigning tasks to requirement dimensions has implications for the interpretation of the resulting cluster structures. Indeed, the assignment of questionnaire tasks to the requirement dimensions is critical to the formation of the resulting clusters. One could argue that by using these dimensions, the resulting clusters differ in the dimensions in which they are allowed to differ and that subjective interpretations and beliefs may have overly biased the task assignment. We addressed these concerns by basing the requirement dimensions on previous results that followed a more inductive approach and by assigning the tasks with great care, secured by an analysis of inter-rater agreement and a consensus solution. Thus, we sought to find an appropriate balance between maintaining the exploratory intent of the analysis and supporting the interpretability of the results. In the following, further research is suggested to confirm and improve the presented results.

A similar concern arises concerning the level of granularity in certain requirement dimension scores (see Table 5.2). Retrospective modification of the questionnaire was not feasible, resulting in few distinct levels in some of the dimensions due to the low number of tasks related to these dimensions in the questionnaire. However, in the subsequent analysis, the results never rely on a single requirement dimension, providing some relief from this concern. Nevertheless, we suggest that all requirement dimensions should be considered a priori in future (quantitative) research on the internal structure of the pPCK. The granularity and thus deviations from Gaussian-like distributions, as well as the varying density of the score data, hindered the application of more sophisticated clustering procedures (e.g., Campello et al., 2013; Spurk et al., 2020). Furthermore, the procedures used to determine an optimal number of clusters for the K-Means algorithm were inconclusive and only roughly favored smaller cluster numbers. The final cluster number of 4 was chosen primarily based on theoretical and practical considerations. Therefore, the competency profiles should only be considered as "latent" groups to a limited extent. A more appropriate perspective might be to consider them as more informed, multidimensional quantiles.

The description of the typical language use of the competency profiles provides evidence for the concurrent validity of the competency profiles with the tendencies being in line with a priori expectations. Similarly, the average scores of the competency profile groups in the

teaching situation requirement dimension provide a compelling argument in support of this assertion, as they align with the typical scores observed in the *analyze* and *evaluate* dimensions. However, in retrospect, it can be argued that the *teaching situation* dimension is superfluous and appears to be an impurity in the otherwise fully theoretically motivated requirement dimensions. Therefore, we re-evaluated the whole score cluster analysis procedure without this dimension. The results for the cognitive activities were almost indistinguishable from those presented in this article.

In addition to the arguments for concurrent validity, the predictability of the competency profiles for unseen data can also be seen as an argument for the validity of their differentiation in the CGT framework. The power of the (rather simple) logistic regression model in predicting the competency profiles also shows that the overlap of the clusters in their two-dimensional PCA visualization (Figure 5.5) does not imply an indistinguishability of the competency profiles. This was also true for several other classification models evaluated for the same purpose as the logistic regression classifier, namely, a support vector machine, a random forest model, and a neural network.

The language analysis yielded several interpretable topics, although it is limited by the digitization process. The initial assessment was carried out in a paper-pencil setting and the responses were later digitized, allowing for the introduction of spelling errors and other inaccuracies. Using an automated approach to correct such errors is not appropriate because (1) it would also correct legitimate errors that were made by the participants, and (2) it is difficult to apply an automated approach to a dataset that contains a significant amount of specialized vocabulary. Nevertheless, the interpretability and appropriateness of the identified topics suggest that the results retain their significance and meaningfulness.

The impact of the competency profile assignments on the topic prevalences (as shown in Figure 5.7) is relatively small, with the differences often being not statistically significant in post hoc test analyses. Nevertheless, we argue that the results combined with the findings from the score cluster analysis provide valuable insights into the competency profiles and their integration into our interpretation is valid. Language data typically embodies a large amount of variety and variation (e.g., Jurafsky & Martin, 2024), so we do not expect the found effects to be large. Our pattern confirmation step does not (yet) make use of the language data and therefore cannot be considered as confirmation of the observations on typical language use. The usage of language data for predictive purposes in an automated scoring workflow is part of our future work (see below) and could then provide further pattern confirmation arguments.

5.6.3 Perspectives and Outlook

Overall, the identified and described competency profiles represent to a certain extent non-hierarchical distinctions of typical pPCK profiles, indicating knowledge areas tailored for different operations within the context of pPCK. In particular, the identified competency profiles reflect a differentiation between more analytical/evaluative and more applicative/creative knowledge. The formation of the competency profiles also indicates that these knowledge areas can be developed to some extent independently of each other, although they do not form genuine “latent” groups.

These findings provide opportunities for further research aimed at describing the internal structure of PCK. First, we suggest that requirement dimensions, such as those relevant to the competency profiles, are included in future quantitative test instruments for pPCK. Based on our findings we argue that a strong emphasis on such a dimension is useful, especially if differentiating specific strengths and weaknesses of participants is relevant for the intended use of the test instrument.

Second, we encourage research to assess the reproducibility of our findings using other pPCK-test instruments and datasets. To gain further insight into the role of language for PCK, it would be beneficial to examine this relationship with a similar methodology but larger data sets. For instance, more complex linguistic features such as whole expressions like "student cognition" could be analyzed rather than just single words, as was the case in our study. The results of this study primarily demonstrate the focus of the competency profiles on specific core ideas of PCK. However, there may be potential connections between the level of PCK demonstrated by a student and the degree of connectedness and sophistication in his/her language use. Similarly, given our data, it remains uncertain whether the identified pPCK domains we proposed above ("$pPCK_{apply}$" and "$pPCK_{eval}$") are specifically associated with particular ePCK domains in the context of the PTR cycle. It is plausible that the applicative/creative pPCK components are more relevant to $ePCK_{plan}$, while the more analytic/evaluative pPCK components are more relevant to $ePCK_{reflect}$. On the contrary, all pPCK components may be necessary to enable effective performance in the form of ePCK, regardless of the concrete focused step in the PTR cycle (or similar distinctions).

Moreover, by choosing (mainly) the cognitive requirement dimensions for the preparation and interpretation of the score cluster analysis, we forced the potential discoveries to be related to these dimensions. However, we cannot and do not rule out the existence of non-hierarchical structures w. r. t. other domains of the task development model, especially the pPCK subscales. Conversely, we even conducted a comparable score cluster analysis for the pPCK subscales[45], which yielded similar results: an overall low-performing group, an overall high-performing group, and two intermediate groups. One of these intermediate groups showed strengths on the *instructional strategies* and *experiments* subscales and one showed strengths on the *student cognition* and *theoretical PCK-related concepts* subscales. However, as already mentioned in the Methods section, the results in the context of the pPCK subscales are less generalizable than the results in the context of the cognitive requirement dimension, because only four selected subscales are targeted in the test instrument, while other subscales are considered in other studies (e.g., Hume et al., 2019).

Although the identified competency profiles cannot be considered as genuine latent groups, i.e., pre-service teachers, in general, may not be strictly limited to the profiles found, there are undoubtedly individuals who exhibit distinct strengths and weaknesses that can be interpreted through our findings. The competency profiles and requirement dimensions therefore represent a step towards empirically based formative assessment (Hattie & Timperley, 2007) that goes beyond a simple differentiation between "lower" and "higher" pPCK. Classifying individual

[45] Abbildung 5.10

student teachers or whole learner groups based on their competency profiles could guide decisions about the selection of specific learning materials and contexts. For example, a group that is more analytical or evaluative could be directed towards more creative tasks, such as producing teaching materials, while more practical or creative students could be encouraged to improve their ability to evaluate or reflect on teaching quality. Such applications of an assessment based on the competency profiles and requirement dimensions could be used to guide the selection of learning opportunities or exercises, as well as to evaluate the effectiveness of teacher education courses or practical training in schools. For instance, a shift in the allocation of a considerable number of participants from the Applying Creatives (pre) to the High Achievers (post) could indicate the intended effectiveness of a course. Contrarily, an unsystematic shift of participants between the Applying Creatives and the Analytic Evaluators profiles could indicate a lack of opportunity for the participants to connect their knowledge in the respective areas.

To make such an assessment feasible and potentially even scalable, automation is required. One option is to convert questionnaires to a closed format, but this could raise uncertainty and concerns about the authenticity of the tasks (e.g., Kulgemeyer et al., 2023) and thus uncertainty about the validity of conclusions drawn from data generated using these newly generated test instruments. Instead, ML and natural language processing techniques could be used to automate the assessment while adhering to already established, (mostly) open-ended test instruments. An automation strategy following the latter approach using a language model is currently being explored for the questionnaire used in this study. During the exploration, we have already drawn two essential conclusions when working on such automation approaches with a data structure as the one described above: (1) an automated scoring step to assign previously manually generated scores to individual tasks seems to be crucial before using these automated scoring results for the comprehensive competency profile classification; (2) the use of a few multiple-choice tasks substantially increases the predictive power, which is helpful in settings where the available training data is too limited to achieve high accuracies for the open-ended tasks. Using transformer language models (Devlin et al., 2019; Vaswani et al., 2017) for the automated scoring step, we are currently achieving $> 70\,\%$ accuracy in competency profiles assignment, approaching human-human-agreement values.

Author Contributions

- *Data acquisition*: Josef Riese (and other researchers not being authors here)
- *Method and analyses*: Jannis Zeller
- *Results interpretation*: Jannis Zeller
- *Initial draft*: Jannis Zeller, Josef Riese
- *Revisions*: Jannis Zeller, Josef Riese
- *Funding acquisition*: Jannis Zeller, Josef Riese

Funding

The project "Professional Competence in Physics' Teacher Training" (ProfiLe-P) was funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) within the framework program "Modelling and Measuring Competencies in Higher Education" (KoKoHs) under the label 01PK15005A-D.

The data used here originates from the above-mentioned project. The manuscript was written as part of a cumulative doctorate funded by a doctoral scholarship from the German Academic Scholarship Foundation (Studienstiftung des Deutschen Volkes).

Conflict of Interest

All authors declare that there are no conflicts of interest.

5.7. Kommentare und Ergänzungen

Zunächst einmal soll hier erneut explizit darauf hingewiesen werden, dass in den explorativen Analysen zum zweiten Zielpaket komplexe Workflows durchlaufen werden. Das Ergebnis des zweiten Artikels sollte daher eher als *eine* Möglichkeit nicht-hierarchische Strukturen des FDW zu beschreiben aufgefasst werden, denn als *die* Möglichkeit. Auch, wenn gezielt die Dimension der kognitiven Anforderungen genutzt wurde, um vor dem Hintergrund der Ergebnisse aus dem ersten Zielpaket die Wahrscheinlichkeit zu erhöhen, dass die Strukturen auch projektunabhängig generalisierbar sind, wird hier weitere entsprechende Forschung zur Überprüfung der Generalisierbarkeit der gefundenen Strukturen als notwendig angesehen.

Die ermittelten Cluster bzw. insbesondere die Analytic Evaluators und Applying Creatives stellen nicht-hierarchische Gruppen dar. Dies wurde im Artikel aus Platzgründen nur recht knapp prosaisch beschrieben und soll daher hier noch einmal ausführlicher nachgereicht werden. In Tabelle 5.6 werden daher paarweise T-Test zum Vergleich der Cluster berichtet. Zwischen den Applying Creatives und den Analytic Evaluators zeigen sich lediglich nicht bzw. nur knapp signifikante Unterschiede bezüglich des Studienfortschritts und des FDW-Gesamtscores. Insbesondere beim Vergleich der Effektstärken kann hier in diesem Sinne definitiv von einer nicht-hierarchischen Unterscheidung ausgegangen werden.

Tabelle 5.6 Paarweise T-Test zum Vergleich der Cluster aus Artikel 2.

	Studienfortschritt (Years of Study)		**FDW-Gesamtscore**	
	$t(\mathbf{df}), p$	**Cohens *d***	$t(\mathbf{df}), p$	**Cohens *d***
Low Achievers vs. Applying Creatives	$t(497) = 6.56$ $p < 0.001$	0.61	$t(497) = 22.25$ $p < 0.001$	2.08
Applying Creatives vs. Analytic Evaluators	$t(320) = 1.67$ $p = 0.10$	-	$t(320) = 2.01$ $p = 0.05$	0.23
Analytic Evaluators vs. High Achievers	$t(277) = 4.38$ $p < 0.001$	0.53	$t(277) = 15.78$ $p < 0.001$	1.90

5.7.1 Alternative Cluster-Modelle und Subskalen

In Abschnitt 5.4.2 wurde beschrieben, dass Dichte-basierte Cluster-Modelle sowie GMMs auf den Datensatz in der bestehenden Form, d. h. aggregiert nach den Subskalen *Reproduzieren*, *Anwenden*, *Analysieren*, *Evaluieren*, *Kreieren* und *Unterrichtssituation*, nicht angewendet werden konnten. Abbildung 5.8 und Abbildung 5.9 visualisieren die dabei auftretende Problematik der „Kollabierung" solcher elaborierteren Modelle aufgrund der geringen Anzahl an Abstufungen – in diesem Fall der Kategorie Evaluieren. Dabei wird ein GMM und ein Dichte-basiertes HDBSCAN[46]-Modell (McInnes et al., 2017) verwendet.

[46] **H**ierarchical **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise

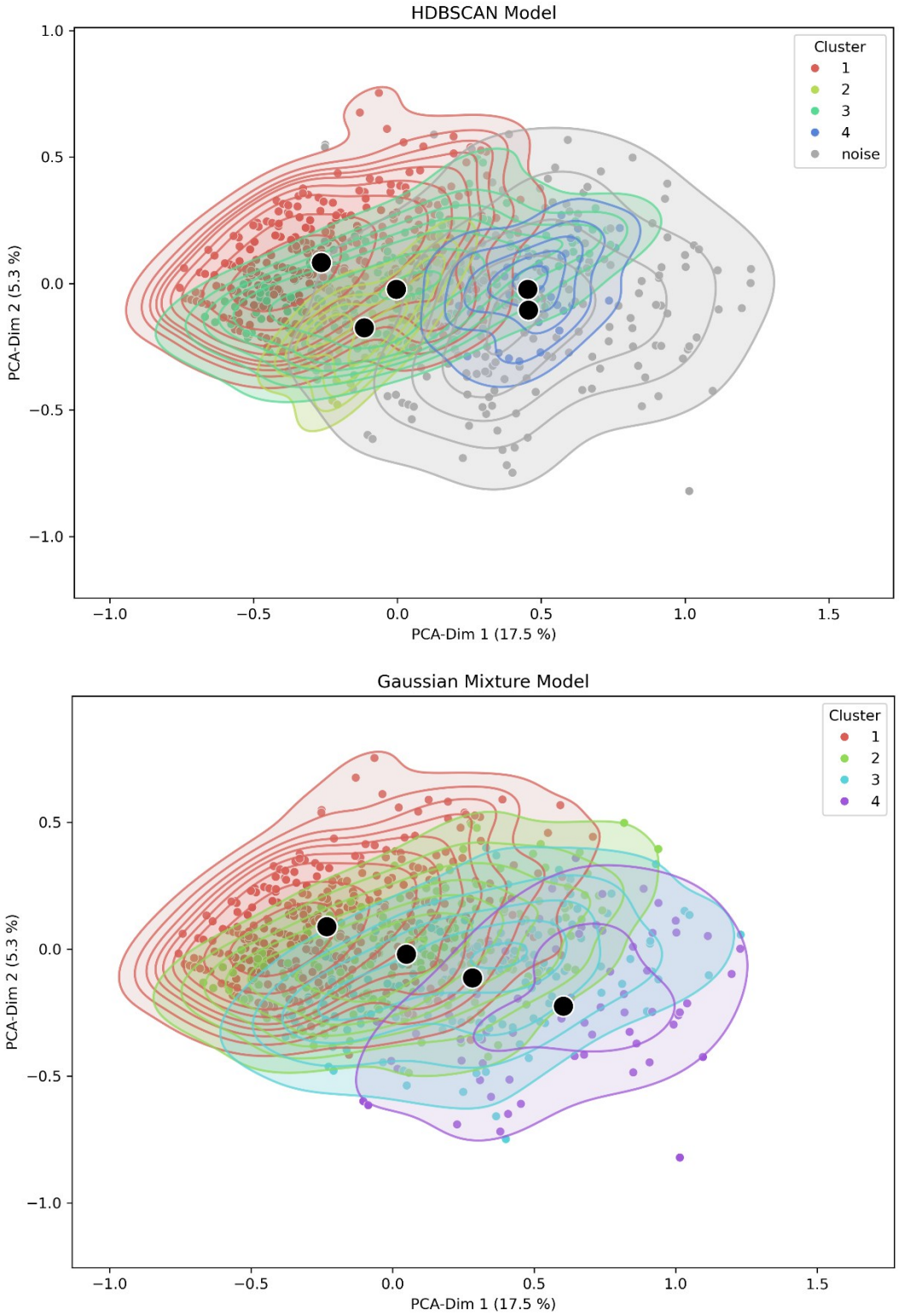


Abbildung 5.8 PCA-Visualisierung von alternativen Score-Clustern bei der Nutzung eines HDBSCAN-Models (oben) und eines GMMs (unten). Im Falle des HDBSCAN-Models deutet sich zudem das beschriebene Problem der zu unsystematischen Dichteverteilungen bereits an, obwohl hier entsprechende Parameter zur „Glättung" bereits eingestellt wurden. Im Falle des GMMs erkennt man eine lineare Struktur. Beides lässt sich mithilfe von Abbildung 5.9 noch einmal expliziter interpretieren.

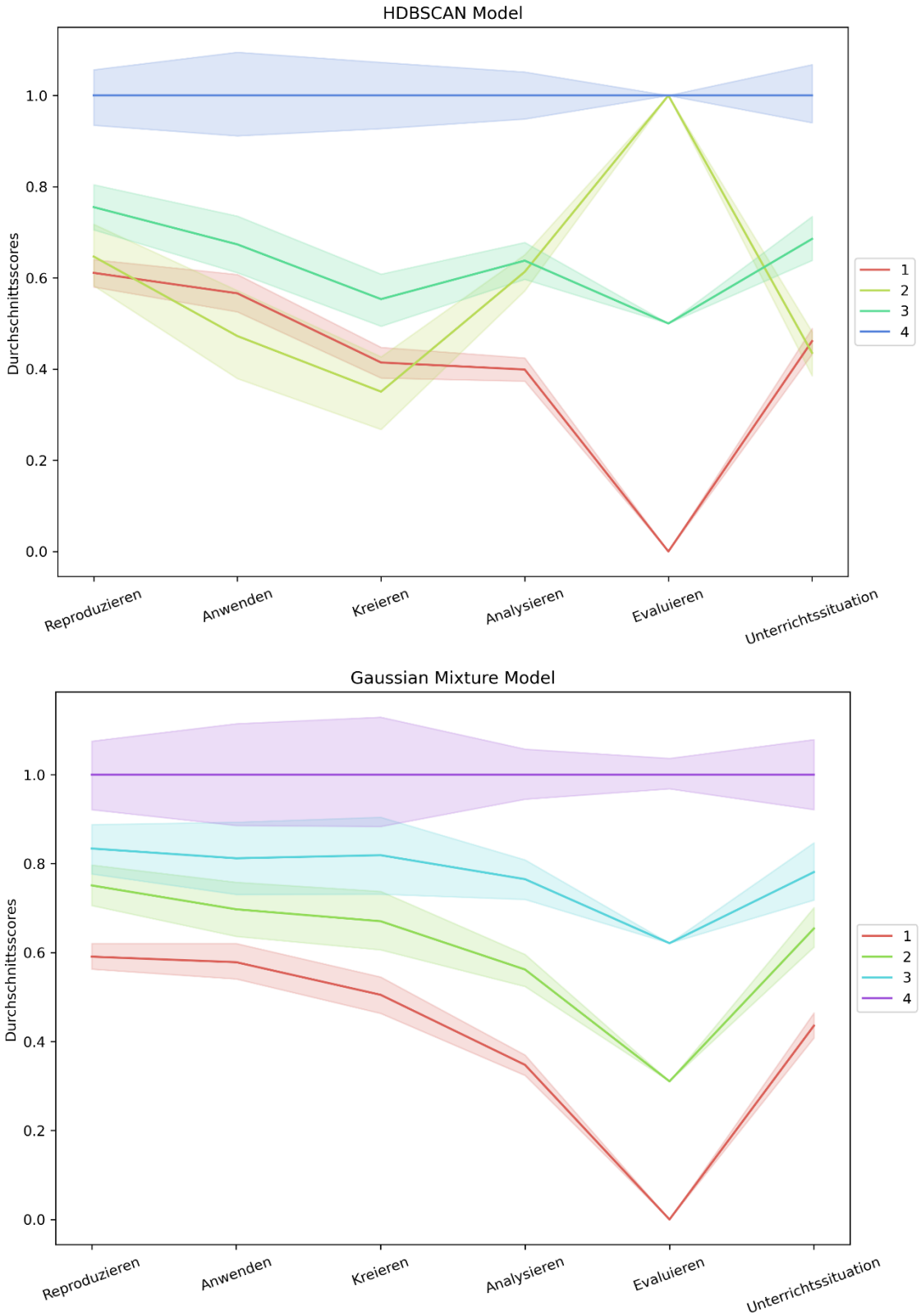


Abbildung 5.9 Zentroid-Linienplots der alternativen Score Cluster bei der Nutzung eines HDBSCAN-Models (oben) und eines GMMs (unten). Man erkennt, wie in beiden Fällen die starke Diskretisierung der Skala Evaluieren dazu führt, dass die Modelle letztlich „kollabieren“ und quasi nur noch die Varianz in dieser einen Kategorie aufklären. Ein K-Means Modell hat sich als gegenüber dieser Problematik als robuster erwiesen und konnte trotzdem informative Cluster liefern.

Diese zusätzlichen Analysen deuten darauf hin, dass mehr Aufgaben bzw. feiner unterteilte Subskalen notwendig (wenn auch nicht unbedingt hinreichend) sind, um elaboriertere Cluster-Modelle einsetzen zu können. Die Ergebnisse des Artikels 2 suggerieren zudem, dass es zu diesem Zweck sinnvoll sein kann, die kognitiven Anforderungen Anwenden und Kreieren bzw. Analysieren und Evaluieren zusammenzufassen. In Artikel 3 (Kapitel 6) wird ebendies angestrebt. Dabei fällt die Wahl der Cluster-Analyse auf eine GMM-basierte LPA (e.g., Spurk et al., 2020), da für diese bereits etablierte Workflows existieren.

Wie bereits in Abschnitt 5.6.3 angedeutet wurde, können auch die Facetten auf analoge Weise wie die in Artikel 2 verwendeten sechs Anforderungskategorien zur Clusterbildung genutzt werden. Ähnlich wie beim K-Means Modell in Artikel 2 gibt es auch hier kaum nennenswerte heuristische Argumente für eine bestimmte Anzahl an Clustern (siehe Elbow- und Silhouette-Plots im digitalen Ergänzungsmaterial). Ein Zentroid-Linienplot eines K-Means Modells mit 4 Clustern ist in Abbildung 5.10 dargestellt. Auch hier zeigt sich eine potenzielle nicht-hierarchische Struktur in Form zweier Gruppen, die sich insbesondere in ihren Kompetenzen hinsichtlich der Facetten Instruktionsstrategien und Experimente unterscheiden. Wie bereits angesprochen, stellen diese im Testinstrument abgedeckten Facetten aber nur eine Auswahl möglicher Facetten dar (Gramzow, 2015). Somit ist eine Übertragbarkeit auf andere Operationalisierungen bzw. Konzeptualisierungen des FDW weniger wahrscheinlich als bei den kognitiven Anforderungen.

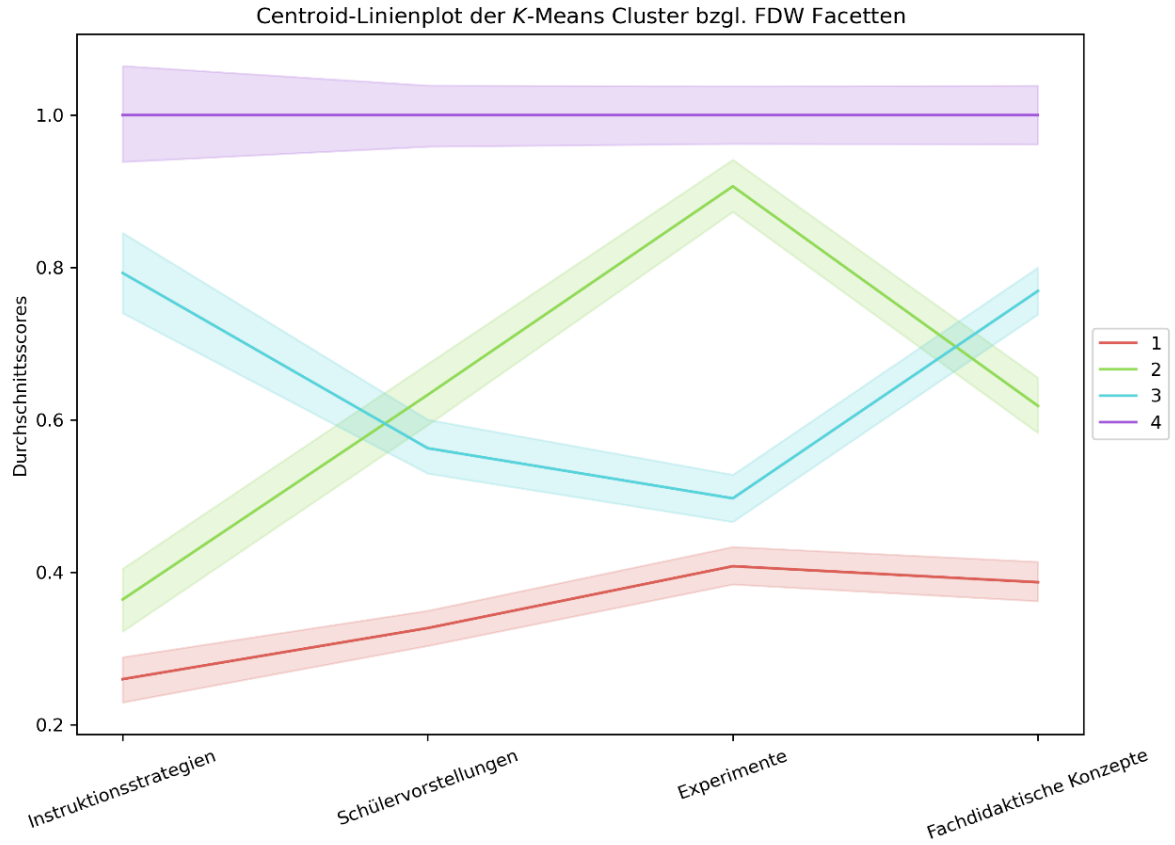


Abbildung 5.10 Zentroid-Linienplot für ein K-Means Cluster-Modell auf Basis der FDW-Facetten.

5.7.2 Hinweise zum Topic Modelling und Alternativen

In Artikel 2 wurde ein STM (Roberts et al., 2019) zur Analyse von typischen Sprachgebrauchsmustern der Personencluster durchgeführt. Zentrales Ergebnis dieser Untersuchung sind die Effekte, die die Zuordnung zu einem Cluster auf die Fokussierung auf bestimmte Topics hat. Dabei wurde bereits angedeutet, dass die Darstellung im Artikel (Figure 5.7) deskriptiv abgeleitet wurde[44]. Die verwendete Software stellt allerdings auch eine Methode bereit, diese Effekte probabilistisch zu untersuchen[47]. Da im Rahmen des probabilistischen STMs ohnehin die Verteilung der Topic-Cluster Variablen ermittelt wird, kann man diese auch direkt aus dem Modell samplen und erhält somit auch ein Maß für die Unsicherheit der Werte (siehe digitales Begleitmaterial). Die Unterschiede der probabilistischen Erwartungswerte der Cluster-Topic-Effekte zu den in Artikel 2 verwendeten, leichter interpretierbaren, frequentistischen Werten sind aber gering (siehe Abbildung 5.11).

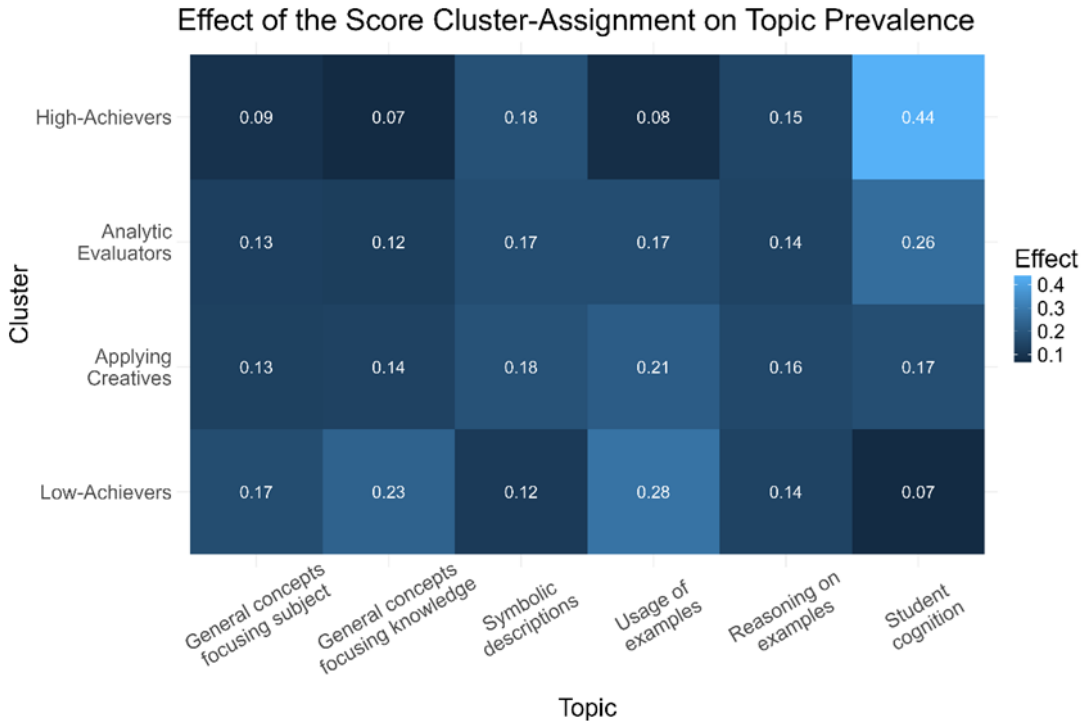


Abbildung 5.11 Darstellung des Effekts, den die Cluster-Zugehörigkeit auf den Anteil hat, den das entsprechende Dokument einem Topic widmet (probabilistische Betrachtung).

Als Alternative zum STM wurden in Artikel 2 auch Deep-Learning-basierte Topic Models, der Namensgebung der verwendeten Software hier auch BERTopic-Modelle (Grootendorst, 2022) genannt, erwähnt. In solchen Modellen wird typischerweise ein Cluster-Modell mithilfe von dimensionsreduzierten[48] Embeddings der Dokumente gebildet. Anschließend werden die

[47] Siehe „`estimateEffect`“ in Roberts et al. (2023, S. 11–13).

[48] Zur Dimensionsreduktion wird hierbei standardmäßig (Grootendorst, 2022) das sog. „UMAP“-Modell (**U**niform **M**anifold **Ap**proximation; McInnes et al., 2020) verwendet, dass gegenüber klassischen Vorgehensweisen wie einer PCA den Vorteil hat, dass sowohl lokale als auch globale Strukturen in der Projektion auf die niederdimensionale Darstellung erhalten bleiben.

charakteristischen Worte aus den Dokumenten, die zu einem Cluster gehören, auf Basis ihrer Frequenz extrahiert. Im Rahmen dieses Projekts wurden neben dem STM auch zwei BERTopic-Modelle erstellt. Diese bieten zwar nicht wie das STM die Möglichkeit die Topic-Cluster-Zusammenhänge schon in der Modellierung zu berücksichtigen, sind aber interessant, da sie die Bag-of-Words Annahme überwinden und sich in unterschiedlichen Kontexten als sehr informativ und reichhaltig erwiesen haben (Grootendorst, 2022).

Im ersten Experiment wurde dabei ein BERTopic-Modell der Gesamtantwortdokumente, analog zum Vorgehen in Artikel 2 erstellt. Das heißt, alle Antworten einer Bearbeitung des Testinstruments wurden gemeinsam als ein Dokument betrachtet. Aufgrund der durchschnittlichen Länge dieser Dokumente von ca. 311 ($SD = 125$, $\max = 892$) Worten, stoßen hier kleine BERT-Sprachmodelle bereits an ihre Grenzen. Stattdessen wurde das „text-embedding-3-small"-Modell von OpenAI (o. D.-b) verwendet[49], das Input-Längen von über 8000 Token verarbeiten kann. Darüber hinaus wurde im Wesentlichen der Standard-Workflow des verwendeten Bertopic-Python-Pakets (Grootendorst, 2022) beibehalten. Tabelle 5.7 zeigt die sich ergebenden Wortlisten zur Charakterisierung möglicher Topics und die Anzahl an zugeordneten Dokumenten. Aus Sicht des Autors sind diese „potenziellen Topics" sehr allgemein und repetitiv. Sie unterscheiden sich, wenn überhaupt, dann lediglich bezüglich fachlicher und nicht fachdidaktischer Konzepte. Sie sind für ein sinnstiftendes Pattern Refinement aus Sicht des Autors nicht brauchbar[50].

Noch eindeutiger ist die Lage, wenn man statt der Gesamtbearbeitungen die einzelnen Antworten als Dokumente nutzt. In diesem Kontext wird jede Antwort demjenigen Cluster zugeordnet, dem auch die Gesamtbearbeitung zugeordnet ist. Da die Einzelantworten deutlich kürzer sind als die Gesamtdokumente, kann hier zur Berechnung der Embeddings wieder ein BERT-Modell verwendet werden. Die sich ergebenden Topic-Begriffslisten deuten bereits darauf hin, dass in diesem Modell vielmehr die jeweilige Aufgabe, zu der eine Antwort gehört, eine Rolle für das Topic spielt, als das Personencluster, zu dem die Gesamtbearbeitung gehört. Zu Illustrationszwecken werden hier anstelle einer Tabelle die Topics als Embeddings-Cluster in Abbildung 5.12 dargestellt, wobei die jeweils wichtigsten Begriffe ebenfalls enthalten sind. Abbildung 5.13 und Abbildung 5.14 visualisieren den Zusammenhang zwischen Clustern und Topics bzw. Aufgabenzuordnungen und Topics noch einmal und bestätigt den Eindruck, der bereits durch die Topic-Begriffslisten entsteht: Die Topics sind primär durch die zur jeweiligen Antwort gehörigen Testaufgabe und nicht durch die Kompetenzprofil-Zugehörigkeit der Antwort-Autorin bzw. Antwort-Autoren charakterisiert. Dass beim BERTopic-Modell anders als beim STM die Clusterzuordnung eines Dokuments erst nach der eigentlichen Modellierung und lediglich deskriptiv genutzt wird, scheint für den hier verfolgten Anwendungszweck ein entscheidender Nachteil zu sein.

---

[49] Bei der Nutzung der OpenAI-API zur Generierung von Embeddings (Abschnitt 5.7.2 & 6.7.6) und zum Finetuning (Abschnitt 6.7.2 & 6.7.8) spielen Datenschutz und Privatsphäre eine Rolle. Es wurde hier (1) bereits bei der Digitalisierung der Testantworten (Abschnitt 3.2, 6.4.1 & 6.7.3) darauf geachtet, dass keine persönlichen oder sensiblen Daten im Datensatz enthalten sind und (2) die OpenAI-API derart konfiguriert, dass keine Daten gespeichert oder zum Training der öffentlichen Modelle verwendet wurden (siehe auch OpenAI, 2024a).

[50] Alle Leser:innen dieser Arbeit sind herzlich eingeladen, selbst kreativ zu werden.

Tabelle 5.7 Charakteristische Begriffe eines Deep-Learning-basierten Topic Models für die Gesamtbearbeitungen als Dokumente.

Topic Nummer	Anzahl an Dokumenten	Charakteristische Begriffe
1	117	schülervorstellungen, physikalische, physikalischen, kraftbegriff, physik, kräfte, bewegungsrichtung, zentripetalkraft, reibungskraft, kreisbewegung
2	109	schülervorstellungen, kräfte, physikalischen, zentripetalkraft, vorstellungen, fragen, bewegung, messfehler, versuch, schneller
3	91	schülervorstellungen, kräfte, reibungskraft, gravitationskraft, physik, zentripetalkraft, beschleunigt, schneller, messfehler, prinzip
4	90	schülervorstellungen, physikalischen, kräfte, physik, zentripetalkraft, schülern, reibungskraft, frage, messfehler, schülers
5	86	kräfte, schülervorstellungen, physik, reibungskraft, zentripetalkraft, beschleunigen, schneller, messfehler, bewegung, messunsicherheiten
6	80	schülervorstellungen, beschleunigt, schülern, kräfte, physik, schneller, schülers, beispiel, schülerinnen, vorstellungen
7	77	kräfte, kreisbewegung, reibungskraft, physikalische, schülervorstellungen, zentripetalkraft, bewegungsrichtung, bewegung, zentrifugalkraft, beschleunigt
8	71	schülervorstellungen, schülern, schülers, fragen, versuch, vorstellungen, beispiel, reibungskraft, fehler, falsch
9	69	schülervorstellungen, physik, beschleunigt, schülern, kräfte, messfehler, schneller, schülers, reibungskraft, zentripetalkraft
10	54	schülervorstellungen, kräfte, reibungskraft, beschleunigt, physik, zentripetalkraft, schneller, frage, zentrifugalkraft, schülern

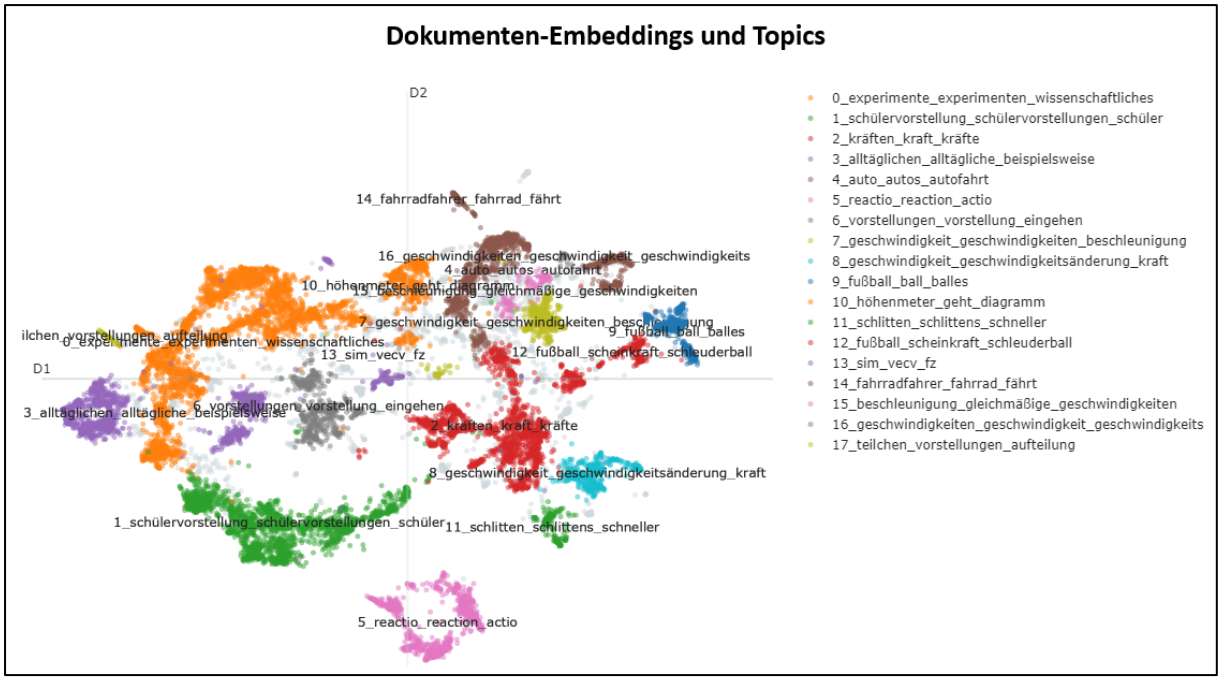


Abbildung 5.12 Darstellung der Dokumente und Topics eines BERTopic Modells mit den Einzelantworten als Dokumente. Verwendet wird hier das das Sentence-BERT-Modell von Reimers und Gurevych (2019), das Abschnitt 6.7.6 noch ausführlicher vorgestellt wird.

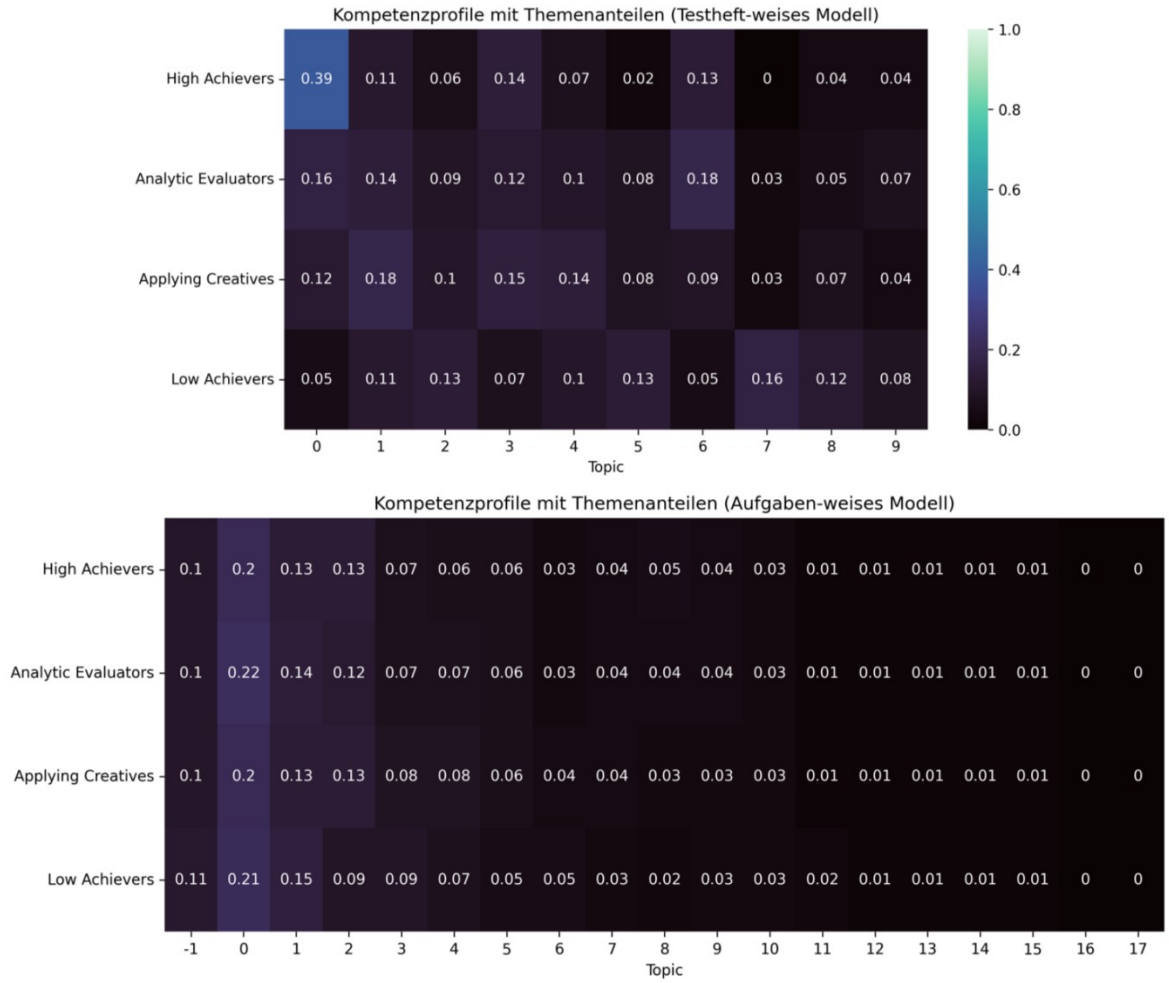


Abbildung 5.13 Cluster-Topic-Zusammenhänge im Testheft-weisen BERTopic-Modell (oben) und im Aufgaben-weisen BERTopic-Modell (unten). Die Werte sind hier etwas anders zu interpretieren als in Figure 5.7 und Abbildung 5.11. Das BERTopic-Modell ordnet jedes Dokument im Cluster Modell der Embeddings einem Topic fix zu. Das STM hingegen ist ein sog. „Soft-Assignment" Modell, d. h. vereinfacht dargestellt, ein Dokument wird den Topics anteilig zugeordnet. Die Werte in den Abbildung 5.13 und Abbildung 5.14 sind daher einfach die Verteilungen aller Dokumente aus den Clustern bzw. Aufgaben auf die Topics. Man erkennt deutlich, dass die fokussierten Topics weitestgehend unspezifisch verteilt sind. Der hervorstechende Fokus des High Achievers Profil auf Topic 0 im Testheft-weisen Modell ist mit den zugehörigen Topic-Begriffe aus Tabelle 5.7 nicht einsichtig.

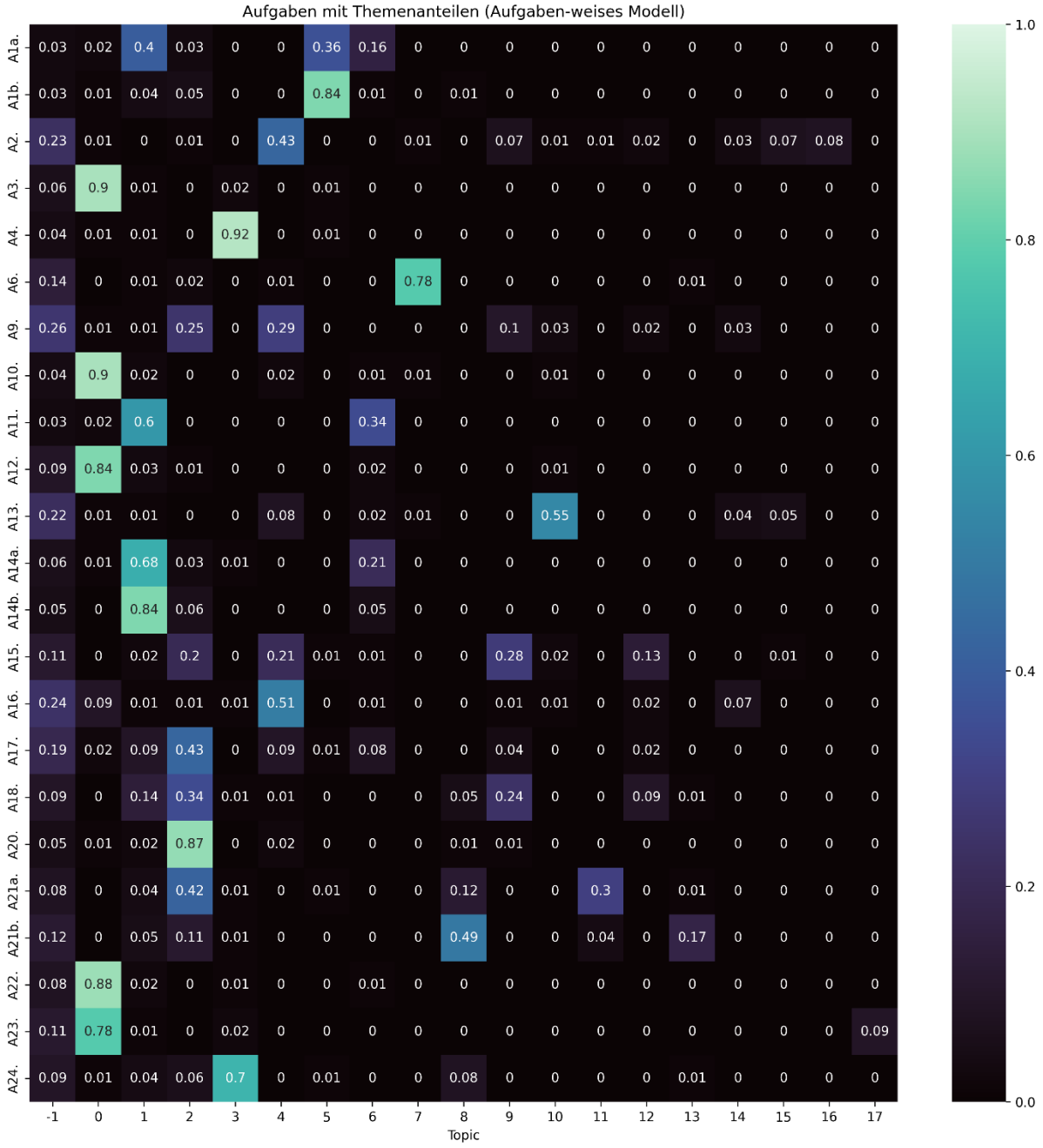


Abbildung 5.14 Aufgaben-Topic-Zusammenhänge im Aufgaben-weisen BERTopic-Modell. Die Darstellung ist analog zu interpretieren wie Abbildung 5.13. Man erkennt deutlich, wie stark die Aufgaben mit einzelnen Topics zusammenhängen. Beispielsweise spielt Topic 0, welches offenbar Experimente im Physikunterricht fokussiert (siehe Abbildung 5.12), eine wichtige Rolle in Antworten zu den Aufgaben A3, A10, A12, A22 und A23. Es überrascht daher nicht, dass die Aufgaben A3, A10, A12 und A22 auch tatsächlich in der Testkonzeption der fachdidaktischen Facette Experimente zugeordnet sind (Gramzow, 2015, S. 276). Andere Aufgaben müssen eher auf Basis ihres konkreten Inhalts betrachtet werden, um die Bezüge zu den entsprechenden Topics zu erklären. Auch, wenn solche Beobachtungen interessante Parallelen (oder auch Diskrepanzen) zwischen der Sprachnutzung von Proband:innen in ihren Antworten und den (intendierten) Inhalten der Testaufgaben ergeben können, sind sie für die Zielpakete dieses Projekts von untergeordneter Bedeutung und werden daher hier nicht ausführlicher dargestellt.

6. Machine-Learning-basiertes automatisiertes Assessment von Kompetenzprofilen des physikdidaktischen Wissens (*Artikel 3*)

Einordnung in das Gesamtprojekt

Der dritte Artikel dieses kumulativen Dissertationsprojekts sollte sich ursprünglich auf die Automatisierung des FDW-Assessments, d. h. primär auf das dritte Zielpaket, fokussieren. Die Cluster-Analysen sollten dabei eher als Ausgangspunkt dienen, an den der Assessment-Workflow angeknüpft werden sollte. Im Review-Prozess des Artikels wurde von Seiten der Reviewenden und der Herausgebenden allerdings eine stärkere Hervorhebung des inhaltlichen Mehrwerts, bzw. des inhaltlichen Erkenntnisgewinns des Projekts gewünscht. Dazu wurde unter anderem vorgeschlagen, die Cluster-Analyse mit in den Artikel aufzunehmen. Anstatt dieselbe Analyse wie in Artikel 2 hier erneut einzureichen, wurde auf Basis der bestehenden Ergebnisse eine neuerliche Cluster-Analyse mit einem veränderten Workflow durchgeführt. Diese ist in Form der Forschungsfrage 2.4 in der Gesamtstrukturierung dieser Arbeit enthalten (siehe Abschnitt 3.1). Zusätzliche Informationen und Ergebnisse zu explorierten alternativen Modellen und Workflows des automatisierten Assessments sind daher aus Platzgründen nicht in Artikel 3 eingeschlossen und folgen in Abschnitt 6.7.

Der Workflow der neuerlichen Cluster-Analyse zu FF2.4 / in Artikel 3 basiert einerseits auf der Beobachtung, dass die Anforderungskategorien Anwenden und Kreieren sowie Analysieren und Evaluieren anscheinend systematisch zusammenhängen (Artikel 2, bzw. Abschnitt 5.5.1, insbesondere Figure 5.6). Darüber hinaus zeigte sich im Rahmen weiterer explorativer Analysen, dass die starke Diskretisierung einzelner Anforderungskategorien (insbesondere Evaluieren) eine wesentliche Limitation des Datensatzes ist und die Anwendbarkeit elaborierterer Cluster-Modelle als dem K-Means-Modell entscheidend beeinträchtigt (Abschnitt 5.7.1). Für eine neuerliche Cluster-Analyse wurden daher die Anforderungskategorien Anwenden und Kreieren sowie Analysieren und Evaluieren zusammengefasst, sodass die einzelnen betrachteten Subskalen weniger stark diskretisiert sind. Im Sinne der CGT kann man die Cluster-Analyse in Artikel 3, d. h. die Analysen zu FF2.4 (Abschnitt 6.4.2 & 6.5.1), also als neuerliche „refined“ Pattern Detection im Rahmen eines zyklischen Durchlaufens der einzelnen Phasen des CGT-Frameworks verstehen.

Die Analysen zur automatisierten Bepunktung des Testinstruments (Abschnitt 6.4.3 & 6.5.2) sowie der darauf aufbauenden automatisierten Zuordnung von Proband:innen zu den Kompetenzprofilen (Abschnitt 6.4.4 & 6.5.3) werden somit auch als Pattern-Confirmation-Schritt interpretiert. Dabei sind die Ergebnisse dieser Pattern Confirmation aber praktisch wesentlich nutzbarer als die Pattern-Confirmation-Ergebnisse in Artikel 2 (FF2.3), da hier der Assessment Workflow nun vollständig automatisiert wird.

Bibliographische Angabe

Zeller, J. & Riese, J. (2025). Machine-Learning-basierte Analyse von latenten Profilen des physikdidaktischen Wissens. *Zeitschrift für Didaktik der Naturwissenschaften, 31,* Artikel 5. https://doi.org/10.1007/s40573-025-00181-y. *Vorab-Print (Stand 01.12.2024)*

Preprint-Statement

This preprint has not undergone peer review (when applicable) or any post-submission improvements or corrections. The Version of Record of this article is published in "Zeitschrift für Didaktik der Naturwissenschaften" and is available online at https://doi.org/10.1007/s40573-025-00181-y.

Hier ist der Artikel auf dem Stand vom 01.12.2024 ähnlich der überarbeiteten Fassung nach der ersten Review-Runde mit zusätzlichen redaktionellen Korrekturen enthalten.

In der hier enthaltenen Version ist der Artikel primär auf die Automatisierung der Zuordnung von Proband:innen zu den Kompetenzprofilen und somit die Entwicklung eines automatisierten FDW-Assessments fokussiert. Im Review-Prozess wurde eine stärkere Fokussierung auf den inhaltlichen Erkenntnismehrwert gewünscht, sodass in der veröffentlichten Version des Artikels ein deutlich stärkerer Fokus auf ausgeschärften inhaltlich-explorativen Analysen – ähnlich zu denen in Artikel 2 – liegt. Das automatisierte Assessment ist in der veröffentlichten Version des Artikels eher „Mittel zum Zweck" und wird nur recht knapp beschrieben. Somit hat sich auch der Titel des veröffentlichten Artikels gegenüber dem Titel des Kapitels 6 in dieser Arbeit verändert.

Zusammenfassung

Das fachdidaktische Wissen (FDW) stellt eine wichtige Komponente des Professionswissens von (angehenden) Lehrkräften dar. Es liegen bereits empirische Forschungsergebnisse zu Einflussfaktoren und zur Entwicklung des FDW sowie zur Bedeutung des FDW für Professionswissen und Qualität professioneller Handlungen vor. Für eine optimale Förderung der Entwicklung des FDW und weitere detailliertere Forschung sind darüber hinaus differenziertere empirisch begründete Beschreibungen der inneren Struktur des FDW notwendig. Bisher sind entsprechende Ansätze allerdings zumeist primär theoretisch-normativ begründet, auf hierarchische Betrachtungen beschränkt, oder nicht in der Lage, tatsächlich *latente* Strukturen zu erfassen. Im vorliegenden Beitrag wird daher ein Ansatz zur datenbasierten Beschreibung latenter Kompetenzprofile des FDW orientiert an der Computational Grounded Theory vorgestellt. Dabei wird zunächst ein Datensatz von 846 Bearbeitungen des Physik-FDW-Testinstruments mit überwiegend offenem Antwortformat aus dem ProfiLe-P+ auf Basis der bisherigen Forschungsergebnisse zur inneren Struktur des FDW vorbereitet. Anschließend wird eine Latent Profile Analysis zur Untersuchung latenter Kompetenzprofile durchgeführt. Um die Ergebnisse im Sinne der Computational Grounded Theory zu bestätigen, wird im Anschluss ein Machine-Learning-basiertes System zur automatisierten Zuordnung von Testbearbeitungen (insbesondere ausgehend von den Freitext-Antworten der Proband:innen) zu den Kompetenzprofilen erstellt. Es zeigen sich vier latente Kompetenzprofile mit nicht-hierarchischem Charakter, die insbesondere auf die Trennbarkeit analytisch-evaluativer und anwendungsorientiert-kreativer Kompetenzen hindeuten. Die automatisierte Zuordnung der Testbearbeitungen zu den Kompetenzprofilen mit einer

Maschine-Mensch-Übereinstimmung von $\kappa = 0{,}587$ (Mensch-Mensch-Baseline: $\kappa = 0{,}624$) kann im Sinne der Computational Grounded Theory als Bestätigung der Validität dieser Strukturen aufgefasst werden. Das dabei entwickelte Machine-Learning-basierte System bietet zudem das Potenzial, für skalierbares automatisiertes inhaltlich reichhaltiges Assessment des FDW genutzt zu werden.

Schlüsselwörter: Fachdidaktisches Wissen · Explorative Analyse · Physik · Machine Learning · Natural Language Processing · BERT-Modell

Machine-Learning-based automated assessment of competency profiles in physics pPCK

Abstract

Personal Pedagogical Content Knowledge (pPCK) represents a crucial component of the professional knowledge of (prospective) teachers. Empirical research has assessed the development and influencing factors of pPCK and shown pPCK's significance for professional knowledge and the quality of professional actions. For the optimal fostering of pPCK (e.g., in teacher education programs) and further research, descriptions of pPCK' internal structure are necessary. However, existing approaches are typically primarily theoretically-normatively grounded, limited to hierarchical views, or unable to capture *latent* structures. We therefore present an approach for data-driven description of latent competency profiles of pPCK, guided by the Computational Grounded Theory. Initially, a dataset of 846 responses to the physics pPCK test instrument from the ProfiLe-P+ - project is pre-processed based on previous research findings. Subsequently, a Latent Profile Analysis is conducted to examine latent competency profiles. To confirm the results in the sense of the Computational Grounded Theory, a Machine-Learning-based system for the automated classification of test responses (particularly from participants' free-text answers) into the competency profiles is developed. Four latent competency profiles, which exhibit a non-hierarchical nature and particularly indicate the separability of analytical-evaluative and application-oriented-creative competencies, are identified. The automated classification of test responses into the competency profiles, with a machine-human agreement of $\kappa = 0.587$ (human-human baseline: $\kappa = 0.624$), can be interpreted as a confirmation of the validity of these structures in the sense of Computational Grounded Theory. Moreover, the Machine-Learning-based assessment-system holds potential for a scalable, automated, content-rich assessment of pPCK.

Keywords: Personal Pedagogical Content Knowledge · Exploratory Analysis · Physics · Machine Learning · Natural Language Processing · BERT-Model

6.1. Einleitung

Erfolgreiche Lehrerbildung stellt im Kontext der Wirkkette schulischer Bildung (Terhart, 2012) und des empirisch belegten Einflusses der Lehrperson auf schulischen Erfolg (Hattie, 2009) eine wichtige Grundlage eines effektiven Bildungssystems dar. Das Lehramtsstudium zielt dabei wesentlich auf die Vermittlung von Professionswissen ab. Eine zentrale Komponente des Professionswissens von Lehrkräften ist neben dem Fachwissen (FW) und dem pädagogischen Wissen (PW) das fachdidaktische Wissen (FDW) (z. B. Baumert & Kunter, 2006; Shulman, 1986). FW beinhaltet dabei das eigentliche „Sachwissen" der jeweiligen Disziplin sowie fachspezifische Arbeitsweisen und Lösungswege. PW umfasst fachunabhängiges Wissen wie beispielsweise Wissen über Klassenführung und Diagnostik. Die Konzeptualisierungen und inhaltlichen Beschreibungen des FDW sind häufig weniger einheitlich (z. B. Gramzow et al., 2013), allgemein kann FDW aber grob als *Wissen über die Vermittlung von bestimmtem Fachwissen and bestimmte Lernende* verstanden werden (siehe Abschnitt 6.2.1, 6.2.2). Konkret für die Naturwissenschaften liegen mittlerweile sowohl im deutschsprachigen (z. B. Riese et al., 2015; Schiering et al., 2019; Tepner et al., 2012) als auch im internationalen Raum (z. B. Hume et al., 2019; Park & Oliver, 2008) Forschungsergebnisse zu Operationalisierungen, Interdependenzen und Einflussfaktoren des FDW vor. Analysen zeigen darüber hinaus die Bedeutsamkeit des FDW sowohl (1) für das Professionswissen als Ganzes (z. B. Hume et al., 2019; Sorge et al., 2019) als auch (2) für die Entwicklung zentraler, unterrichtsbezogener Fähigkeiten (Kulgemeyer et al., 2020; Schröder et al., 2020) sowie (3) für die kognitive Aktivierung von Schüler:innen (Förtsch et al., 2016).

Aufgrund der sowohl theoretisch angenommenen als auch empirisch belegten Bedeutsamkeit des FDW gibt es Bestrebungen, die innere Struktur des FDW inhaltlich zu beschreiben. Dazu sind neben theoretisch-normativen Modellierungen im Rahmen von häufig als „Facetten" bezeichneten Subskalen (z. B. Park & Oliver, 2008; Riese et al., 2017; Sorge et al., 2019) auf empirischer Seite bislang vor allem Niveaumodelle auf Basis von Item-Response-Modellen entwickelt worden (z. B. Schiering et al., 2023; Schiering et al., 2019; Zeller et al., 2024). Projektübergreifend zeigte sich dabei, dass FDW in niedrigen Niveaus auf reproduktive Aspekte beschränkt bleibt, sich in hohen Niveaus aber hin zu evaluierenden und kreativen Aspekten erweitert (Zeller et al., 2024). Um auch nicht-hierarchische Strukturen im Kontext dieser Beobachtungen beschreiben zu können, wurde in einer der hier vorgestellten Analyse vorangegangenen Untersuchung ein Cluster-Modell des FDW erstellt, welches auch distinkte nicht-hierarchische Strukturen aufdecken (Zeller & Riese, 2025) . Dabei konnten die resultierenden prototypischen Personengruppen aber aufgrund methodischer Limitationen nicht als tatsächlich „latente" Strukturen verstanden werden (Abschnitt 6.2.2).

Im vorliegenden Artikel wird nun ein erweiterter Ansatz vorgestellt, der aufbauend auf den bisherigen Erkenntnissen zur inneren Struktur des FDW und mithilfe einer erweiterten Methodik tatsächlich latente nicht-hierarchische Strukturen erfasst. Dazu wird orientiert an der Computational Grounded Theory (CGT) nach Nelson (2020) zunächst eine explorative Analyse unter intensiver Nutzung von Theorie- und Expertenwissen durchgeführt. Für diesen Zweck wird eine Latente Profilanalyse (LPA, z. B. Spurk et al., 2020) eines Datensatzes mit 846 Bearbeitungen eines größtenteils offenen FDW-Testinstruments (Gramzow et al., 2013)

aus dem Projekt ProfiLe-P+ (Vogelsang et al., 2019) durchgeführt. Anschließend werden die gefundenen latenten Profile im Rahmen der sog. „Pattern Confirmation“ der CGT (siehe Abschnitt 6.2.3) in ihrer Gültigkeit bestätigt. Zu diesem Zweck wird der CGT folgend die Performanz eines automatisierten Systems auf der Basis von Machine-Learning-(ML)-Modellen zur Zuordnung von Proband:innen zu den latenten Profilen evaluiert.

Das in dieser Untersuchung verwendete Testinstrument besteht zu einem großen Anteil aus Aufgaben in offenem Antwortformat (Gramzow et al., 2013); auch ähnliche Testinstrumente weisen häufig mindestens anteilig ein offenes Aufgabenformat auf (z. B. Kröger, 2019). Die Auswertung zu Forschungs- und Feedbackzwecken erzeugt somit bislang meist einen hohen händischen Kodieraufwand. Das in diesem Beitrag im Rahmen der CGT-Pattern Confirmation verwendete automatisierte System basiert unter anderem auf einem ML-Modell zum automatisierten Scoren des Testinstruments. Als Nebenprodukt der Bestätigung der untersuchten latenten FDW-Strukturen entstand somit ein allgemeines automatisiertes Assessment-System für das FDW im Fach Physik. Abgesehen von einer zeitökonomischen Messung des FDW zu Forschungszwecken ermöglicht dieses System damit beispielsweise auch ein automatisiertes Feedback zum Stand des FDW zum Zweck der formativen Diagnostik. Der vorgestellte Workflow bietet also insgesamt das Potenzial, als Blaupause für den Transfer bisheriger Forschungsergebnisse und Messverfahren mit offenen Testitems in die Lehrpraxis zu dienen.

6.2. Theoretischer Hintergrund

6.2.1 Konzeptualisierung des Fachdidaktischen Wissens

Es existieren unterschiedliche Ansätze, FDW zu konzeptualisieren. Gemein ist den Meisten die auf (Shulman, 1986, 1987) zurückgehende grundlegende Auffassung von FDW als demjenigen Wissen, das zur Vermittlung von bestimmtem Fachwissen an bestimmte Lernende notwendig ist. Im englischsprachigen Raum hat sich parallel zu FDW dabei das Konstrukt des *Pedagogical Content Knowledge* (PCK) entwickelt, das eng verwandt, aber nicht deckungsgleich mit FDW ist (z. B. Vollmer & Klette, 2023). FDW lässt sich aber im Rahmen des international etablierten „Refined Consensus Model“ (RCM) des PCK (Carlson et al., 2019; Hume et al., 2019) interpretieren. Grob zusammengefasst konzeptualisiert das RCM das Konstrukt PCK in den drei Domänen *collective* PCK (cPCK), *personal* PCK (pPCK) und *enacted* PCK (ePCK). Dabei beschreibt cPCK die kollektive, explizierbare Wissensbasis der fachdidaktischen Community („Lehrbuchwissen“), pPCK das persönliche internalisierte (aber immer noch explizierbare) Wissen der Einzelpersonen („testbarer Wissensstand“) und ePCK das individuelle, ggf. implizite Wissen einer Lehrkraft, das in einer konkreten Situation der Planung, Durchführung und Reflexion von Unterricht zugrunde liegt („aus der Handlung rekonstruierbar“; Carlson et al., 2019, S. 83–90). Im Sinne des RCM werden Operationalisierungen und die durch die Testinstrumente abgebildeten FDW-Konstrukte dabei meist *als pPCK* interpretiert (z. B. Kulgemeyer et al., 2023; Schiering et al., 2023). Auch dieser Beitrag schließt sich dieser Auffassung an. Im Folgenden wird somit der Begriff des FDW im Sinne eines pPCK genutzt. Neben dem RCM wird häufig auch das sog. Kontinuumsmodell

(„Model of Competence“, MoC) nach Blömeke et al. (2015) zur Konzeptualisierung des FDW verwendet. Das MoC beschreibt professionelle Kompetenz als ein Kontinuum zwischen kognitiven Dispositionen und gezeigter Performanz in konkreten Handlungssituationen. FDW lässt sich im Rahmen dieses Modells eher auf der Seite der kognitiven Dispositionen verorten (z. B. Kulgemeyer et al., 2023).

Zur Operationalisierungen des FDW für die Naturwissenschaften und im Fach Physik (z. B. Gramzow et al., 2013; Kröger, 2019; Tepner et al., 2012) werden meist Strukturmodelle auf Basis der folgenden drei Dimensionen genutzt:

1. *Fachinhalte* (z. B. Mechanik, Elektrizitätslehre etc.): Die grundsätzliche Abhängigkeit des FDW vom jeweiligen zu vermittelnden Fachinhalt wird in allen gängigen Konzeptualisierungen des FDW angenommen (Baumert & Kunter, 2006; Riese et al., 2017; Shulman, 1986, 1987; Sorge et al., 2019).
2. *Facetten* bzw. *fachdidaktische Inhalte*: Diese zentrale Dimension dient zur Beschreibung unterschiedlicher inhaltlicher Themenfelder, die FDW umfasst. Die Auswahl relevanter Facetten des FDW wird dabei zumeist auf Basis von theoretisch-normativen Modellierungen (z. B. Magnusson et al., 1999; Park & Oliver, 2008), Analysen von Curricula der Lehrerbildung bzw. Literatur-Reviews (z. B. Gramzow et al., 2013; Kröger, 2019) und Expertenbefragungen zur Einschätzung der curricularen Validität entsprechender Testaufgaben (z. B. Gramzow et al., 2013) getroffen. Dabei werden in verschiedenen Studien häufig unterschiedliche Facetten fokussiert (Übersicht bei Kirschner, 2013). In den meisten Ansätzen, werden aber die zentralen Facetten *Schülervorstellungen* und *Instruktionsstrategien*, die bereits bei Shulman (1987) zu finden sind, in die Betrachtung eingeschlossen.
3. *Kognitive Aktivität* bzw. *Wissensarten*: Üblicherweise wird in den Modellen zur Entwicklung von FDW-Testinstrumenten, der Empfehlung von Klieme et al. (2003) folgend, eine Dimension zur Anreicherung entsprechender Testaufgaben mit Anforderungen unterschiedlicher kognitiver Komplexität genutzt. In den im deutschsprachigen Raum etablierten FDW-Testinstrumenten werden dabei beispielsweise sog. kognitive Aktivitäten (z. B. Blömeke et al., 2008b; Riese et al., 2017) wie beispielsweise *Reproduzieren, Anwenden* und *Analysieren* oder auch sog. Wissensarten (Kröger, 2019; Tepner et al., 2012) wie *deklaratives* oder *prozedurales Wissen* genutzt.

Mithilfe der beschriebenen Konzeptualisierungen und Operationalisierungen wurde das FDW in unterschiedlichen Studien systematisch erhoben. Dabei zeigten sich unter anderem (a) Zuwächse des FDW in Studium und Vorbereitungsdienst, (b) Unterschiede im FDW zwischen verschiedenen Lehramtstypen (z. B. Großschedl et al., 2015; Riese & Reinhold, 2012), (c) Zusammenhänge des FDW mit FW und PW (z. B. Sorge et al., 2019) sowie (d) Zusammenhänge des FDW mit gezeigter Performanz in konkreten Handlungssituationen (z. B. Förtsch et al., 2016; Kulgemeyer et al., 2020; Schröder et al., 2020). Das FDW hat sich somit insgesamt als bedeutsam erwiesen – sowohl für das Professionswissen und dessen Entwicklung als Ganzes, als auch für die Handlungsqualität von Lehrpersonen bzw. die Unterrichtsqualität im naturwissenschaftlichen Unterricht.

6.2.2 Empirische Analyse der inneren Struktur des Fachdidaktischen Wissens

Neben den beschriebenen Analysen zur Bedeutung des FDW für Professionswissen und für die Handlungsqualität, für die eher die Gesamteinschätzung des FDW der Probanden (z. B. in Form von Summenscores) relevant war, wurden auch weiterführende Studien zur empirisch fundierten, kriterienorientierten Beschreibung der inneren Struktur des FDW durchgeführt. Solche inhaltlichen Beschreibungen bieten einen Ansatzpunkt für die Weiterentwicklung der Konzeptualisierung des FDWs als Teil des Professionswissens von Lehrkräften. Darüber hinaus können diese Erkenntnisse für ein differenziertes Assessment des FDW auf Subskalenebene, als Ausgangsbasis für inhaltsbezogenes Feedback sowie für die inhaltliche Charakterisierung individueller Stärken und Schwächen von Einzelpersonen genutzt werden.

Nach dem Vorbild großer Schulleistungsstudien wie PISA oder TIMSS wurden bereits in unterschiedlichen Studien empirisch basierte inhaltliche Beschreibungen von Niveaustufen des FW (z. B. Bernholt, 2010; Woitkowski & Riese, 2017) und des PW (z. B. König, 2009) auf Basis von Item-Response-Modellen erstellt. Daran angelehnt analysierten Schiering et al. (2019, 2023) sowie Zeller et al. (2022) Niveaustufen des FDW mithilfe des Scale Anchoring Verfahrens (z. B. Mullis et al., 2016). Die im Rahmen solcher unabhängigen Analysen gefundenen Parallelen konnten in einer projektübergreifenden Betrachtung bestätigt werden (Zeller et al., 2024): In niedrigen Niveaustufen beschränkt sich das FDW primär auf reproduktive Aspekte, während in höheren Niveaustufen kreative und evaluierende Elemente hinzukommen. Die Ergebnisse deuten zudem darauf hin, dass eine genaue Betrachtung der Dimensionen wie der kognitiven Aktivierung (s. o.) unter Einbeziehung zusätzlicher Anforderungsbereiche, wie den Stufen der Taxonomie nach Anderson und Krathwohl (2001), sinnvoll und für die genauere Untersuchung der Feinstruktur des FDW vielleicht sogar notwendig ist (Zeller & Riese, 2024). Die Methodik dieser auf Item-Response-Modellen basierenden Analysen ist allerdings auf hierarchische Betrachtungen beschränkt, sodass so beispielsweise keine Unterteilung von typischen Proband:innen mit Stärken im Kreieren oder Evaluieren von Unterrichtselementen möglich war.

Um die Limitation der ausschließlich hierarchischen Beschreibungen im Rahmen von Niveaumodellen zu überwinden, führten Zeller und Riese (angenommen) eine nicht-hierarchische Cluster-Analyse des FDW mithilfe des K-Means-Algorithmus (MacQueen, 1967) unter Betrachtung der kognitiven Anforderungsbereiche *Reproduzieren*, *Analysieren*, *Anwenden*, *Evaluieren* und *Kreieren* (angelehnt an Anderson & Krathwohl, 2001) durch. Das Cluster-Modell auf Basis der Scores sowie eine darauf aufbauende computerbasierte Sprachanalyse der Antworten der Proband:innen zu den offenen Aufgaben des zugrundeliegenden Testinstruments deuteten auf die Trennbarkeit dieser fünf kognitiven Anforderungen als Teilkompetenzen des FDW hin und zeigten die Existenz von Personengruppen mit prototypischem Antwortverhalten und prototypischen Kompetenz-ausprägungen im Rahmen dieser Teildimensionen. Insgesamt zeigte sich sowohl in der inhaltlichen Re-Analyse des Testinstruments zur Zuordnung der Aufgaben zu den kognitiven Anforderungsbereichen als auch in den Personen-Clustern, dass Kompetenzen im Evaluieren häufig mit Kompetenzen im Analysieren und Kompetenzen im Kreieren häufig mit

Kompetenzen im Anwenden einhergehen.

Die Analyse von Zeller und Riese (angenommen) ist allerdings durch die teilweise niedrige Anzahl an Aufgaben in den oben genannten kognitiven Anforderungskategorien limitiert. Diese Einschränkung des Testinstruments führte dazu, dass Modelle zur Untersuchung echt *latenter* Strukturen, wie beispielsweise die im Rahmen von LPAs typischerweise genutzten Gaussian-Mixture-Models (GMM, z. B. Spurk et al., 2020), nicht konvergierten und eine K-Means-Analyse genutzt werden musste. Die erhaltenen Cluster sind somit eher als „datengestützte Leistungsquantile" in den kognitiven Anforderungen aufzufassen, denn als latente Gruppen.

Die bestehenden Ansätze zur empirisch gestützten inhaltlichen Beschreibung der inneren Struktur des FDW sind also bisher limitiert. Im vorliegenden Beitrag werden daher zusätzliche vorbereitende Schritte bei der Datenverarbeitung verwendet, um eine erweiterte Analyse nicht-hierarchischer Strukturen mithilfe einer LPA (z. B. Spurk et al., 2020) unter der Nutzung von GMMs (z. B. Murphy, 2022) des FDW-Datensatzes aus dem Projekt ProfiLe-P+ (Vogelsang et al., 2019) durchzuführen. Vergleichbare explorative Analysen werden zur Absicherung ihrer Aussagekraft häufig unter Nutzung der CGT nach (Nelson, 2020) strukturiert (z. B. Tschisgale et al., 2023), die daher im folgenden Abschnitt vorgestellt wird.

6.2.3 Machine-Learning-basierte Analysen im Rahmen der Computational Grounded Theory

Explorative ML-basierte Analysen bergen zwar das Potenzial, bislang unerkannte Strukturen in den jeweils untersuchten Konstrukten aufzudecken, es stellt aber eine Herausforderung dar, die Interpretierbarkeit der Ergebnisse zu gewährleisten (z. B. Sherin, 2013; Zhai et al., 2020b). In ihrem systematischen Review von ML-Anwendungen in der naturwissenschaftsdidaktischen Forschung stellen Zhai et al. (2020b) dementsprechend fest, dass ein Großteil der Forschungsprojekte bislang primär auf die Entlastung von menschlichen Ratern bei basalen Aufgaben des Assessments ausgerichtet ist. Gleichzeitig arbeiten aber Zhai et al. (2020a) das Potenzial von explorativen Methoden, die im ML-Kontext auch als *Unsupervised-Learning-Methoden* bezeichnet werden, zur Untersuchung bisher unerkannter Strukturen heraus. Auch Kubsch et al. (2022) unterstreichen diese Potenziale im Rahmen der Entwicklung ihres Frameworks zu Einordnung von ML-basierten Analysen. Um den methodischen Workflow des hier vorgestellten Projekts und die einzelnen Schritte der zur Strukturierung der Analyse herangezogenen CGT darzustellen, werden im Folgenden einige Begriffe aus dem ML-Kontext eingeführt bzw. in den ML-Kontext eingeordnet.

Im ML-Bereich wird zwischen sog. dem *Supervised Learning* und *Unsupervised Learning* unterschieden (z. B. Géron, 2019). Im Supervised Learning geht es um die automatisierte Vorhersage bestimmter Ziel-Variablen (auch *Targets* oder *Labels*) mithilfe unabhängiger Variablen (sog. *Features*). Das kann beispielsweise ein Regressionsmodell zur Vorhersage von Studienerfolg (Target) auf Basis von Prädiktoren wie der Abiturnote und dem IQ (Features) sein. Die Erstellung eines solchen Modells auf Basis eines vorhandenen Datensatzes wird auch als *Training* bezeichnet. Dazu wird meist eine sog. *Loss-Funktion* (auch kurz *Loss*) optimiert, die von den Parametern des Modells abhängig ist. Im Falle eines (hier beispielhaft

zweidimensionalen) Regressionsmodells ist der Loss üblicherweise die Mean-Squared-Error-Funktion (MSE), die von den Regressionsgewichten des Modells (w_1, w_2 und b), sowie den Trainingsdaten (Targets $y^{(i)}$ und Features $x_1^{(i)}, x_2^{(i)}$, $i = 1 \dots N$) abhängt:

$$\mathrm{MSE}(w_1, w_2, b) = \sum_{i=1}^{N} \left(y^{(i)} - w_1 x_1^{(i)} - w_2 x_2^{(i)} - b \right)^2 .$$

Zur Anpassung des Modells an die Daten werden die Modellparameter mithilfe mathematischer Verfahren so optimiert, dass die Loss Funktion minimiert wird (z. B. Géron, 2019). Für Klassifikationsmodelle, bei denen die Targets diskrete Kategorien anstelle von kontinuierlichen Größen sind, existieren andere Loss-Funktionen (z. B. die sog. Cross-Entropy), die hier aus Platzgründen nicht ausführlicher dargestellt werden. Das grundsätzliche Vorgehen beim Training bleibt aber gleich.

Bei großen Datensätzen und hochdimensionalen Feature-Variablen (beispielsweise Sprachdaten) können in einem Optimierungsschritt aufgrund von Limitationen der Rechenkapazität meist nicht alle verfügbaren Datenpunkte auf einmal verwendet werden. Man geht dann dazu über, den Gesamtdatensatz in kleinere Einheiten, sog. *Batches* aufzuteilen und in einem einzelnen Optimierungsschritt jeweils nur einen einzelnen Batch zu nutzen. So wird iterativ der Datensatz durchlaufen, wobei man, wenn einmal der gesamte Trainingsdatensatz durchlaufen worden ist, auch von einer *Epoch* an Training spricht. Zur Einschätzung der Vorhersagegenauigkeit (auch *Performanz*) des Modells wird meist nicht direkt der Loss, sondern andere, leichter interpretierbare Metriken, wie beispielsweise die Varianzaufklärung R^2 bei Regressionsmodellen oder die prozentuale Übereinstimmung (*Accuracy*) und Cohens κ zwischen Labeln und Vorhersagen bei Klassifikationsmodellen genutzt.

ML-Modelle mit einer hohen Anzahl an Parametern können Spezifika des Trainingsdatensatzes sehr genau abbilden[51], man spricht auch von *Overfitting* (z. B. Géron, 2019). Eine hohe Performanz des Modells für die Trainingsdaten gewährleistet daher noch keine hohe Performanz für Daten, die während des Trainings nicht genutzt wurden. Für die tatsächliche Nutzung eines Modells ist aber gerade die Performanz für solche „ungesehenen" Daten von Interesse (z. B. Breiman, 2001). Um anschließend an das Training das Modell zu evaluieren, wird daher ein separater Evaluierungs- oder Test-Datensatz verwendet, der vom Training ausgeschlossen ist. Dieses Vorgehen lässt sich auch zur sog. *k-Fold-Cross-Validierung* (CV) erweitern, bei der der verfügbare Gesamtdatensatz in k gleich große Segmente unterteilt wird. Das Modell wird dann k-mal neu trainiert, wobei jeweils eines der Segmente zu Evaluierungszwecken zurückgehalten wird. Die Evaluierung erfolgt dann im Anschluss auf Basis der Modellvorhersagen für die jeweiligen Evaluierungsdaten aus dem wiederholten Training.

Anders als beim Supervised Learning liegen beim Unsupervised Learning keine a priori bekannten Targets vor, sondern es geht um die Untersuchung von Mustern und Strukturen in Daten (z. B. Duda et al., 2001). Dazu können unterschiedliche Modelle verwendet werden,

[51] Man stelle sich beispielsweise ein Polynom hohen Grades vor, welches an vergleichsweise wenige Datenpunkte angepasst wird.

deren Trainingsalgorithmen sich mitunter unterscheiden. Als Beispiel dient hier das GMM, bei dem davon ausgegangen wird, dass folgender (hier vereinfacht dargestellter) Prozess die Daten generiert, d. h., dass die Daten der sich dadurch ergebenden Verteilung folgen (z. B. Murphy, 2022):

1. Für jeden Datenpunkt wird eines von K Clustern gewählt. Die Wahrscheinlichkeit für jedes einzelne Cluster ist durch einen entsprechenden Parameter gegeben[52].
2. Der tatsächliche Datenpunkt ergibt sich aus einer von K Normalverteilungen, deren jeweilige Mittelwert- und Kovarianz-Parameter sich von Cluster zu Cluster unterscheiden können.

Mithilfe eines Algorithmus, der hier aus Platzgründen nicht näher erläutert werden kann (z. B. Murphy, 2022) können aus den Daten diejenigen Cluster-Zuordnungen und Parameter der K Normalverteilungen ermittelt werden, die die beobachteten Daten am wahrscheinlichsten beschreiben. Zur Evaluierung der Passung solcher Modelle zu den Daten existieren unterschiedliche Metriken, wobei häufig das sog. *Bayesian Information Criterion* (BIC) verwendet wird. Ein höherer BIC-Score bedeutet eine höhere Wahrscheinlichkeit, dass das Modell die Daten adäquat beschreibt.

Eine Schwierigkeit bei der Anwendung von explorativen Unsupervised-Learning-Methoden stellt insbesondere die mitunter hohe Dimensionalität der verwendeten Daten dar (z. B. Géron, 2019; Sherin, 2013). Beispielsweise wird eine Cluster Analyse eines Score-Datensatzes zu einem Testinstrument mit über 20 Aufgaben nur wenig interpretierbare Cluster liefern, die stark von einzelnen Aufgaben abhängen. Sherin (2013) rät daher, zur Erhöhung der Interpretierbarkeit und somit Nutzens solcher Methoden, die Computer-basierten algorithmischen Auswertungsschritte bereits im Analyseprozess mit menschlichem Expertenwissen und menschlicher Interpretationskraft zu verknüpfen. Um eine solche Verknüpfung zu strukturieren, schlägt Nelson (2020) die CGT vor. Im Wesentlichen besteht ihr Ansatz aus drei Schritten:

- *Pattern Detection*: Es werden explorative Methoden zur Untersuchung potenziell bisher unerkannter Strukturen in den Daten angewendet.
- *Pattern Refinement*: Die identifizierten Strukturen werden durch inhaltliche Detailanalysen und / oder Einbeziehung von menschlichem Expertenwissen ausgeschärft.
- *Pattern Confirmation*: Die Performanz von ML-Modellen zur Vorhersage der identifizierten Strukturen wird evaluiert. Dies dient zur Bestätigung der beobachteten Strukturen hinsichtlich ihrer Reliabilität und Validität.

Die ursprünglich publizierte Beschreibung der CGT ist stark auf Text-Daten ausgerichtet, es wird aber betont, dass für die jeweilige Analyse und konkreten Daten insbesondere Pattern Detection und Pattern Refinement eher als Teile eines iterativen Prozesses betrachtet werden

[52] Die Zuordnung zu den Kategorien folgt also einer verallgemeinerten Bernoulli Verteilung (auch „Categorical Distribution“, Murphy, 2022).

können bzw. müssen (Nelson, 2020). Konkret in der Physikdidaktik nutzten Tschisgale et al. (2023) die CGT erfolgreich zur explorativen Analyse von Problemlösestrategien von Schülerinnen und Schülern auf Basis von Text-Daten. Sie stellen dabei die Potenziale der CGT und die Vorteile dar, die eine ML-basierte Analyse gegenüber einer manuellen explorativen Analyse haben kann.

6.3. Ziele und Forschungsfragen

Für die Weiterentwicklung der Konzeptualisierung des FDWs als Teil des Professionswissens von Lehrkräften ist eine detailliertere inhaltliche Beschreibung der inneren Struktur des FDW und eine differenziertere Analyse von Zusammenhangsstrukturen innerhalb des Professionswissens von Lehrkräften notwendig. Darüber hinaus können empirisch abgesicherte Beschreibungen solcher Strukturen als Basis für Feedback genutzt werden, das über eine reine quantitative Gesamteinschätzung hinaus geht. Die bestehenden Ansätze für eine solche Untersuchung der inneren Struktur des FDW sind aber bisher limitiert. Mithilfe von Niveauanalysen (z. B. Schiering et al., 2023; Zeller et al., 2024) konnten zwar projektübergreifend Kompetenzniveaus ermittelt werden, diese sind allerdings methodisch auf hierarchische Beschreibungen beschränkt. Darüber hinaus liefern die Niveauanalysen primär entweder projektspezifische, meist wenig generalisierbare oder nur recht grobe Beschreibungen. Insbesondere können sie nicht zwischen Lernenden mit Stärken bzw. Schwächen bei bestimmten kognitiven Anforderungen wie dem Evaluieren und Kreieren unterscheiden. Der bislang genutzte nicht-hierarchische Ansatz auf Basis von K-Means-Cluster-Analysen (Zeller & Riese, 2025) kann eine solche Unterscheidung vornehmen. Hier wird die Aussagekraft der Ergebnisse allerdings dadurch eingeschränkt, dass das bislang genutzte bzw. aufgrund von Limitationen des Datensatzes einzig nutzbare Cluster-Modell keine tatsächlich latenten Strukturen beschreiben kann. Bei den im Rahmen dieser Analyse beschriebenen Clustern handelt es sich also eher um „datenbasierte Leistungsquantile".

Die vorgestellte Studie verfolgt dementsprechend das Ziel, nicht-hierarchische, *tatsächlich latente* Strukturen des FDW in Form von Personengruppen mit typischen Ausprägungen des FDW zu beschreiben. Dazu wird aufbauend auf einem Datensatz von 846 Bearbeitungen des größtenteils offenen FDW-Testinstruments aus dem ProfiLe-P+ - Projekt (Vogelsang et al., 2019; siehe auch Abschnitt 6.4.1) eine LPA auf Basis von GMMs durchgeführt. Zunächst werden die Daten dafür auf Basis von bisherigen Modellierungen und Ergebnissen sowie aus methodischen Gründen (siehe Abschnitt 6.4.2) zu Summenscores in den kognitiven Anforderungskategorien *Reproduzieren*, *Anwenden-Kreieren* und *Analysieren-Evaluieren* akkumuliert. Dies kann im Rahmen der CGT als Element eines „vorgezogenen" Pattern Refinements verstanden werden. Verortet man die hier vorgestellte Analyse im Gesamtprojekt, so zeigt sich ein zyklisch-iteratives Vorgehen in Fortführung der Vorläuferanalyse (Zeller & Riese, 2025), das ebenfalls im Rahmen der CGT begründet werden kann. Die (bisher hypothetischen) Cluster dieser LPA werden *latente Kompetenzprofile* genannt. Die erste Forschungsfrage widmet sich im Sinne der CGT der „Detection" dieser Cluster:

> ***FF1** (~ Pattern Detection)*: Welche latenten FDW-Kompetenzprofile lassen sich durch eine GMM-basierte LPA in den FDW-Score-Daten des Projekts

> X bezüglich der kognitiven Anforderungskategorien *Reproduzieren*, *Anwenden-Kreieren* und *Analysieren-Evaluieren* finden*?*

Die Kompetenzprofile werden dabei durch ihre durchschnittlichen Summenscores bezüglich der betrachteten Anforderungskategorien interpretiert.

Um die Kompetenzprofile im Sinne der CGT zu bestätigen, soll die Performanz von ML-Modellen bei ihrer automatisierten Erfassung evaluiert werden. Einerseits könnten zu diesem Zweck wie bei Tschisgale et al. (2023) diejenigen Daten verwendet werden, welche auch bei der Erstellung des Cluster-Modells genutzt wurden[53]. Im Falle des vorliegenden Beitrags wären das die Scores. Ein stärkeres Argument für die Robustheit der Zuordnung von Proband:innen zu den Kompetenzprofilen als Pattern Confirmation wäre aber eine automatisierte Zuordnung direkt auf Basis der authentischen Test-Bearbeitungen der Proband:innen. Diese bestehen zu größtenteils aus den Text-Antworten auf die offenen Aufgaben des Testinstruments (siehe Abschnitt 6.4.1). Explorative Analysen zur Zuordnung von Proband:innen zu den Kompetenzprofilen haben gezeigt, dass ein dafür nutzbares ML-System eine deutlich höhere Performanz erreicht, wenn zunächst die automatisierte Bepunktung der offenen Aufgaben des Testinstruments und erst darauf aufbauend die Zuordnung zu den Kompetenzprofilen in den Blick genommen wird. Eine automatisierte Bepunktung bzw. Klassifikation von Textelementen wurde in der Naturwissenschaftsdidaktik bereits mehrfach durch sog. *Finetuning* von BERT-Sprachmodellen erfolgreich vorgenommen (Camus & Filighera, 2020; Wulff et al., 2021; Zhai et al., 2020b). Beim Finetuning wird ein vortrainiertes Sprachmodell für eine konkrete Aufgabe nach-trainiert (Details in Abschnitt 6.4.3). Die Pattern Confirmation wird also mithilfe der folgenden Forschungsfragen strukturiert:

> ***FF2a*** *(~ Pattern Confirmation 1)*: Welche Maschine-Mensch-Übereinstimmung erreicht ein BERT-Sprachmodell (Devlin et al., 2019) bei der Vorhersage von FDW-Scores unter Nutzung eines typischen Finetuning-Workflows auf Basis von 846 Bearbeitungen des FDW-Testinstruments?
>
> ***FF2b*** *(~ Pattern Confirmation 2)*: Wie hoch ist die Maschine-Mensch-Übereinstimmung einer automatisierten Zuordnung von Bearbeitungen des FDW-Testinstruments zu einem prototypischen FDW-Kompetenzprofil auf Basis der maschinellen Score-Vorhersagen (FF2a)?

Details zur Auswahl des Sprachmodells und dessen Training bzw. Finetuning (FF2a) werden in Abschnitt 6.4.3 erläutert. Durch die Trennung des automatisierten Scorings von der Cluster-Vorhersage in zwei separate Modelle können detailliertere Informationen aus dem Datensatz in Form der Scores in die Entwicklung des Systems aufgenommen werden. Neben der Nutzung dieser Modelle im Rahmen der Pattern Confirmation können sie zudem als Basis eines skalierbaren automatisierten FDW-Assessments dienen.

[53] Bei Tschisgale et al. (2023) wurden Cluster in numerischen Repräsentationen (sog. *Embeddings*) von Sätzen untersucht. Für das Pattern Refinement wurden diese Cluster ausgehend von den Embeddings zugeordnet.

6.4. Methode

6.4.1 Testinstrument und Datensatz

Die im vorliegenden Beitrag verwendeten Daten entstammen dem ProfiLe-P+ - Projekt (Vogelsang et al., 2019) und wurden im Zeitraum von 2016 bis 2019 an 12 deutschsprachigen Universitäten erhoben. Ziel des Projekts war unter anderem die Erfassung des FDWs angehender Physiklehrkräfte. Dazu wurde das im Vorgängerprojekt ProfiLe-P (Riese et al., 2015) von Gramzow (2015) entwickelte FDW-Testinstrument bestehend aus 23 offenen und 4 Multiple-Choice (MC) Aufgaben eingesetzt. Die Entwicklung des Testinstruments fand durch ein intensives Literaturreview gängiger Modellierungen und Operationalisierungen des FDW statt. Zur Validierung der ausgewählten Facetten und auch der Testaufgaben im Allgemeinen wurden sowohl qualitative Untersuchungen wie Think-Aloud-Studien und Expertenbefragungen als auch quantitative Untersuchungen insbesondere auf Basis von Item-Response-Modellen durchgeführt (Gramzow, 2015). Das Testinstrument erfasst physikdidaktisches Wissen in den vier Facetten *Schülervorstellungen*, *Instruktionsstrategien*, *Fachdidaktische Konzepte* (z. B. didaktische Rekonstruktion) sowie *Experimente und Vermittlung eines angemessenen Wissenschaftsverständnisses* (kurz *Experimente*) und den kognitiven Anforderungen *Reproduzieren*, *Anwenden* und *Analysieren* (Abbildung 6.1). Da im Projekt ProfiLe-P(+) Zusammenhänge in einem exemplarisch fokussierten fachphysikalischen Inhaltsbereich differenziert auf Subskalenebene (vgl. Riese et al., 2017) betrachtet wurden, konzentriert sich der verwendete FDW-Tests auf den physikalischen Fachinhalt *Mechanik*. Eine Beispielaufgabe des Testinstruments inklusive einer beispielhaften Antwort aus dem Datensatz zeigt Abbildung 6.2. Der finale Datensatz besteht aus 846 Bearbeitungen dieses Testinstruments durch Physik-Lehramtsstudierende der Sekundarstufe im Bachelor- und Masterstudiengang in Quer- und Längsschnitt, wobei diese Bearbeitungen als unabhängige virtuelle Proband:innen (Davier et al., 2008) betrachtet werden. Demographische Eckdaten sind in Tabelle 6.1 dargestellt.

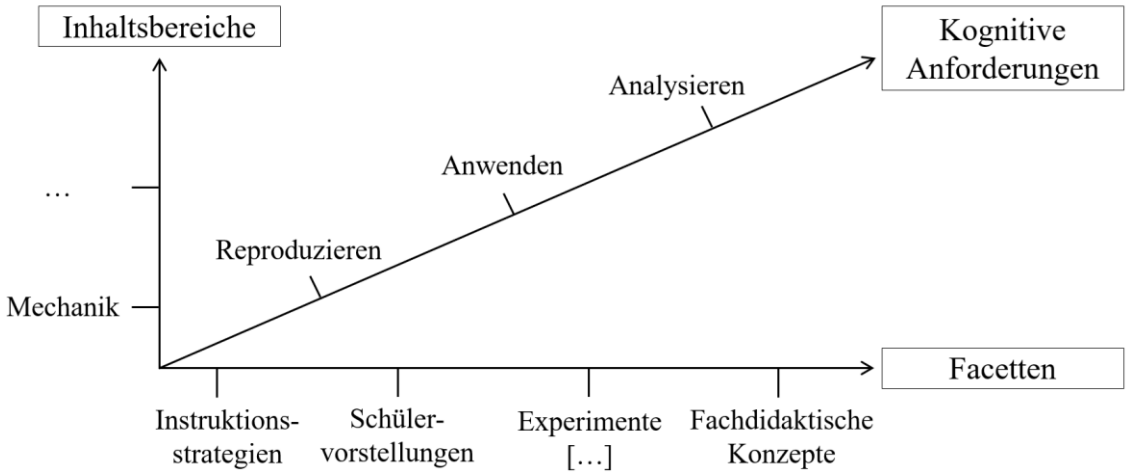
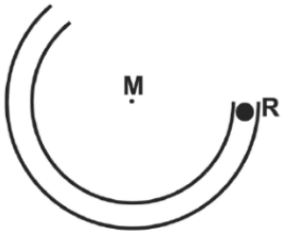


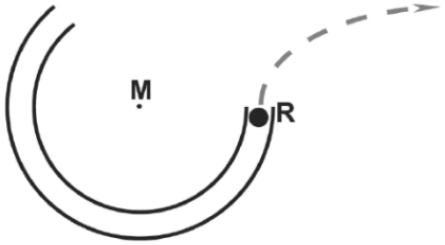
Abbildung 6.1 Itementwicklungsmodell des FDW-Testinstruments des ProfiLe-P(+) - Projekts nach (Gramzow et al., 2013).

Aufgabe 15

Schüler sollen folgende Situation betrachten: Ein Ball rollt in der dargestellten Rinne (Draufsicht) und verlässt diese am Punkt R.



Schüler *A* zeichnet folgende Bahn, die der Ball nach Verlassen der Rinne beschreiben soll:



Lösung von Schüler A

Angenommen, der Schüler versteht die Zeichnung korrekt als Draufsicht: Welche fachlich nicht korrekte Vorstellung des Schülers *A* liegt bei der gezeichneten Bahnkurve zugrunde?

Antwort einer Probandin / eines Probanden:

"Der Schüler denkt, dass die Zentripetalkraft nach außen und nicht zum Kreismittelpunkt wirkt."

Abbildung 6.2 Beispielaufgabe des FDW-Testinstruments mit Vignette und beispielhafte Antwort aus dem Datensatz (nach Gramzow et al., 2013). Diese Aufgabe ist der Facette Schülervorstellungen und der kognitiven Anforderung Analysieren zugeordnet.

Tabelle 6.1 Demographische Eckdaten des Datensatz bezogen auf die Einzelbearbeitungen (virtuelle Probanden).

Gesamtanzahl Bearbeitungen	Fachsemester Physik	Anzahl Bachelor	Anzahl Master	Anteil weiblich
846	4,1 (3,5)	672	174	34 %

Die Text-Antworten wurden mithilfe eines Kodiermanuals (Gramzow, 2015) bepunktet und nachträglich für Computer-basierte Analysen und das Training des automatisierten Auswertungssystems digitalisiert. Die Verwendung von offenen Aufgaben erhöht zwar den Aufwand bei der Auswertung der Testbearbeitungen, ermöglicht aber eine breitere Abbildung kognitiver Anforderungen (vgl. Krüger & Krell, 2020). Die nachträgliche Schließung entsprechender Testinstrumente eröffnet zudem Fragen nach der Authentizität der entstehenden geschlossenen Aufgaben (Kulgemeyer et al., 2023).

Die Berechnung der Metriken und Visualisierungen der Interrater-Übereinstimmungen als „Mensch-Mensch“ Baseline zur Evaluierung der ML-Modelle (FF2a, FF2b) basiert auf einer Doppelkodierung von 267 Bearbeitungen des Testinstruments. Die Interrater-Übereinstimmung inklusive der MC-Aufgaben beträgt für das Testinstrument $\kappa = 0{,}761$ und es weist eine interne Konsistenz von Cronbach's $\alpha = 0{,}80$ bezogen auf die durch die trainierten Kodierer:innen erstellten Scores auf. Es hat somit vergleichbare quantitative Eigenschaften wie vergleichbare FDW-Tests aus anderen Studien (z. B. Krauss et al., 2008; Kröger, 2019).

6.4.2 FF1: Latente Profilanalyse des FDW

Zur Identifikation latenter Kompetenzprofile des FDW wird eine LPA (z. B. Spurk et al., 2020) durchgeführt. Im Rahmen von LPAs werden typischerweise mehrere Modelle einer Modellklasse mit unterschiedlichen Spezifikationen (z. B. Cluster-Anzahl) anhand des BIC verglichen und eines der best-passenden Modelle theoretisch basiert ausgewählt. Die genaue Modellklasse ist dabei nicht festgelegt, gemein ist aber allen für LPAs genutzten Modellklassen, dass sie die Cluster-Zugehörigkeit als latente Variable modellieren (müssen). Für die hier vorliegende Datenstruktur aus FDW-Scores, die bezüglich dreier Subskalen aggregiert sind (s. u.), bietet sich die häufig in LPAs genutzte Modellklasse der GMMs an.

Aus den bislang durchgeführten Studien (Schiering et al., 2023; Zeller & Riese, 2025; Zeller et al., 2024) ist bekannt, dass kognitive Anforderungskategorien dazu geeignet sind, empirisch basiert projektübergreifend anwendbare inhaltliche Beschreibungen des FDW zu generieren. Im vorangegangenen Ansatz zur Beschreibung nicht-hierarchischer Strukturen wurden die Aufgaben des Testinstruments re-analysiert und orientiert an der Taxonomie kognitiver Anforderungen nach Anderson und Krathwohl (2001) den fünf kognitiven Anforderungskategorien *Reproduzieren, Anwenden, Analysieren, Evaluieren* und *Kreieren* zugeordnet (Zeller & Riese, 2025) . Wegen des aus Gründen der Testökonomie begrenzten Umfang des Testinstruments konnten den Kategorien teilweise nur wenige Aufgaben zugeordnet werden, sodass latente GMMs, die kontinuierliche Daten voraussetzen, nicht sinnvoll verwendet werden konnten. Im Rahmen der hier vorgestellten Analyse wurde daher die Beobachtung genutzt, dass Kompetenzen im Anwenden häufig mit Kompetenzen im Kreieren zusammenhängen und das Kompetenzen im Analysieren häufig mit Kompetenzen im Evaluieren zusammenhängen (Zeller & Riese, 2025; Zeller et al., 2024). In der hier vorgestellten Analyse wurden die fünf Anforderungskategorien daher zu den drei Kategorien *Reproduzieren*, *Anwenden-Kreieren* und *Analysieren-Evaluieren* zusammengefasst. Zusätzlich zu den bereits genannten Argumenten für dieses Vorgehen ermöglicht es die spätere unmittelbare Interpretation der Cluster. Die Re-Analyse zur Zuordnung der Aufgaben zu den Anforderungskategorien wurde von drei Expert:innen durchgeführt. Die Übereinstimmungswerte dieser drei Personen bei der Zuordnung der Testaufgaben zu diesen zusammengefassten Kategorien sind in Tabelle 6.2 dargestellt. Auf Basis dieser Zuordnungen als Diskussionsgrundlage wurde von den drei Expert:inenn gemeinsam eine Konsens-Zuordnung erstellt. Die Anzahl an Aufgaben und erreichbaren Punkten pro Anforderungskategorie bezogen auf diese Konsens-Zuordnung sind in Tabelle 6.3 dargestellt.

Tabelle 6.2 Übereinstimmungsmaße (Cohens κ) der Zuordnung der Testaufgaben zu den kognitiven Anforderungskategorien der drei Expert:innen

Übereinstimmung	Reproduzieren	Anwenden-Kreieren	Analysieren-Evaluieren
κ_{12}	0,84	0,74	0,92
κ_{13}	0,83	0,67	0,55
κ_{23}	0,83	0,76	0,62

Tabelle 6.3 Aufgaben- und Punkteanzahl in der Konsens-Zuordnung der Testaufgaben zu den kognitiven Anforderungskategorien.
Die Zuordnungen sind nicht vollständig disjunkt.

	Reproduzieren	Anwenden-Kreieren	Analysieren-Evaluieren
Aufgabenanzahl	12	7	11
Erreichbare Punktzahl	23	12	14

Um den Datensatz nicht durch unzureichende Bearbeitungen zu verzerren, wurden Personen, die weniger als 50 % der Aufgaben bearbeitet haben, von der LPA zu FF1 ausgeschlossen. Es blieben dadurch 785 Bearbeitungen für die LPA. Die Score-Daten wurden zunächst im Rahmen dieser Anforderungskategorien akkumuliert. Die akkumulierten Scores in diesen Subskalen wurden anschließend auf das Intervall [0, 1] normiert, um die Konvergenz der GMMs zu erleichtern. Im Rahmen der LPA wurden insgesamt 40 GMMs mit einer Cluster-Anzahl von 1 bis 10 an die Daten angepasst. Pro Cluster-Anzahl wurden dabei die folgenden vier Konfigurationen der Kovarianzmatrizen der jeweiligen Normalverteilungen der GMMs modelliert[54]:

- *„Spherical“*: Es gibt keine Kovarianzen zwischen den Skalen (d. h. Anforderungskategorien) und die Varianz ist in allen Skalen ist gleich.
- *„Diagonal“*: Es gibt keine Kovarianzen zwischen den Skalen, die Varianz in den Skalen kann sich aber unterscheiden.
- *„Tied“*: Es gibt Kovarianzen zwischen den Skalen, diese Kovarianzen sind aber für alle Cluster gleich.
- *„Full“*: Es gibt Kovarianzen zwischen den Skalen, die sich von Cluster zu Cluster unterscheiden können.

Für diese Modellierungen wurden die sich ergebenden BIC-Scores für die Auswahl eines

[54] Für diese Analyse wurde das Python Paket scikit-learn (Pedregosa et al., 2011; siehe auch https://scikit-learn.org/stable/modules/mixture.html) verwendet.

geeigneten Modells verglichen. Aus einer theoretischen Perspektive heraus, wäre es zu erwarten, dass durchaus Kovarianzen zwischen den Scores in den Anforderungskategorien bestehen, wobei es keinen Grund gäbe, anzunehmen, dass diese Kovarianzen sich zwischen den Clustern unterscheiden. Dieses Setting würde einer *Tied*-Kovarianzmodellierung entsprechen. Gleichzeitig sind für das Ziel der Beschreibung nicht-hierarchischer Strukturen Clustermodelle ab einer Cluster-Anzahl von vier Clustern besonders interessant, da Modellierungen mit weniger Clustern i. d. R. lediglich hierarchische Abstufungen ergeben.

6.4.3 FF2a: Automatisiertes Scoren des FDW-Testinstruments

Wie in Abschnitt 6.2.3 beschrieben muss zur Analyse der Performanz eines ML-Modells zunächst ein geeigneter Split zwischen Trainings- und Evaluierungsdaten erstellt werden. Da hier insgesamt darauf abgezielt wird, Aussagen über die Kompetenzprofile, d. h. insbesondere Aussagen auf Personen-Ebene und nicht nur auf Aufgaben-Ebene zu treffen, muss auch dieser Split personenweise erfolgen. Da in diesem Projekt ein für ML-Zwecke vergleichsweise kleiner Datensatz vorliegt, werden in den Analysen zu FF2a und FF2b wieder alle 846 Bearbeitungen genutzt, wobei die in der LPA ausgeschlossenen Bearbeitungen nachträglich dem jeweils passendsten Cluster zugeordnet wurden[55]. Darüber hinaus wurde eine 10-Fold-CV (siehe Abschnitt 6.2.3) durchgeführt, die eine Balance zwischen erhöhtem Zeitaufwand für das wiederholte Training des Modells und erhöhter Verlässlichkeit der erhaltenen Performanzschätzungen bietet.

Als Modell für das automatisierte Scoring bietet sich ein Sprachmodell (auch „Language Model“, bzw. LM) an. LMs sind Neuronale Netze (siehe z. B. Géron, 2019) mit einer großen Anzahl an trainierbaren Parametern (einige 10 Mio. bis mehrere 100 Mrd.) zur Verarbeitung von Sprache. Es hat sich gezeigt, dass LMs klassische ML-Modelle in der Performanz bezüglich einer Vielzahl an Sprachverarbeitungsaufgaben inklusive des automatisierten Scorens von offenen Testaufgaben systematisch übertreffen (z. B. Camus & Filighera, 2020). LMs werden mithilfe allgemeiner Sprachverarbeitungsaufgaben, wie beispielsweise dem Vorhersagen des nächsten Wortes bei gegebenen Satzanfängen o. Ä., unter der Nutzung großer Datenmengen „vor“-trainiert (z. B. Hoffmann et al., 2024). Dadurch „erlernen“ LMs eine allgemeine Repräsentation von Sprache, durch die sie sich flexibel an konkrete Anwendungsfälle anpassen können. Das anschließende Training des LMs für einen solchen Anwendungsfall wird auch als *Finetuning* bezeichnet. Im Falle des automatisierten Scorings besteht das Finetuning aus einem klassischen Supervised-Learning. Das Python-Paket „huggingface transformers“ (Wolf et al., 2020) bietet einen großen Umfang an Tools für solches Finetuning und implementiert insbesondere typische Workflows.

Für die vorliegende Studie wurde das sog. BERT-Modell (Devlin et al., 2019) gewählt. Das Modell steht in einer deutschen Variante open-source für huggingface transformers zur Verfügung und wird dort bereitgestellt durch das „Münchener DigitalisierungsZentrum“

[55] Die genutzte Software bietet bei GMMs die Möglichkeit, nachträgliche Clusterzuordnungen für Daten, die nicht während des Cluster-Bildungsprozesses genutzt wurden, vorzunehmen.

(MDZ) der Bayerischen Staatsbibliothek[56]. Neben der leichten Zugänglichkeit dieses BERT-Modells und der hohen Vertrauenswürdigkeit des MDZ als Bezugsquelle gibt es weitere Gründe, die für die Nutzung dieses Modells sprechen. Zunächst wurde das BERT-Modell im deutschsprachigen Raum bereits mehrfach erfolgreich im Rahmen naturwissenschaftsdidaktischer Sprachanalyse für explorative Analysen und auch Klassifkationsprobleme genutzt (z. B. Tschisgale et al., 2023; Wulff et al., 2023). Kürzlich konnten zudem Latif et al. (2024) zeigen, dass das deutsche BERT-Modell geeignet ist, um offene Antworten auf naturwissenschaftliche Fragen der PISA-Studien automatisiert zu scoren[57]. Die Nutzung des BERT-Modells erscheint also auch für den hier vorliegenden Datensatz als vielversprechend.

Das BERT-Modell wurde mit dem Aufgabentext als Input und dem Score als Output nach dem Vorbild von Latif et al. (2024) auf Basis aller Aufgaben gemeinsam trainiert. Es gibt also nur ein gemeinsames finegetunetes BERT-Modell für alle Aufgaben[58]. Insgesamt liegen für das Training 15600 Bearbeitungen einzelner Aufgaben vor. Die Text-Antworten zu den Aufgaben umfassen im Mittel ca. 17 Worte ($\text{Min} = 1$, $\text{Max} = 99$). In Abbildung 6.3 ist ein Histogramm der Antwortlängen dargestellt. Um darzustellen, welche zentralen Begriffe in den Antworten typischerweise auftreten, ist in Abbildung 6.4 zudem eine Wortwolke, die die verwendeten Begriffe entsprechend ihrer Häufigkeit skaliert dargestellt[59]. Zentral sind vor allem Begriffe aus dem Bereich des Lehrens und Lernens (z. B. „Vorstellung“ und „Schüler“) sowie Begriffe aus dem im Testinstrument adressierten Fachinhalt Mechanik (z. B. „Geschwindigkeit“ und „Kraft“). Die Verteilung der Score-Labels innerhalb des gesamten Datensatzes ist in Tabelle 6.4 dargestellt.

Tabelle 6.4 Verteilung der Score-Labels im Gesamtdatensatz (nur Textaufgaben).

Score	**Absolute Häufigkeit**	**Relative Häufigkeit** (gerundet)
0	8800	0,56
1	5128	0,33
2	1672	0,11

[56] https://huggingface.co/dbmdz/bert-base-german-uncased

[57] Anders, als der Name „SciEdBERT“ des Modells von Latif et al. (2024) vermuten lassen könnte, ist nicht zu erwarten, dass die Weiternutzung dieses Modells für das hier genutzte FDW-Testinstrument einen Vorteil gegenüber dem „normalen“ BERT-Modell bietet. Das liegt daran, dass die Aufgaben im hier vorliegenden Testinstrument einen deutlich anderen Inhalt abbilden: hier geht es um FDW, bei SciEdBERT um Fachwissen auf (mittlerem) schulischem Niveau. Eine zur Absicherung dieser Vermutung durchgeführte 3-Fold-CV konnte wie erwartet keine Performanzzuwächse durch die Nutzung von SciEdBERT feststellen.

[58] Anders als bei Latif et al. (2024) konnten durch ein zusätzliches aufgabenweises Finetuning, während dem dann 23 unterschiedliche BERT-Modelle generiert wurden, im Rahmen einer explorativen 3-Fold-CV keine Performanzzuwächse erreicht werden.

[59] Bei Erstellen der Wortwolke (Abbildung 6.4) wurden sog. *Stopwords*, d. h. häufig auftretende Worte, ohne große inhaltliche Bedeutung wie „der“, „die“, „das“, „aber“, „wie“ etc. vernachlässigt.

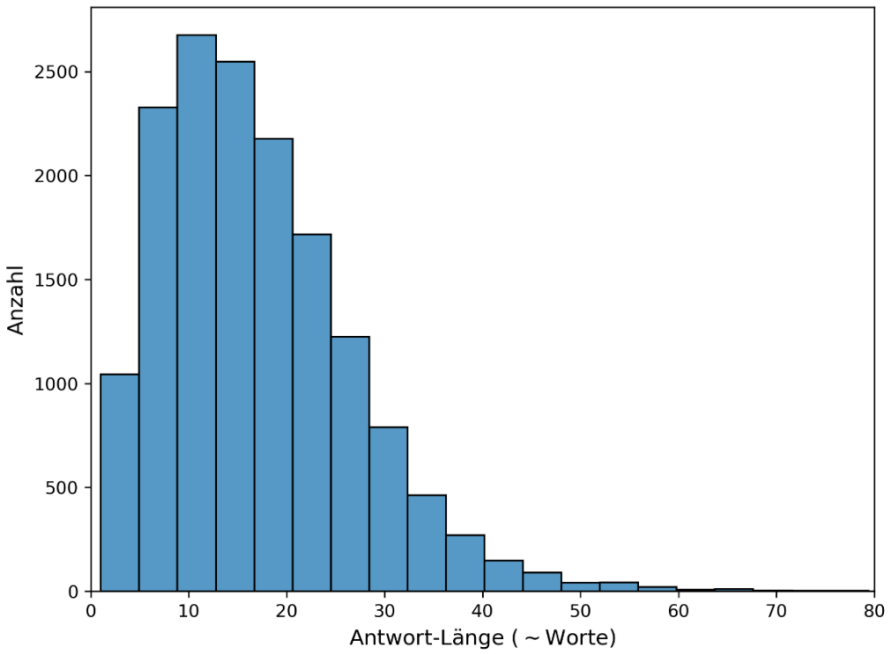


Abbildung 6.3 Histogramm der Längen der einzelnen Antworten zu den offenen Aufgaben des FDW-Testinstruments.

Abbildung 6.4 Wortwolke zur Darstellung zentraler Begriffe in den Antworten zu den offenen Aufgaben des FDW-Testinstruments.

Das Finetuning des BERT-Modells wurde dem Standard-Workflow des huggingface-transformers Python Pakets[60] folgend implementiert. Dabei wurden die Trainingsparameter wie Batch-Größe o. Ä. nicht verändert.

[60] Ein Tutorial, in dem dieser Workflow grob vorgestellt wird, ist unter https://huggingface.co/docs/transformers/tasks/sequence_classification zu finden.

6.4.4 FF2b: Automatisierte Zuordnung zu Kompetenzprofilen

Zur Vorhersage der Kompetenzprofile wurde hier zunächst direkt das GMM aus FF1 verwendet, da so der Vergleich zur Mensch-Mensch Übereinstimmung als Baseline ermöglicht wird. Dieses „wahre" Cluster-Modell liegt vollständig aus den Analysen zu FF1 vor und wurde dementsprechend nicht mehr trainiert. D. h., es wurden die Kompetenzprofile auf Basis der maschinell vorhergesagten Scores direkt mit den Kompetenzprofilen auf Basis der menschlichen Scores verglichen. Zur Bestimmung der Mensch-Mensch Referenzwerte wurden die Scores von Kodierer:in 1 als „wahre Scores" und die Scores von Kodierer:in 2 als „Vorhersagen" betrachtet.

Ergänzend wurde ein logistisches Regressionsmodell (LR-Modell) zur Vorhersage der Kompetenzprofile auf Basis der maschinellen Score-Vorhersagen trainiert, um im Sinne der CGT sicherzugehen, dass auch ein „neues" ML-Modell zur Vorhersage der Kompetenzprofile in der Lage ist. Der Vergleich zu einem analogen Modell für den Mensch-Mensch-Datensatz wäre hier aber nicht zielführend, da der Mensch-Mensch Datensatz (267 Test-Bearbeitungen) deutlich kleiner ist als der für das automatisierte Scoren verfügbare Datensatz (846 Test-Bearbeitungen). Für das Training des LR-Modells konnten die CV-Splits aus dem Training des Scoring-Modells (FF2a) wiederverwendet werden, da sie personenweise erstellt wurden. Zu diesem Zweck wurden während des Trainings des Scoring-Modells neben den Evaluierungs-Vorhersagen auch die Trainings-Vorhersagen zu jedem CV-Split abgespeichert. Diese wurden nun zum Training des LR-Modells genutzt. Auch das LR-Modell wurde dementsprechend um Rahmen der CV 10-mal neu trainiert. Somit ist sichergestellt, dass auch beim Training des LR-Modells keine Vermischung von Trainings- und Evaluierungsdaten stattfindet.

6.5. Ergebnisse

6.5.1 FF1: Latente Kompetenzprofile des FDW

Zur Pattern Detection wurde eine LPA durchgeführt, bei der 40 GMMs an die Daten angepasst und anhand ihres BIC-Scores verglichen wurden (FF1). In Abbildung 6.5 erkennt man, dass die beiden höchsten BIC-Werte für die Konfigurationen „2 Cluster, Kovarianz: Full" ($\mathrm{BIC} = 1491$) und „4 Cluster, Kovarianz: Tied" ($\mathrm{BIC} = 1483$) erreicht werden. Der Unterschied in diesen beiden BIC-Scores ist klein, sodass auf Basis des BIC beide Modelle zur Beschreibung latenter Strukturen herangezogen werden können. Aus theoretischen und pragmatischen Gründen ist es zielführender, das „4 Cluster, Kovarianz: Tied"-Modell weiter zu untersuchen, da eine „Tied"-Kovarianzmodellierung der theoretischen Erwartung am ehesten entspricht und vier Cluster zur Beschreibung nicht-hierarchischer Strukturen geeignet sind (siehe Abschnitt 6.4.2). Die Cluster-Datenpunkte dieses Modells sind in einem Paarplot in Abbildung 6.6 dargestellt. Man erkennt deutlich, dass die Dimensionen Anwenden-Kreieren und Analysieren-Evaluieren für die Zuordnung zu den Clustern zentral sind (Abbildung 6.6, untere mittlere Kachel). Die Cluster-Zentren sind in Abbildung 6.7 inklusive ihrer jeweiligen Mittelwertsstreuung dargestellt.

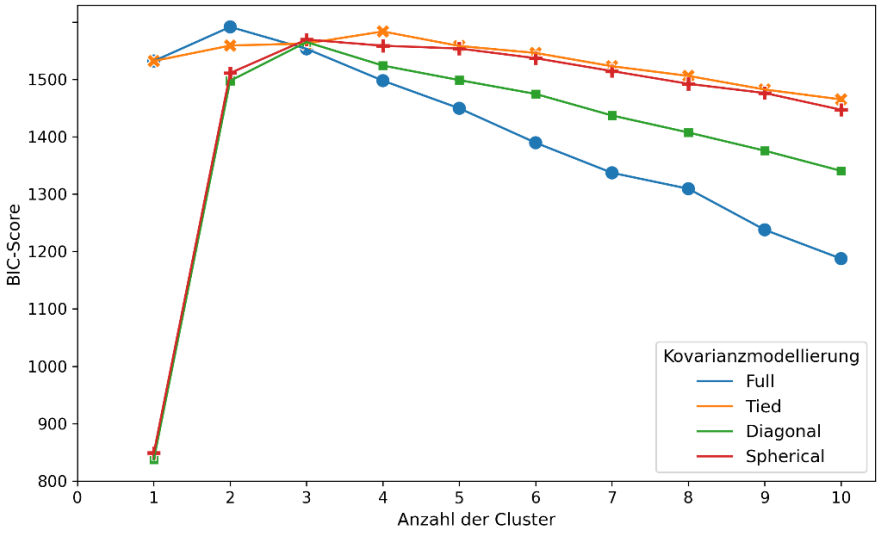


Abbildung 6.5 Darstellung der BIC-Scores für die 40 Gaussian Mixture Models der Latent Profile Analysis.

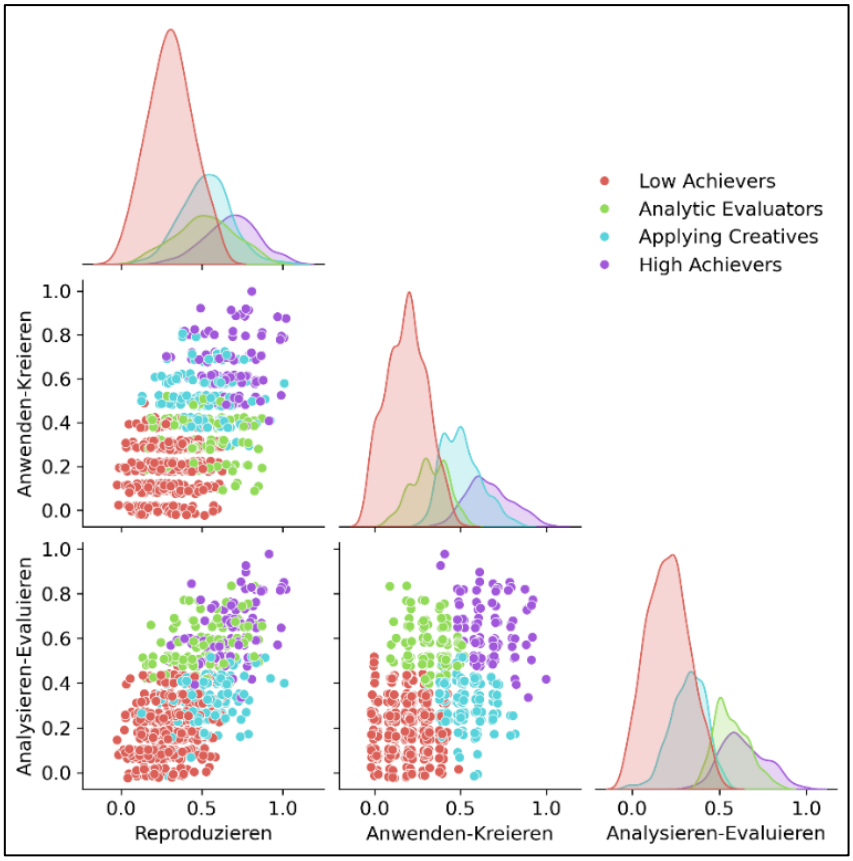


Abbildung 6.6 Paarplot-Darstellung der einzelnen FDW-Score Datenpunkte mit Kompetenzprofilen. Auf der Diagonale werden die jeweiligen geschätzten Wahrscheinlichkeitsverteilungen der Scores der einzelnen Cluster in den jeweiligen Anforderungskategorien mithilfe sog. Kerndichteschätzungen (≈ kontinuierliche Histogramme) dargestellt. Auf den Nicht-Diagonalelementen werden die Score-Paare der einzelnen Testbearbeitungen immer im Rahmen von zwei Anforderungskategorien gegeneinander aufgetragen. Die Scores sind dabei kategorienweise auf das Intervall $[0, 1]$ skaliert. Um eine etwas bessere Darstellung der einzelnen Punkte der Punktwolken zur erhalten, wurde zu den Daten hier jeweils im Intervall $[-0{,}025; 0{,}025]$ gleichverteiltes Rauschen von addiert. Die Bezeichnungen der Cluster werden in Abschnitt 6.5.1 eingeführt.

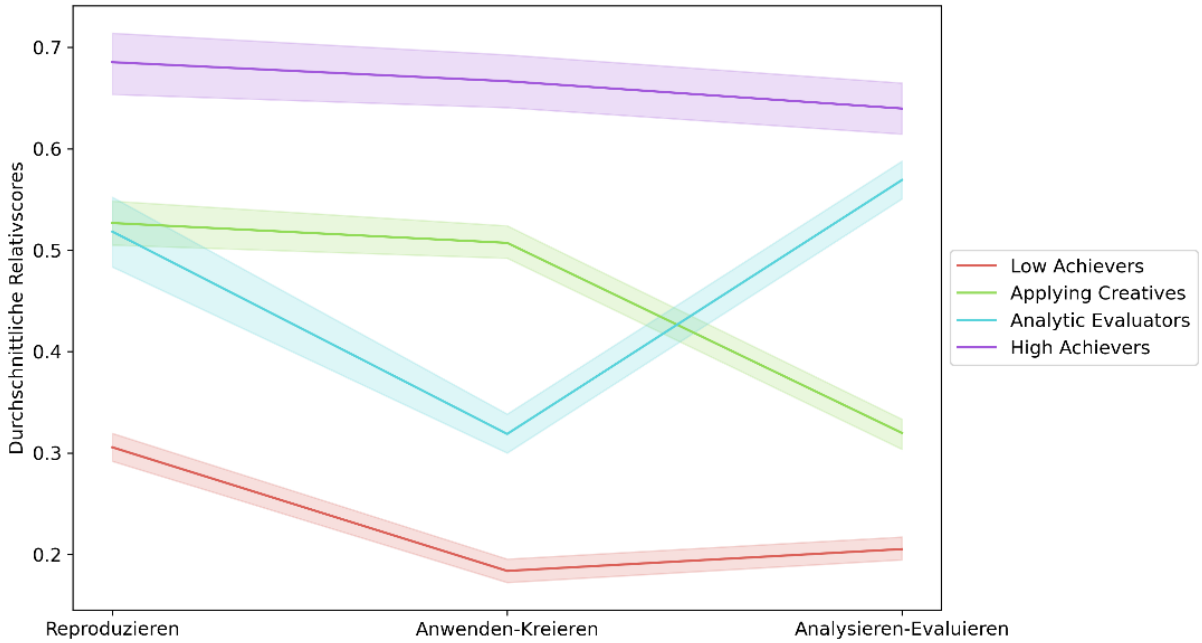


Abbildung 6.7 Linienplot der Score-Mittelwerte der Kompetenzprofile. Auch hier sind die Werte auf das Intervall $[0, 1]$ normiert, d. h., dass beispielsweise die High Achievers durchschnittlich etwa 66 % des Maximalscores in der Dimension Anwenden-Kreieren erreichen. Die hell eingefärbten Bänder um die Mittelwertslinien stellen die Mittelwertsstreuung dar.

Die latenten Kompetenzprofile werden auf Basis ihrer durchschnittlichen Scores in den kognitiven Anforderungsdimensionen (Abbildung 6.7) benannt. Aufgrund der Parallelen zu den (nicht-latenten) Personenclustern aus der Vorgängerstudie (Zeller & Riese, 2025) werden die bereits dort verwendeten englischen Bezeichnungen wiederverwendet. Zudem lassen sich die Kompetenzprofile im Englischen griffiger und insbesondere leicht geschlechtsneutral bezeichnen:

1. *Low Achievers*: Insgesamt in allen Teildimensionen niedriges Kompetenzniveau.
2. *Applying Creatives*: Stärken im Kreieren von Unterrichtselementen und Anwenden von FDW auf beschriebene (Unterrichts-) Situationen u. Ä.
3. *Analytic Evaluators*: Stärken im Analysieren und Bewerten beschriebener (Unterrichts-) Situationen oder beschriebenem Handeln einer Lehrperson u. Ä.
4. *High Achievers*: Insgesamt in allen Teildimensionen hohes Kompetenzniveau.

Die durchschnittlichen FDW-Scores und absolvierten Fachsemester des Physik-Lehramtsstudiums sowie der Anzahl an Proband:innen in den Kompetenzprofilen sind in Tabelle 6.5 dargestellt.

6.5.2 FF2a: Maschine-Mensch Übereinstimmung des Scoring-LMs

Der erste Schritt der Pattern Confirmation ist die Erstellung eines ML-Modells zur automatischen Bepunktung der offenen Aufgaben des verwendeten Testinstruments (FF2a). Dazu wurde hier ein BERT-Sprachmodell zur Bepunktung dieser Aufgaben finegetuned. Das

Modell wurde für drei Epochs trainiert. Die Entwicklung der Werte der verwendeten Cross-Entropy-Loss-Funktion und der Accuracy über das Training sind in Abbildung 6.8 gemittelt über alle CV-Splits gegen die durchlaufenden Epochs aufgetragen. Man erkennt beginnendes Overfitting ab dem Beginn der dritten Epoch, d. h. das BERT-Modell fängt an, Details des Trainingsdatensatzes „auswendig zu lernen". Da die hier vor allem relevante diskrete Accuracy-Metrik bezüglich der Evaluierungsdaten aber während der dritten Epoch noch weiter leicht ansteigt, zusätzliches Training über die dritte Epoch hinaus aber keine weiteren Performanzzuwächse bewirkte, wird das 3-Epoch Modell verwendet.

Die Performanz des Scoring-Modells ist in Tabelle 6.6 und Abbildung 6.9 dargestellt. Insgesamt erreicht das Modell Maschine-Mensch-Übereinstimmungswerte, die 80 bis 90 % der Mensch-Mensch-Übereinstimmungswerte entsprechen (Tabelle 6.6). Schließt man die MC-Aufgaben in diese Betrachtung mit ein und bewertet fehlende Antworten mit „0 Punkten", so erreicht das Scoring-System eine gute (Döring, 2023) Übereinstimmung von $\kappa = 0.680$ (Mensch-Mensch-Baseline: $\kappa = 0.761$). Das Scoring-Modell zeigt bis auf eine etwas „strengere" Bewertung keine systematischen Verzerrungen der Vorhersagen (Abbildung 6.9).

Tabelle 6.5 Vergleich der latenten Kompetenzprofile in Hinsicht auf Fachsemester, FDW-Gesamtscore und Umfang. Die Spalte „*N* (Gesamtdatensatz)" schließt die 41 Bearbeitungen des Testinstruments, die aus der Erstellung des Clustermodells ausgeschlossen wurden, mit ein (siehe Abschnitt 6.4.2, 6.4.3).

Kompetenzprofil	**Fachsemester Physik**		**FDW-Gesamtscore**		***N***	***N*** (Gesamtdatensatz)
	M	***SD***	***M***	***SD***		
Low Achievers	2,87	2,56	12,86	2,07	411	470
Applying Creatives	5,31	3,59	18,97	2,93	166	167
Analytic Evaluators	5,28	3,92	19,32	2,58	112	113
High Achievers	6.96	3,71	25,44	3,53	96	96

Tabelle 6.6 Scorer Performanz. Hier wurden keine Missings oder MC-Aufgaben betrachtet.

	Mensch-Maschine-Übereinstimmung	**Mensch-Mensch-Baseline**
Anzahl Daten (*N*)	15 600	4 748
Accuracy	0,751	0,813
Cohens *κ*	0,560	0,665

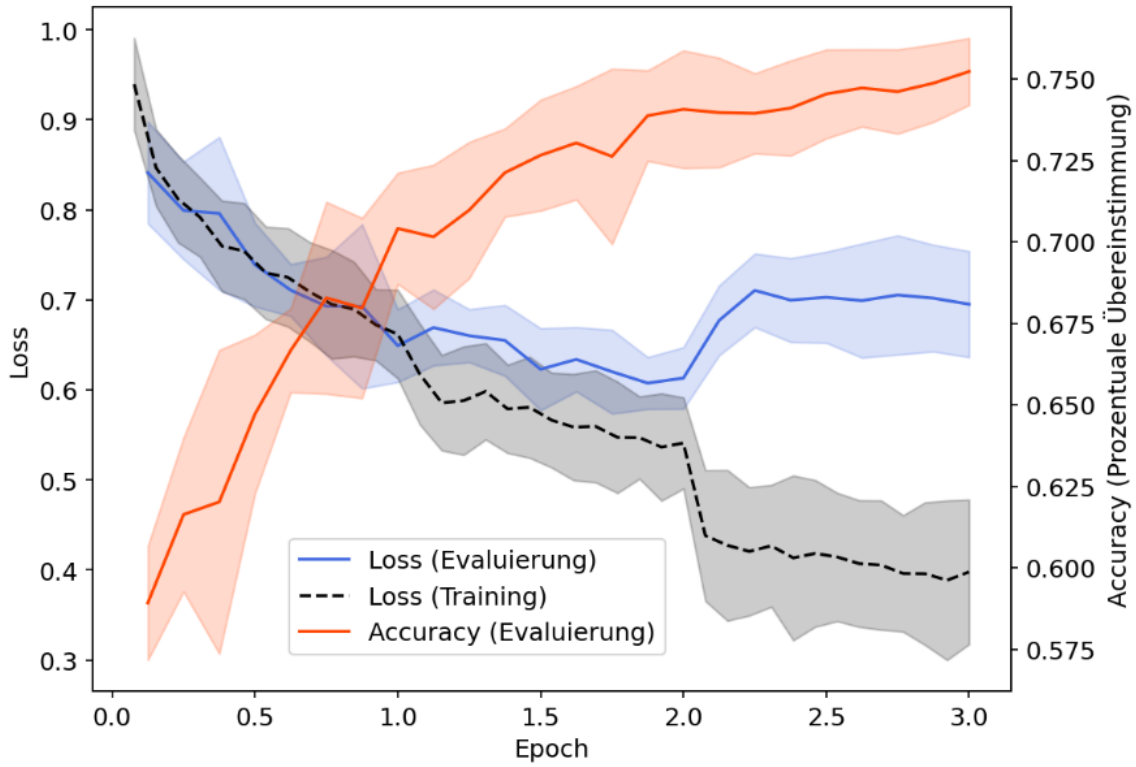


Abbildung 6.8 Learning Curves des Scorer-Trainings "gemittelt" über die 10 CV-Splits. Die eingefärbten Bereiche stellen die Standardabweichungen der jeweils 10 Werte (10 CV-Splits) pro Log-Punkt dar.

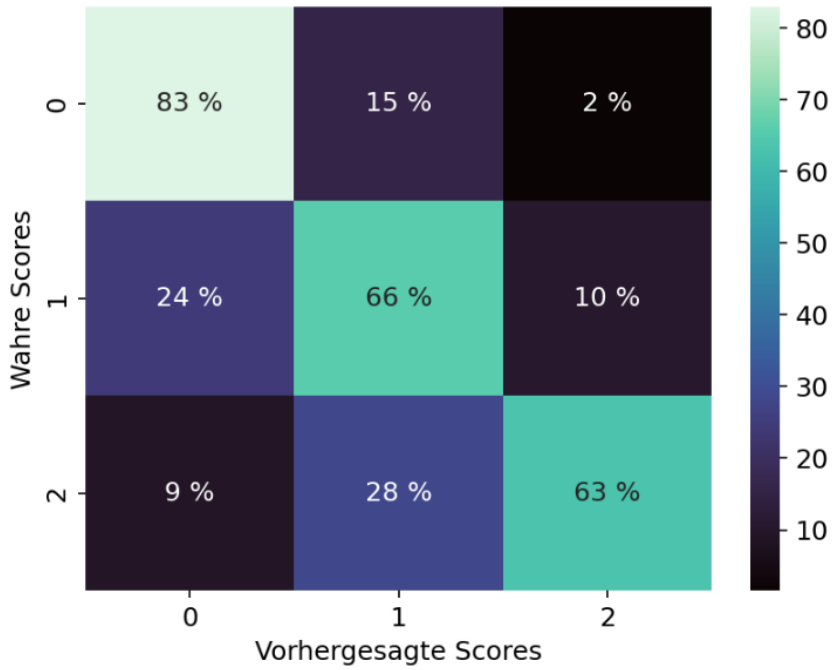


Abbildung 6.9 Darstellung der Score-Vorhersageübereinstimmung als Heatmap. Die abgebildeten Werte sind Zeilen-weise normiert, d. h. der Eintrag „15 %" in der oberen mittleren Zelle ist beispielsweise wie folgt zu interpretieren: 15 % der Bearbeitungen, die von dem/der Kodierer:in mit null Punkten bewertet wurden (die also 0 als Target haben) bewertet der BERT-Scorer mit einem Punkt.

6.5.3 FF2b: Maschine-Mensch Übereinstimmung der latenten Kompetenzprofile

Der zweite Schritt der Pattern Confirmation ist die ML-basierte Zuordnung zu den Kompetenzprofilen auf Basis der automatischen Bepunktung (FF2b). Die Ergebnisse der Evaluierung der automatisierten Zuordnung zu den Kompetenzprofilen mithilfe des „wahren" GMMs und des LR-Modells sind in Tabelle 6.7 und Abbildung 6.10 dargestellt. In Tabelle 6.7 sind zudem die Übereinstimmungswerte der beiden menschlichen Kodierer:innen auf Basis einer Zuordnung mit dem GMM dargestellt. Tatsächlich erreicht das LR-Modell mit $\kappa = 0{,}612$ sogar eine etwas größere Performanz als die Zuordnung auf Basis des „wahren" GMMs mit $\kappa = 0.587$. Die Maschine-Mensch-Übereinstimmung auf Basis des GMMs entspricht dabei 94 % der Mensch-Mensch-Übereinstimmung auf Basis des GMMs ($\kappa = 0{,}624$). Diese Werte können als gute Übereinstimmungen eingeordnet werden (Döring, 2023).

Neben der Klassifikation auf Basis der Bearbeitungen des Testinstruments, d. h. der Sprachantworten zu den offenen Aufgaben des Testinstruments und den Antworten im Rahmen der MC-Aufgaben, wurde ergänzend die Vorhersage direkt auf Basis der manuell kodierten Scores evaluiert. Damit wird wie bei Tschisgale et al. (2023) die Vorhersage der Cluster auf Basis der für die Erstellung des Cluster-Modells genutzten numerischen Daten angestrebt. Die Komplexität dieser Klassifikationsaufgabe ist gegenüber der Zuordnung ausgehend von den „rohen" Test-Bearbeitungen deutlich verringert und ein logistisches Regressionsmodell erreicht hierbei im Rahmen einer 10-fold-CV ($N_{\text{eval}} = 846$) exzellente Übereinstimmungswerte (94,5 %, $\kappa = 0{,}918$).

Insgesamt kann die CGT-Pattern-Confirmation unter Beachtung der Komplexität des zu erfassenden Konstrukts des FDW bzw. der zu erfassenden Kompetenzprofile insbesondere beim Vergleich der Maschine-Mensch-Übereinstimmung mit der Mensch-Mensch-Baseline als erfolgreich angesehen werden.

Tabelle 6.7 Performanz der automatisierten Kompetenzprofil-Zuordnungen. Auch hier wurden zur Bestimmung der Mensch-Maschine-Übereinstimmung alle Validierungsdaten zusammen betrachtet.

	Logistische Regression	**Wahres GMM-Cluster Modell (FF1)**	**Mensch-Mensch-Baseline**
Anzahl Testhefte (*N*)	846	846	267
Accuracy	0,759	0,743	0,787
Cohens κ	0,612	0,587	0,624

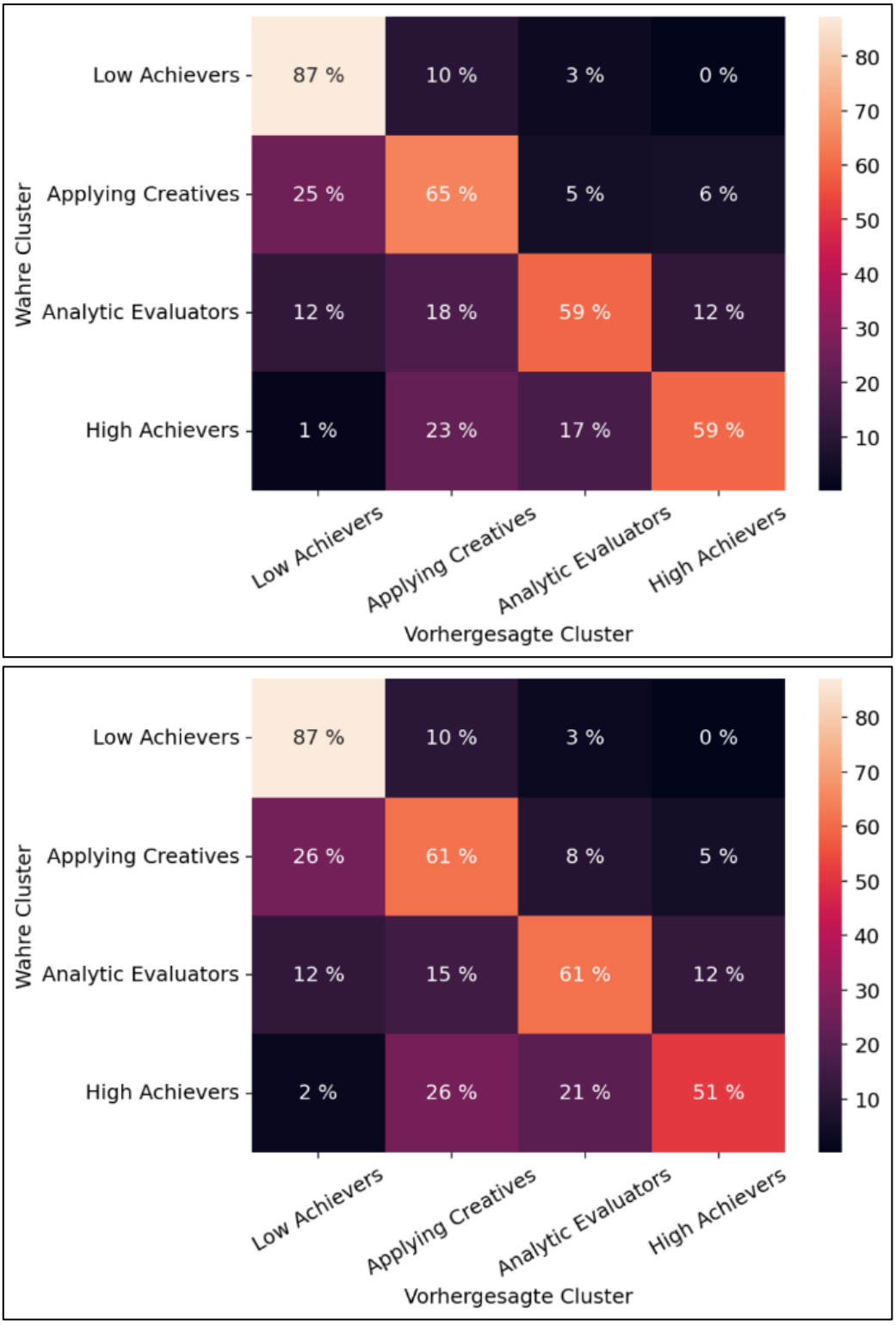


Abbildung 6.10 Darstellung der Kompetenzprofil-Vorhersageübereinstimmungen als Heatmap. Oben sind die Werte für das Logistische Regressionsmodell berichtet, unten die für das „wahre“ GMM-Cluster Modell. Die abgebildeten Werte sind analog zu interpretieren, wie in Abbildung 6.9 beschrieben.[61]

[61] In Abbildung A5 ist auch der entsprechende Plot für die Mensch-Mensch-Übereinstimmung dargestellt.

6.6. Diskussion

6.6.1 Zusammenfassung und Einordnung

Für die empirisch fundierte Weiterentwicklung der Modellierung des Fachdidaktischen Wissens (FDW) sind datenbasierte Beschreibungen der inneren Struktur des FDW notwendig (Kapitel 2). Zu diesem Zweck wurde im vorliegenden Beitrag eine explorative empirische Analyse nicht-hierarchischer Strukturen des FDW angehender Physiklehrkräfte durchgeführt. Im Rahmen von FF1 wurden vier latente nicht-hierarchische Kompetenzprofile ermittelt, deren Robustheit und Validität in den Analysen zu FF2a und FF2b bestätigt wurden. Dazu wurde die Gesamtanalyse im Sinne der Computational Grounded Theory (CGT, Nelson, 2020) strukturiert. Die explorative Analyse der Kompetenzprofile (FF1) stellt in diesem Framework eine Pattern Detection dar. Menschliches Expertenwissen und Interpretationskraft auf Basis des Forschungsstands ist hier nicht in einem separaten CGT-Pattern Refinement Schritt, sondern bereits zur Vorbereitung der Pattern Detection in Form der Re-Analyse der Testaufgaben im Kontext kognitiver Anforderungskategorien und einer entsprechenden Aggregierung der Scores eingeflossen. Im Rahmen der Pattern Confirmation wurde ein Assessment System auf Basis eines BERT-Sprachmodells (FF2a) sowie eines Klassifikationsmodells (FF2b) erstellt, das Testbearbeitungen den Kompetenzprofilen automatisiert zuordnen kann und somit im Sinne der CGT die gefundenen Strukturen bestätigt.

Die vier erhaltenen Kompetenzprofile *Low Achievers*, *Applying Creatives*, *Analytic Evaluators* und *High Achievers* stellen somit vier latente Gruppen von Proband:innen mit prototypischen Stärken und Schwächen dar. Die Applying Creatives und Analytic Evaluators zeigen dabei deutlich eine nicht hierarchische Struktur im Sinne von FF1 (Abbildung 6.7, Anhang E). Auffällig ist darüber hinaus, dass selbst die High Achievers noch einiges Verbesserungspotenzial bezogen auf die Maximalpunktzahlen des Testinstruments haben (Abbildung 6.7). Ähnliches wurde bereits bei früheren Einsätzen des Testinstruments beobachtet (Gramzow et al., 2013; Riese et al., 2017). Beeinflusst ist diese Beobachtung aber auch durch den hohen Anteil an Anfängerstudierenden im Datensatz; Studierende in den ersten zwei Studienjahren bilden ca. 62 % der Gesamtstichprobe.

Die gefundenen Kompetenzprofile und die zugehörigen kognitiven Anforderungskategorien lassen sich zudem im Rahmen des RCM of PCK auf das ePCK zurückbeziehen. Von ePCK wird angenommen, dass es sich im Rahmen des sog. „Plan-Teach-Reflect-Cycles“ (PTR-Cycle, Alonzo et al., 2019) iterativ entwickelt. Empirische Ergebnisse stützen dieses Modell (z. B. Behling et al., 2022b). Dem PTR-Cycle folgend werden somit auch entsprechende einzelnen ePCK-Komponenten, d. h. ePCK-plan, ePCK-teach und ePCK-reflect unterschieden. Die kognitiven Anforderungskategorien, die sich im Rahmen der hier vorgestellten Studie für das FDW (im Sinne eines pPCK) als bedeutsam erwiesen haben, können auch als empirische Hinweise auf die Existenz einer entsprechenden inneren Struktur des pPCK gedeutet werden, wie beispielsweise die Trennung von „pPCK-apply“ und „pPCK-analyze“. Für eine gesichertere Aussage sollten an dieser Stelle aber weitere ggf. konfirmatorische Analysen insbesondere auch mit anderen Datensätzen durchgeführt werden. Darüber hinaus sollte die Beziehung zwischen den potenziellen pPCK-Komponenten und den ePCK-Komponenten in

den Blick genommen werden. Es ist bisher unklar, ob einzelne pPCK-Komponenten auch mit einzelnen ePCK-Komponenten korrespondieren, oder, ob ggf. erst die Integration der pPCK-Bestandteile zur Steigerung oder Entwicklung von ePCK führt.

Die fachdidaktischen Facetten stellen prinzipiell eine Alternative zu den kognitiven Anforderungsdimensionen bei der theoriegeleiteten Akkumulierung der Scores als Vorbereitung der Cluster Analysen dar. Fachdidaktische Facetten bzw. Inhalte werden, wie bereits beschrieben, in unterschiedlichen FDW-Testinstrumenten im naturwissenschaftlichen Bereich genutzt (z. B. Kröger, 2019; Tepner et al., 2012) und haben sich im Rahmen von IRT-Modellierungen bereits als trennbare Subskalen erwiesen (Riese et al., 2017). Allerdings sind die Facetten gegenüber der Nutzung der kognitiven Anforderungen weniger gut auf andere Testinstrumente bzw. Operationalisierungen generalisierbar, da in den Einzelprojekten meist eine unterschiedliche Auswahl von Facetten betrachtet wird (z. B. Kirschner, 2013). Für die kognitiven Anforderungskategorien weisen die Ergebnisse der projektübergreifenden Analyse von Zeller et al. (2024) hingegen auf eine Übertragbarkeit der Kategorien auf unterschiedliche Testinstrumente hin. Das Generalisierungspotenzial der kognitiven Anforderungskategorien sollte dementsprechend auch genutzt werden, um zu überprüfen, ob ähnliche Analysen anderer FDW-Testinstrumente die hier vorgestellten Ergebnisse in Form der latenten Kompetenzprofile unterstützen. Dies gilt insbesondere vor dem Hintergrund der Limitation der hier berichteten Ergebnisse aufgrund der Beschränkung des verwendeten Testinstruments auf den Fachinhalt Mechanik.

Der Bedarf an weiteren Analysen zur Überprüfung der Reproduzierbarkeit der vorgestellten Ergebnisse gilt insbesondere vor dem Hintergrund, dass der vorgestellten Analyse ein komplexer methodischer Workflow zugrunde liegt und sie aufgrund ihres explorativen Charakters viele „Moving Parts“ umfasst. Aus methodischer Sicht wäre hier auch die Nutzung eines Testinstruments interessant, welches eine gleichmäßigere Anzahl an Aufgaben in den einzelnen kognitiven Anforderungskategorien umfasst, um eine bessere Auflösung der Kompetenzprofile im Rahmen der GMMs (oder ähnlicher Modelle) erreichen zu können. Darüber hinaus wäre die Nutzung von Methoden zum Umgang mit ungleich verteilten Datensätzen (z. B. Lemaître et al., 2017) mit Blick auf das Training des BERT-Modells für das automatische Scoring lohnend, um zu untersuchen, ob sich die in Abschnitt 6.5.2 beschriebene „Strenge“ des Systems abmildern lässt.

6.6.2 Ausblick

Wie bereits beschrieben, wird theoriebasiertes menschliches Expertenwissen und menschliche Interpretationskraft im Sinne der CGT in der hier vorgestellten Analyse bereits während bzw. vor dem eigentlichen Pattern Detection Schritt einbezogen. Ein zusätzliches Pattern Refinement zur weiteren Detailbeschreibung der Kompetenzprofile wurde hier aus Platzgründen nicht vorgestellt. Analog zur Vorgängeranalyse (Zeller & Riese, 2025) ließen sich aber auch hier mithilfe von Topic Models (z. B. Blei, 2012; Roberts et al., 2019) praktikabel Zusammenhänge zwischen der Kompetenzprofil-Zugehörigkeit und der Nutzung bestimmter Begriffe bzw. Fokussierung auf bestimmte Konzepte untersuchen. Eine vorläufige Analyse dieser Art zeigt beispielsweise, dass die High Achievers sich in ihren Antworten auf

die offenen Testaufgaben deutlich stärker auf Konzepte aus dem Bereich der Schülervorstellungen (Begriffe wie „Schülervorstellungen“, „kognitiv“, „Konflikt“ oder „Alltagserfahrungen“) fokussieren als die übrigen Kompetenzprofile.

Neben der Nutzung für die Pattern Confirmation bietet auch das hier vorgestellte BERT-Modell zum automatisierten Scoring Ansatzpunkte, selbst Gegenstand weiterer Untersuchungen zu werden. Dazu könnten (ohne Anspruch auf Vollständigkeit) Analysen (1) zur Bedeutung bestimmter Merkmale von Antworten (Antwortlänge, Wortwahl etc.) für die Performanz des Scoring-Modells (z. B. Zesch et al., 2023), (2) zur Erklärung von bestimmten Modellentscheidungen (z. B. Gombert et al., 2023) sowie (3) zur Fairness des Modells (z. B. Barocas et al., 2023) durchgeführt werden. Erste Ansätze in diese Richtungen wurden bereits erprobt:

1) Bei einer Aufgaben-weisen Betrachtung zeigte sich, dass die Maschine-Mensch-Übereinstimmung nicht nennenswert mit der durchschnittlichen Antwortlänge in den Aufgaben, aber signifikant mit der Mensch-Mensch-Übereinstimmung zusammenhängt (Spearman-Korrelation von 0.612^{**} zwischen den Maschine-Mensch-κs und Mensch-Mensch-κs). Aufgaben, bei denen eine hohe Interrater-Reliabilität besteht, scheinen also auch besonders reliabel automatisiert bepunktet zu werden.
2) Ähnlich dem Ansatz von Gombert et al. (2023) wurden erste Analysen zur Bedeutsamkeit bestimmter Worte für die Bepunktung durch das BERT-Modell durchgeführt. Dabei wurde die sog. Attribution-Metrik verwendet, die beschreibt, wie stark jedes einzelne Wort für oder gegen die Klassifikation des eingegebenen Textes „arbeitet“ (Sundararajan et al., 2017[62]). Die Ergebnisse dieses Vorgehens lassen sich wie in Abbildung 6.11 gezeigt visualisieren. Bei einer ersten Aufgaben-übergreifenden Aggregierung der Bedeutsamkeit der Worte konnte bislang keine interessante Systematik festgestellt werden, was wahrscheinlich daran liegt, dass in den einzelnen Aufgaben unterschiedliche fachliche und fachdidaktische Konzepte relevant sind.

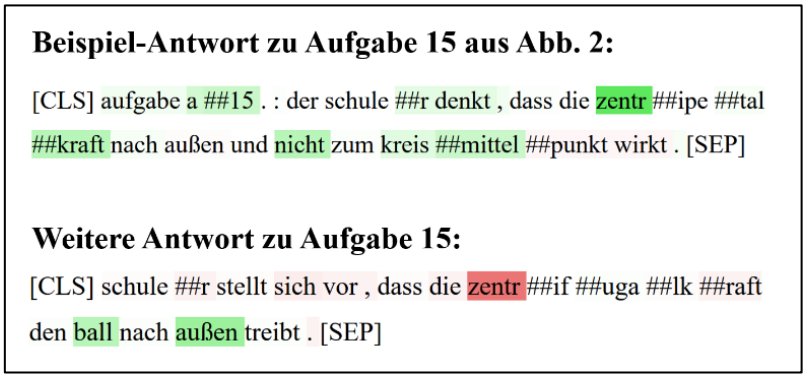


Abbildung 6.11 Darstellung der Attribution der einzelnen Worte offener Testantworten zu Aufgabe 15 auf die jeweilige Klassifikation durch den Scorer. Eine grüne (bzw. rote) Einfärbung bedeutet, dass das Wort die Entscheidung für die Score-Entscheidung positiv (bzw. negativ) beeinflusst hat. Die ungewöhnliche Formatierung des Textes hängt mit der Vorverarbeitung der Inputs durch das BERT-Sprachmodell zusammen.

[62] Dazu wurde das Python-Paket transformers-interpret (https://github.com/cdpierse/transformers-interpret) verwendet.

3) Zur Analyse der Fairness von ML-Modellen werden üblicherweise potenziell benachteiligte Gruppen und die Performanz bezüglich dieser Gruppen betrachtet (z. B. Barocas et al., 2023). Im hier verwendeten Datensatz liegen kaum derartige Daten vor. Für die Geschlechter ergab eine Betrachtung keine bedeutsamen Unterschiede in der Performanz ($\kappa_W = 0{,}57 \pm 0{,}05,\ \kappa_M = 0{,}55 \pm 0{,}05$[63]) und auch keinen Unterschied bezüglich der zu „strengen" Bewertung.

Das erhaltene Modell lässt sich im Rahmen eines automatisierten Systems für das Assessment des FDW flexibel einsetzen. Ein entsprechendes Webtool unter Nutzung von Open-Source Software mit einem Interface zur digitalen Bearbeitung des Testinstruments und einer automatisierten Erstellung eines Reports ist bereits angelegt. Es liefert die Rückmeldungen zu Bearbeitungen des Tests je nach verfügbarer Hardware innerhalb weniger Sekunden. Dabei ist geplant, Zuordnungen zu den Kompetenzprofilen und Summenscores bezüglich der kognitiven Anforderungen und Facetten in ein formativ nutzbares Feedback einzuschließen. Die Rückmeldung von Scores in den einzelnen Aufgaben wird nicht angestrebt, denn Testinstrumente wie das genutzte lassen üblicherweise keine reliable Einschätzung auf Ebene der Einzelaufgaben zu.

Das Assessment-System aus Scoring- und Cluster-Modell kann genutzt werden, um Studierenden unmittelbares, inhaltliches Feedback zu ihren Kompetenzstand im FDW zu bieten. Es kann dabei helfen, Verbesserungspotenziale zu identifizieren und ggf. gezielt Lerngelegenheiten zu empfehlen. Über eine individuelle Nutzung hinaus könnte das System auch für Lehrende von Interesse sein, die mithilfe einer Einordnung ihrer Lerngruppen gezielt Lehrinhalte auswählen oder Materialien gestalten können. Auch für weitere Forschungszwecke könnte ein automatisiertes digitales System den bislang eher schwierigen Zugriff auf große Stichproben erleichtern und gleichzeitig den Aufwand bei der Kodierung offener Aufgaben minimieren.

Der hier dargestellte Workflow zur Erstellung und Evaluierung eines automatisierten Assessment-Systems auf Basis eines Testinstruments mit offenen und geschlossenen Aufgaben, inklusive der explorativen Untersuchung von Kompetenzprofilen, ist nicht auf Konstrukte wie das FDW beschränkt. Die Abstrahierung des genutzten Python-Codes für die Analysen und das Webtool für eine flexible Übertragung auf andere Testinstrumente ist in Arbeit.

[63] Die Unsicherheiten dieser Übereinstimmungswerte wurden mithilfe der Bootstrap-Methode aus den Vorhersagen ermittelt.

Beiträge der Autoren

- *Datenakquise*: Josef Riese (und andere Wissenschaftler:innen, die an diesem Artikel nicht konkret beteiligt waren)
- *Methode und Analysen*: Jannis Zeller
- *Ergebnisinterpretation*: Jannis Zeller
- *Ursprünglicher Entwurf*: Jannis Zeller, Josef Riese
- *Review und Überarbeitung*: Jannis Zeller, Josef Riese
- *Fördermittelbeschaffung*: Jannis Zeller, Josef Riese

Förderung

Professionskompetenz im Lehramtsstudium Physik“ (Akronym *ProfiLe-P+*) wurde vom Bundesministerium für Bildung und Forschung im Rahmen des BMBF-Rahmenprogramms ”Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor - Validierungen und methodische Innovationen“ (Akronym *KoKoHs*) unter dem Kennzeichen 01PK15005A-D gefördert.

Die hier verwendeten Daten stammen aus dem o. g. Projekt. Das Manuskript ist im Rahmen einer kumulativen Promotion entstanden, die mit einem Promotionsstipendium der Studienstiftung des deutschen Volkes gefördert wurde.

Interessenskonflikt

Jannis Zeller und Josef Riese erklären, dass keine Interessenkonflikte vorliegen

6.7. Kommentare und Ergänzungen

Da der Fokus von Artikel 3 eher auf den inhaltlichen Erkenntnissen liegt / liegen sollte, mussten viele methodische und technische Anmerkungen stark gekürzt oder vollständig gestrichen werden. Daher folgen nun zu Artikel 3 recht umfangreiche zusätzliche Informationen und Analysen.

6.7.1 Zusätzliche Daten zu den latenten Kompetenzprofilen

Die latenten Kompetenzprofile stellen den Kern der inhaltlichen Analyse des FDW dar und in Tabelle 6.5 wurden bereits einige zusätzliche Informationen über durchschnittliche FDW-Gesamtscores und absolvierte Fachsemester berichtet. In Tabelle 6.8 werden nun noch weitere demographische Daten ergänzt. Auffällig ist hierbei vor allem der hohe Anteil an weiblichen Probandinnen unter den High Achievers und ihr geringer Anteil unter den Analytic Evaluators. In Anhang E sind zudem weitere Werte aus anderen Erhebungen des ProfiLe-P+ Projekts dargestellt, die hier aus Platzgründen nicht alle systematisch eingeführt werden.

Tabelle 6.8 Zusätzliche demographische Daten zu den Kompetenzprofilen.

	Abschluss-note	**Letzte Punktzahl (Schule)**			**Außerschulische Lehrerfahrung**	**Anteil weiblich**
		Physik	**Mathematik**	**Deutsch**		
Anzahl an Daten	649	474	510	504	841	845
Low Achievers	2,35	11,45	10,97	9,19	78 %	34 %
Applying Creatives	2,07	12,38	12,05	10,42	86 %	35 %
Analytic Evaluators	2,09	11,90	11,48	9,86	84 %	24 %
High Achievers	1,83	12,72	12,42	10,63	92 %	43 %

6.7.2 Keine direkte Vorhersage von Clustern ohne Scoring

In Artikel 3 wurde nur am Rande angedeutet, dass eine direkte Vorhersage der Kompetenzprofile ausgehend von den Gesamtantworttexten problematisch ist. Um dies zu evaluieren, wurden vor allem zwei Experimente mit den Text-Daten durchgeführt. Zunächst wurden dazu die Einzelantworten der Personen zu „Gesamttexten“ zusammengefasst, wobei für jede Einzelantwort der Zusatz „Aufgabe X: …“ hinzugefügt wurde, um die einzelnen Aufgaben sprachlich voneinander abzugrenzen. Das Setting ist demnach ein klassisches Supervised-Learning-Problem mit den Kompetenzprofil-Zuordnungen als Labels.

Im ersten Ansatz wurde dasselbe BERT-Modell zur Vorhersage der Kompetenzprofile trainiert, das auch in Artikel 3 zum automatischen Scoring verwendet wird. Problematisch ist, dass dieses Modell nur Texte mit einer Länge von bis zu 512 Token (ca. 320 Worte) verarbeiten kann (man spricht hier auch von *Kontextlänge*), die Gesamttexte aber bis zu 1493 Token (ca.

928 Worte) umfassen. Alle längeren Dokumente werden ab dem 513ten Token gekappt. Bei den Dokumenten, die über 512 Token lang sind, werden dadurch im Durchschnitt ca. 23 % des Dokuments vernachlässigt. Insgesamt werden so ca. 18 % der Gesamttextdaten aller Proband:innen zusammengenommen nicht genutzt. Das BERT-Modell wurde anschließend im Rahmen einer 3-fold-CV (siehe Abschnitt 2.4) analog zum Vorgehen in Artikel 3 evaluiert. Es erreicht eine Accuracy von 57,0 % und ein Cohens κ von 0,262. Die Confusion Matrix ist in Abbildung 6.12 oben dargestellt. Die Übereinstimmung ist deutlich schlechter als die Ergebnisse mit dem automatisierten Scoring als Zwischenschritt.

Aufgrund der Limitation des BERT-Modells durch die geringe Kontextlänge wurde als Alternative über die OpenAI-API ein GPT4o-mini Modell mit einer Kontextlänge von 128.000 Token (OpenAI, 2024b) ebenfalls im Rahmen einer 3-fold-CV trainiert und evaluiert[49]. Es erreicht mit einer Accuracy von 63,3 % und einem Cohens κ von 0,375 ebenfalls nur mittlere Übereinstimmungswerte. In Abbildung 6.12 (unten) erkennt man, dass insbesondere das Analytic Evaluators Kompetenzprofil von beiden Modellen nicht reliabel erkannt wird und auch insgesamt hohe Fehlerquoten auftreten.

Aufgrund dieser Ergebnisse wurde bereits früh im Projekt zum „Scoring-First-Clustering-Second"-Workflow übergegangen, dem auch in Artikel 3 gefolgt wird. Ein wichtiger Eckpfeiler dieses Workflows ist die strikte Trennung von Trainings- und Evaluierungsdaten über beide Vorhersageschritte (erst Scores, dann Kompetenzprofile / Cluster) hinweg, wie auch in Abschnitt 6.7.3 noch einmal detaillierter beschrieben wird.

6.7.3 Zusätzliche Anmerkungen zum Workflow

Zunächst muss betont werden, dass der verwendete Fragebogen in einem analogen Pencil-Paper-Format durchgeführt wurde. Zur computerbasierten Analyse der Antwort-Texte wurden die Papier-Testhefte dementsprechend durch Hilfskräfte digitalisiert. Die Bepunktung der Testhefte ist bereits in früheren Projektphasen manuell vorgenommen worden (Vogelsang et al., 2019). Mithilfe von Personencodes aus dem Demographie-Teil des Fragebogens konnten die Testhefte lückenlos den bestehenden Scores zugeordnet werden. Die Digitalisierung durch die Hilfskräfte wurde im Rahmen der Möglichkeiten engmaschig überwacht, Tippfehler u. Ä. können aber nicht ausgeschlossen werden. Es ist allerdings unwahrscheinlich, dass solche marginalen „Verfälschungen" einen großen Einfluss auf die verwendeten Modelle haben. Eine zukünftig möglicherweise volldigitale Bearbeitung des Testinstruments könnte die Antwortstrukturen aber systematisch verändern. Für die Assessment Modelle empfiehlt sich daher bei einer Nachnutzung im Rahmen eines volldigitalen Settings in jedem Fall gerade zu Beginn ein Monitoring inklusive einer Evaluierung, um sicherzustellen, dass die Performanz der ML-Modelle auch in einem volldigitalen Format den Erwartungen entspricht.

Die Zuordnung der Aufgaben zu den Anforderungskategorien wurde zur Vorbereitung der latenten Profilanalysen analog zu Artikel 2 ebenfalls gemäß des Zuordnungsmanuals (Anhang B) vorgenommen. Dabei wurden die Kategorien zusammengefasst, indem die Aufgaben nach dem „Inklusiven-Oder"-Prinzip zugeordnet wurden. Beispielsweise wird eine Aufgabe der Kategorie Analysieren-Evaluieren zugeordnet, wenn sie der Kategorie Analysieren und / oder der Kategorie Evaluieren zugeordnet ist.

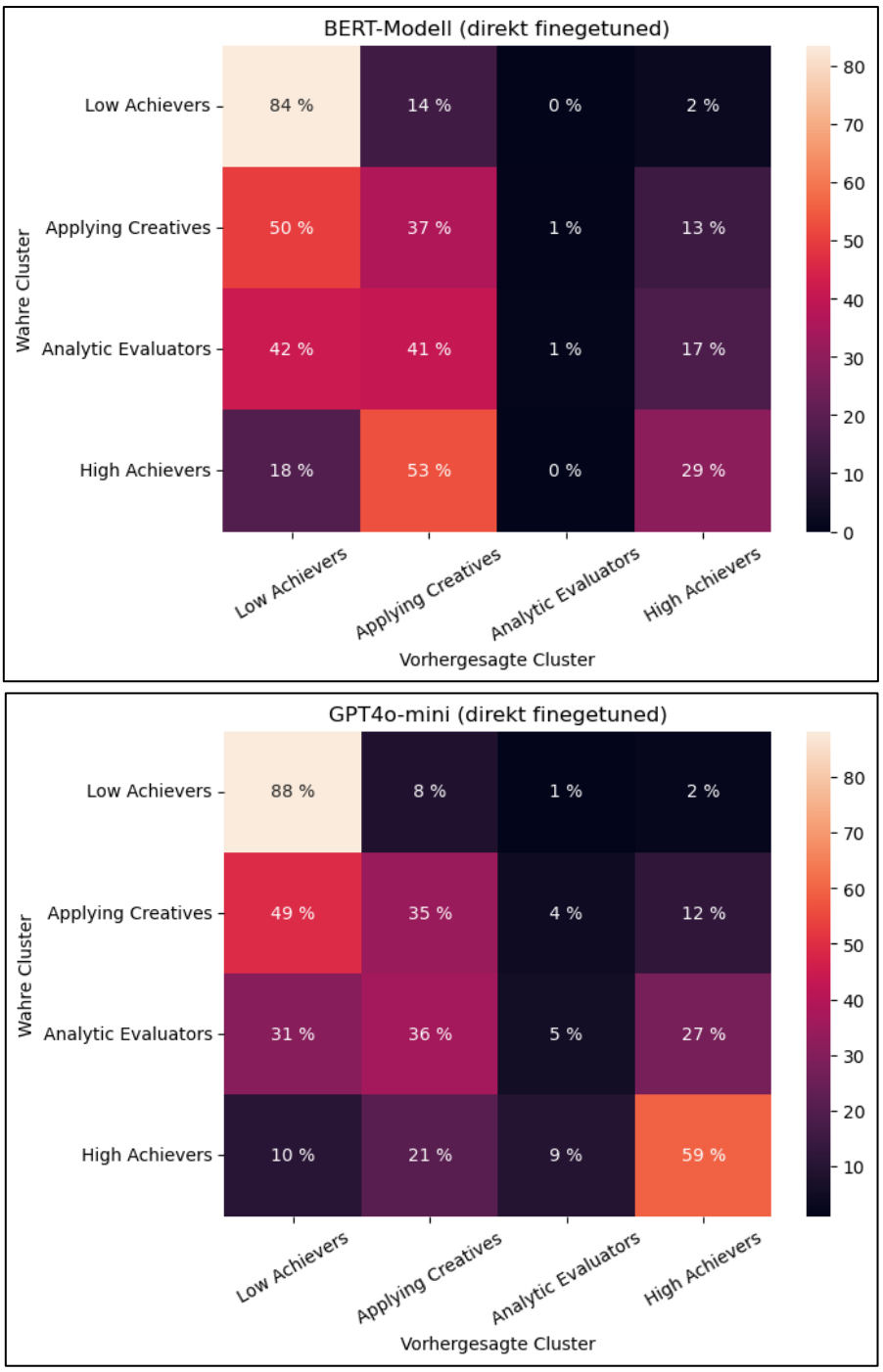


Abbildung 6.12 Confusion-Matrizen der direkten Kompetenzprofil-Vorhersage mit BERT und GPT4o-mini.

Bei der automatisierten Bepunktung des Testinstruments wurde in Artikel 3 und auch in den hier noch folgenden zusätzlichen Analysen nur eine Punktzahl von 0 bis 2 Punkten pro Aufgabe berücksichtigt. Tatsächlich können in Aufgabe 23 allerdings bis zu drei Punkte erreicht werden. Bezogen auf alle Testhefte erreichen allerdings nur 26 Personen drei Punkte in Aufgabe 23, was ca. 3 % der Stichprobe entspricht. Ein ML-Modell zur automatisierten Bepunktung, das alle Aufgaben bepunkten können soll, müsste dementsprechend einen marginalen Anteil an Antworten mit 3 Punkten erkennen (0.17 % bezogen auf die Werte in Tabelle 6.4). Diese Kategorie tritt also deutlich zu selten auf, um durch ein Klassifikationsmodell reliabel erkannt

zu werden. Daher werden für die Analysen zum automatisierten Assessment alle „3 Punkte“ in Aufgabe 23 durch „2 Punkte“ ersetzt. Dadurch entstehende Limitationen werden im Rahmen der Personen-weisen Evaluierung der Zuordnung von Kompetenzprofilen (Abschnitt 6.5.3) sowie der Vorhersage von Subskalenscores (Abschnitt 6.7.4) automatisch berücksichtigt und quantifiziert.

Da nur ein gemeinsames Scoring-Modell für alle Aufgaben verwendet wird, besteht zudem die Möglichkeit, dass das Modell Punktzahlen vergibt, die das Testinstrument eigentlich nicht vorsieht. So kann beispielsweise in Aufgabe 11 nur 1 Punkt erreicht werden, das Modell kann grundsätzlich aber auch 2 Punkte vorhersagen. Für das in Artikel 3 vorgestellte BERT-Modell geschieht eine solche Vergabe unzulässiger Punktzahlen allerdings bemerkenswerterweise für kein einziges der 15.600 verarbeiteten Antwort-Score-Paare – weder in den Trainings- noch den Evaluierungsvorhersagen. Dies wird in Abschnitt 6.7.6 noch einmal aufgegriffen.

In Artikel 3 wurde bereits darauf hingewiesen, dass die CV-Splits personenweise erfolgen, da im Assessment-Setting typischerweise Aussagen auf Personen-Ebene (Kompetenzprofil, Summenscore, Subskalenscore etc.) von Interesse sind. Neue Daten, die das Modell in einem tatsächlichen Assessment-Workflow verarbeiten müsste, wären ebenfalls in Form einer vollständigen Testbearbeitung strukturiert. Die CV-Splits Personen-weise durchzuführen ist also eine notwendige Maßnahme um sog. Data Leakage, also die Vermischung von Trainings- und Evaluierungsdaten (Kapoor & Narayanan, 2023; Kaufman et al., 2012; siehe auch Tschisgale et al., 2025) zu verhindern und verlässlichere Schätzwerte für die Performanz des gesamtem Assessment-Systems zu erhalten, auch wenn dadurch die Komplexität des Workflows deutlich zunimmt.

Um die Schätzwerte für die Assessment-Performanz in einen nützlichen Vergleich einzubetten, wurden in Artikel 3 Mensch-Mensch-Übereinstimmungswerte bzgl. der doppeltkodierten Teilstichprobe von 267 Testheften berichtet (siehe Abschnitt 6.4.4). Für die Bepunktung der einzelnen Aufgaben können hier für einen Vergleichswert direkt die jeweiligen Scores der beiden Kodierer:innen als Vorhersagen bzw. Targets verwendet werden. Für die Evaluierung von weiterführenden Teilen des Assessment Systems, beispielsweise der Vorhersage von Kompetenzprofilen oder Summenscores bzgl. der Subskalen des Testinstruments (s. u.), muss allerdings weitergedacht werden. Grundsätzlich lassen sich unter Nutzung der CV-Splits aus der Automatisierung des Scorings auch solche weiteren, sog. „Downstream-Modelle“ trainieren bzw. evaluieren. Dazu werden im Workflow der Analyse nicht nur die zur Evaluierung des Scoring Systems notwendigen Vorhersagen bzgl. der Evaluierungs-Splits abgespeichert, sondern zusätzlich auch alle Vorhersagen bzgl. der jeweiligen Trainings-Splits. Bei einer k-fold-CV wird dadurch zwar die notwendige Datenmenge zum Speichern der Vorhersagen um den Faktor k erhöht, allerdings können die Datensplits so weitergenutzt werden: Um beispielsweise ein logistisches Regressionsmodell zu evaluieren, welches auf Basis der Scores das jeweilige Kompetenzprofil vorhersagen soll, liegen nun wieder k CV-Splits vor, die direkt verwendet werden können. Dieser Workflow funktioniert, ohne dass die Scoring-Modelle aller Splits erneut trainiert oder genutzt werden müssen – sie müssen nicht einmal mehr verfügbar sein. Dabei werden dann beim Training sowohl des Scoring-Modells als auch des Downstream-Modells ausschließlich die

Trainingsdaten verwendet und Data Leakage somit ausgeschlossen.

Ein Problem bei diesem Vorgehen ist allerdings, dass die Performanz eines solchen Downstream-Modells von der Größe des verwendeten Datensatzes abhängen kann. Für die Automatisierung wird aber der Gesamtdatensatz von 846 Testheften verwendet, was gut dem Dreifachen des doppeltkodierten Datensatzes entspricht. Tatsächlich „trainierte" Downstream-Modelle für den Gesamtdatensatz mit einem analogen trainierten Modell für den Mensch-Mensch-Datensatz zu vergleichen, würde die Performanz des Downstream-Modells gegenüber der jeweiligen Mensch-Mensch-Übereinstimmung also ggf. überschätzen. Daher werden für Vergleiche von Downstream-Modellen mit der Mensch-Mensch-Übereinstimmung meist Modelle gewählt, die nicht mehr weiter trainiert werden. Im Falle der Kompetenzprofil-Vorhersage ist das das „wahre" GMM aus der explorativen Analyse und im Falle der Vorhersage von Subskalenscores werden die Scores der Einzelaufgaben einfach gemäß der Zuordnung der Testaufgaben zu den Subskalen summiert. Die jeweiligen Maschine-Mensch-Übereinstimmungsmaße lassen sich dann sinnvoll mit den Mensch-Mensch-Übereinstimmungsmaßen vergleichen. Im weiteren Verlauf werden darüber hinaus dann aber teilweise trotzdem noch „lernende" Downstream-Modelle wie logistische Regressionsmodelle genutzt, um zu zeigen, dass ggf. durchaus höhere Übereinstimmungswerte erreicht werden können, indem das Downstream-Modell „Fehler" des Scoring-Modells ausgleicht. Es ist dann aber zu erwarten, dass ein analoges Modell für einen gleichgroßen Mensch-Mensch-Datensatz eine entsprechend des vorherigen Vergleichs erhöhte Performanz erreichen würde. Die Performanz von trainierten Downstream-Modellen ist also weniger im Verhältnis zu den vorher berichteten Mensch-Mensch-Übereinstimmungen zu interpretieren, sondern eher absolut.

Auch, wenn eine direkte Zuordnung der Antworttexte zu den Kompetenzprofilen nicht bzw. nur mit unzufriedenstellender Genauigkeit möglich ist (Abschnitt 6.7.2), so liefert dieser zweistufige Ansatz dennoch ein Pattern-Confirmation-Argument in folgendem Sinne: Es wird eine latente Zusammenhangsstruktur zwischen den Antworttexten und den Kompetenzprofilen „mediiert" durch die Bepunktung der Aufgaben und MC-Scores gefunden. Die Antworttexte können dabei als Repräsentationen des Wissens der Proband:innen als kognitives Konstrukt aufgefasst werden (Halliday, 1978). Dadurch liefert die vorgestellte Pattern Confirmation neben der praktischen Nutzbarkeit für ein Assessment sowie den Argumenten für Robustheit und Generalisierbarkeit der Kompetenzprofile hier insbesondere auch ein Argument für die kognitive Validität (Messick, 1995) der Kompetenzprofile.

6.7.4 Zusätzliche Analysen zu den bestehenden Modellen

Bei der Nutzung von ML-Modellen für praktische Anwendungen besteht häufig die Problematik, dass sich die Daten aus dem realen Anwendungsfall systematisch von den beim Training genutzten Daten unterscheiden. Dieses Phänomen wird auch als *Distribution Shift* bezeichnet (z. B. Koh et al., 2021; Webb et al., 2018; siehe auch Martin & Graulich, 2024). In Standard-Evaluierungsworkflows (wie der CV) wird meist ein zugrundeliegender Gesamtdatensatz verwendet, der randomisiert in Trainings- und Evaluierungssegemente unterteilt wird. So können aber potenzielle Distribution Shifts nicht abgebildet werden, da per Konstruktion die Evaluierungs- und Trainingsdaten dabei stets aus derselben Grundgesamtheit

stammen und somit im mathematischen Sinne derselben Verteilung folgen. Daher ist es bei der Evaluierung von ML-Modellen sinnvoll, wenn möglich, neben den „klassischen" Evaluierungstechniken auch eine sog. externale Evaluierung (z. B. Varshney, 2019) anzustreben. Dabei werden Daten aus anderen Bezugsquellen herangezogen, um die Robustheit des Modells gegenüber Distribution Shifts zu evaluieren.

Die oben bereits angesprochene Nutzung der hier entwickelten FDW-Assessment-Modelle in einem volldigitalen Format stellt einen möglichen Distribution Shift dar. Hier liegen allerdings bislang keine Daten vor, die für eine entsprechende externale Evaluierung geeignet wären. Der doppeltkodierte Teildatensatz (267 Testhefte) bietet aber eine Möglichkeit für eine externale Validierung, um die Robustheit des Modells im Allgemeinen einzuschätzen. Die Modellvorhersagen zu den Evaluierungssplits können neben den Bepunktungen von Kodiererin 1 auch mit der alternativen Bepunktung von Kodierer 2 verglichen werden. In einer entsprechenden Evaluierung stimmten die maschinellen Scores (BERT-Modell) in 72,3 % der Fälle ($\kappa = 0{,}508$) mit den Scores des zweiten Kodierers überein, was immer noch als gute Übereinstimmung aufzufassen ist. Auch die Zuordnung zu den Kompetenzprofilen auf Basis des „wahren" GMM (s. o.) lieferte noch gute Übereinstimmungswerde (73,0 % Accuracy, $\kappa = 0.532$). Es kann also begründet davon ausgegangen werden, dass das Modell nur in einem eher geringen Maße „Kodierer:in-spezifische" Strategien erlernt hat, obwohl im Training ausschließlich die Daten der ersten Kodiererin verwendet wurden. Das spricht grundsätzlich für die Robustheit des Modells.

Neben den Kompetenzprofilen sind auf Personen-Ebene auch Summenscores für ein automatisiertes Assessment von Interesse. Dabei ist es naheliegend, die Vorhersagegüte bezüglich theoretisch fundierter Subskalen, d. h. hier den fachdidaktischen Facetten sowie den kognitiven Anforderungskategorien, zu betrachten. Auch der Gesamtscore kann mit in diese Betrachtung aufgenommen werden. Um einen direkten Vergleich zur Mensch-Mensch-Baseline zu ermöglichen, werden hier erneut keine zusätzlichen Modelle trainiert (s. o.), sondern lediglich die bestehenden Score-Vorhersagen zu Summenscores bzgl. der betrachteten Skalen aggregiert[64]. Es werden die 10 Skalen *Gesamtscore*, *Instruktionsstrategien*, *Schülervorstellungen*, *Experimente*, *Fachdidaktische Konzepte*, *Reproduzieren*, *Anwenden*, *Analysieren*, *Evaluieren* und *Kreieren* betrachtet. Für die Mensch-Mensch-Übereinstimmung werden dabei analog zum Vorgehen in Artikel 3 die Bepunktungen von Kodiererin 1 als Targets und die Bepunktungen von Kodierer 2 als Vorhersagen betrachtet. Die Zusammenhänge zwischen den Vorhersagen und den Targets sind in Abbildung 6.13 dargestellt. Man erkennt deutlich, dass die Annahme von linearen Zusammenhängen zwischen den Targets und den Vorhersagen der Subskalenscores angemessen ist. In Tabelle 6.9 sind die Übereinstimmungen mithilfe von Korrelationen und R^2-Werten quantifiziert. Die maschinellen Vorhersagen weisen sowohl im Vergleich zur Mensch-Mensch-Baseline als auch absolut betrachtet hohe Korrelationen mit den menschlichen Bepunktungen auf. Insgesamt kann also davon ausgegangen werden, dass Proband:innen mithilfe des Scoring-Modells valide

[64] Im digitalen Begleitmaterial ist sind auch die Ergebnisse unter Nutzung von linearen Regressionsmodellen zur Vorhersage der Summenscores enthalten. Die Übereinstimmungswerte sind marginal besser, als die hier berichteten.

und reliable Rückmeldungen über ihre Kompetenzen in den unterschiedlichen, durch die Subskalen abgedeckten Kompetenzbereichen erhalten können.

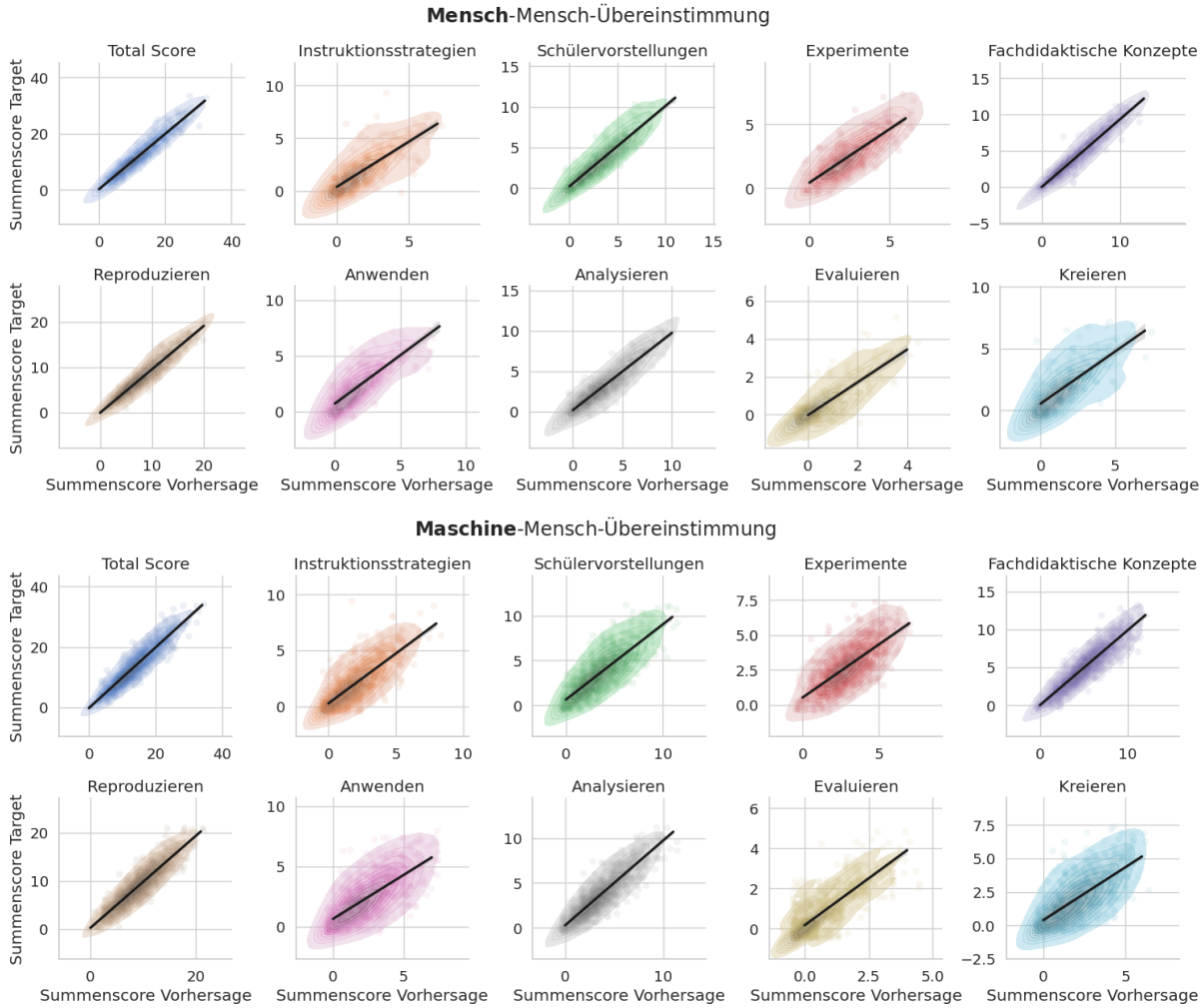


Abbildung 6.13 Darstellung der Zusammenhäng zwischen Summenscore-Targets und -Vorhersagen. Im oberen Plot sind die Mensch-Mensch-Übereinstimmungen, im unteren Plot die Maschine-Mensch-Übereinstimmungen dargestellt. Die schwarzen Linien stellen jeweils Ausgleichsgeraden dar. Die schraffierten Bereiche repräsentieren die Verteilung der Datenpunkte. Man erkennt, dass in allen Fällen ein linearer Zusammenhang angenommen werden kann, auch wenn für die Mensch-Mensch-Daten die Verteilungen der Datenpunkte etwas stärker um die jeweiligen Ausgleichsgeraden konzentriert sind.

Tabelle 6.9 Quantifizierung der Übereinstimmungswerte der Summenscore-Vorhersagen.

	Maschine-Mensch		**Mensch-Mensch**	
	Korrelation (*Person* r)	$\boldsymbol{R^2}$ (*Vorhersage* $\propto$ *Target*)	**Korrelation** (*Person* r)	$\boldsymbol{R^2}$ (*Vorhersage* $\propto$ *Target*)
Minimum	0,71***	0,50	0,75***	0,56
Median	0,84***	0,70	0,88***	0,77
Maximum	0,93***	0,86	0,96***	0,91

Um die Übertragbarkeit der Beobachtungen bezüglich des Sprachgebrauchs der K-Means-Kompetenzprofile aus Artikel 2 auf die latenten Kompetenzprofile aus Artikel 3 einzuschätzen, wurde hier zudem erneut ein STM (Roberts et al., 2019) erstellt. Um eine Vergleichbarkeit zu den Analysen in Artikel 2 herzustellen, wurde das Modell analog – insbesondere ebenfalls mit 6 Topics – konfiguriert. Aus den Wortlisten der Topics lassen sich analog zum Vorgehen in Artikel 2 (Abschnitt 5.5.2) Kurzbeschreibungen / Titel für die Topics ableiten (Tabelle A5). Die sich ergebenden Topics sind ähnlich zu denen in Artikel 2. Auch hier wurde anschließend der Zusammenhang zwischen der Kompetenzprofilzugehörigkeit und den Topics quantifiziert (Abbildung 6.14). Dabei sind deutliche Parallelen zu den Ergebnissen der K-Means Analyse (Figure 5.7) zu erkennen. Insbesondere die Fokussierung der High-Achievers auf das Topic Schülervorstellungen ist hier sogar stärker.

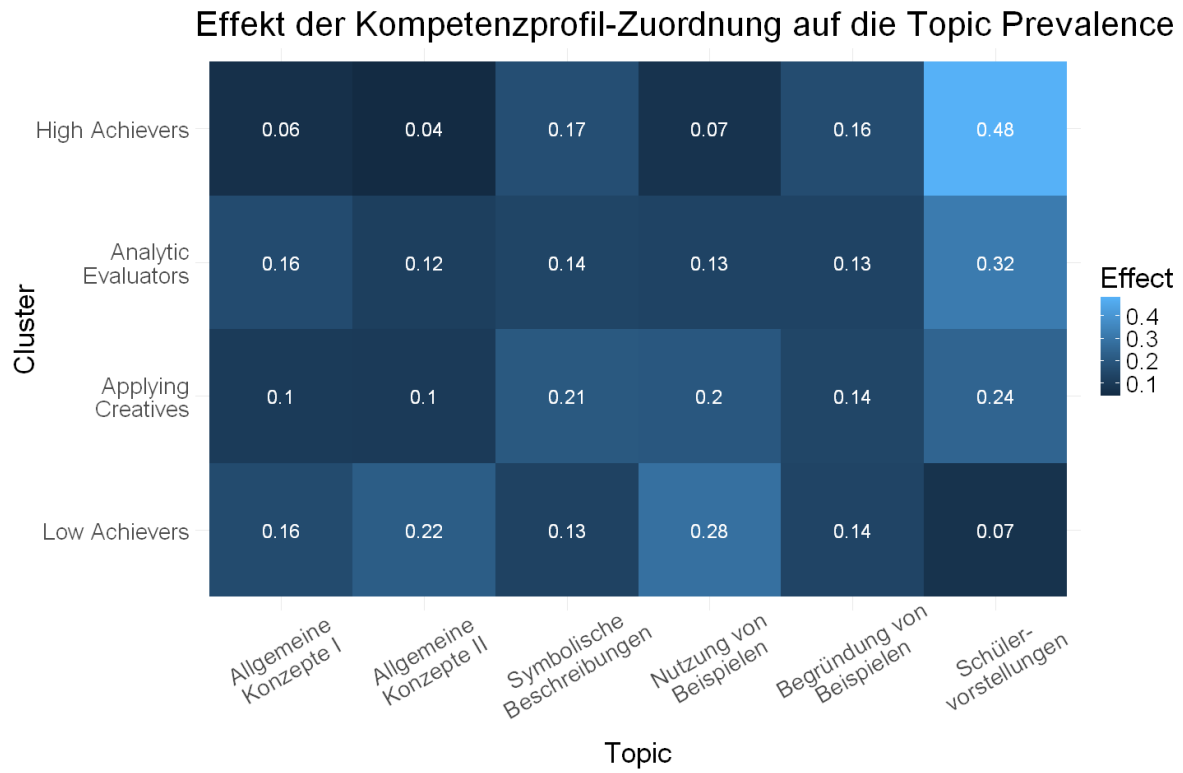


Abbildung 6.14 Zusammenhang zwischen Kompetenzprofilzugehörigkeit (LPA) und Topics.

Zuletzt sei hier noch angemerkt, dass ausführlichere Interpretierbarkeitsbetrachtungen und Fairnessanalysen nicht mehr Teil dieses Projekts sind, auch wenn entsprechende erste Analysen im Ausblick von Artikel 3 genannt wurden. Die Ansätze im digitalen Begleitmaterial zu dieser Arbeit können allerdings als Ausgangspunkt für etwaige Folgeprojekte dienen.

6.7.5 Aufgabenweise Performanzanalysen

Zhai et al. (2021b) arbeiten heraus, dass im Kontext des automatisierten Scorings weitere Forschung zu Einflussfaktoren auf die Modellperformanz notwendig ist. Zesch et al. (2023) stellen im Rahmen ihrer Untersuchung unterschiedlicher Testinstrumente (auf Schulniveau) fest, dass unter anderem die durchschnittliche Länge der Antworten zu einer Aufgabe einen

Einfluss auf die Performanz eines Scoring-Modells haben kann: Je länger die Antworten zu einer Aufgabe im Mittel sind, umso geringer ist die Modellperformanz. Im Rahmen des vorliegenden Projekts liegen einige weitere interessante Daten vor, die auf die Modellperformanz bezogen werden können. Somit kann hier ein Beitrag zum Forschungstand geleistet werden, auch, wenn die Ergebnisse aufgrund der vergleichsweisen kleinen Anzahl an Aufgaben eher als Indizien aufzufassen sind. Dazu wird hier die Performanz des BERT-Scoring Modells aus Artikel 3 aufgabenweise genauer betrachtet. Als potenzielle Einflussfaktoren werden herangezogen:

1. Die Anzahl an Antworten, die für die jeweilige Aufgabe vorhanden sind. Beispielsweise sind im Datensatz zu Aufgabe 17 nur 454 Antworten vorhanden, während zu Aufgabe 3 825 Antworten verfügbar sind.
2. Die „Schiefe“ (Kokoska & Zwillinger, 2000, Abschnitt 2.2.24.1) der Verteilung der Scores der jeweiligen Aufgaben. Die Schiefe ist größer, je ungleichmäßiger die Punktzahlen zwischen 0, 1 und 2 verteilt sind.
3. Die durchschnittliche Länge der Antworten zur jeweiligen Aufgabe.
4. Die Inter-Rater-Reliabilität in Form des Mensch-Mensch-Cohens-κ für die jeweilige Aufgabe.

Zur Quantifizierung der Modellperformanz wird das Maschine-Mensch-Cohens-κ verwendet, da so die ungleichmäßige Verteilung der Score-Labels berücksichtigt wird. Für die Analyse wurde ein z-standardisiertes lineares Regressionsmodell mit den o. g. vier Prädiktoren und dem Maschine-Mensch-Cohens-κ als abhängige Variable erstellt (Tabelle 6.10, Abbildung 6.15). Relevante Einflussfaktoren sind diesem Modell zufolge die durchschnittliche Antwortlänge sowie die Mensch-Mensch-Übereinstimmung: Die durchschnittliche Antwortlänge hängt signifikant negativ und die Mensch-Mensch-Übereinstimmung signifikant positiv mit der Modellperformanz zusammen. Beide Einflussfaktoren weisen eine große Effektstärke (Regressions-β) auf. Die Anzahl an verfügbaren Antwort-Score Paaren für das Modelltraining sowie die Schiefe der Score-Verteilung zeigen hier (erstaunlicherweise) keinen signifikanten Zusammenhang zur Modellperformanz.

Tabelle 6.10 Einflussfaktoren auf die (aufgabenweise) Performanz des BERT-Scoring-Modells (Regression). Da hier die aufgabenweise Performanz untersucht wird, entspricht das N des Regressionsmodells ($F(4, 18) = 5{,}596$, $p = 0.004^{**}$) der Anzahl an Aufgaben mit offenem Antwortformat im Testinstrument, d. h., $N = 23$.

Prädiktor	$\boldsymbol{\beta}$	$\boldsymbol{T}$	$\boldsymbol{p}$**-Wert**
Intercept	0	0	1
Anzahl an Antworten	$-0{,}096$	$-0{,}605$	0,553
Schiefe der Punktzahl-Verteilung	0,037	0,227	0,823
Durchschnittliche Antwortlänge	$\mathbf{-0,471}$	$\mathbf{-2,683}$	$\mathbf{0,015}^{*}$
Mensch-Mensch-κ	$\mathbf{0,417}$	$\mathbf{2,307}$	$\mathbf{0,034}^{*}$

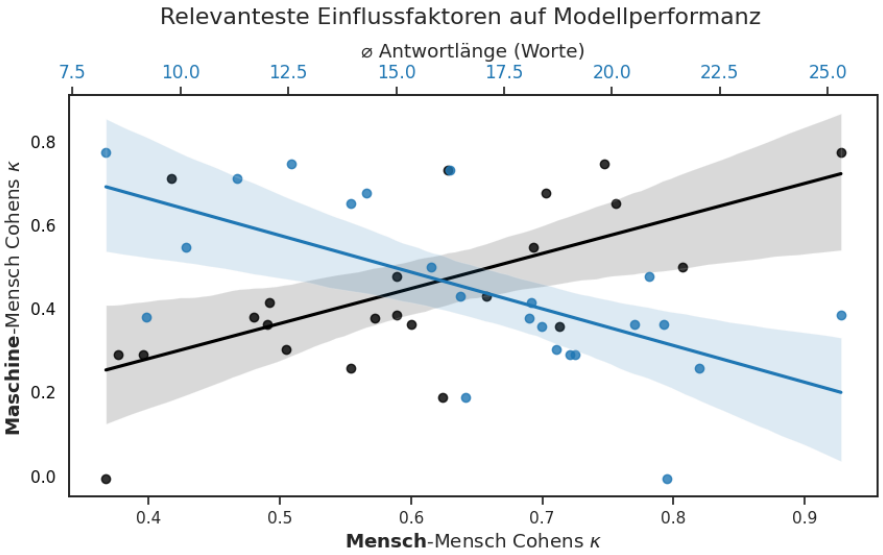


Abbildung 6.15 Relevanteste potenzielle Einflussfaktoren auf die (aufgabenweise) Performanz des BERT-Scoring-Modells (Regression).

6.7.6 Embedding Basierte Scoring-Modelle

Im Folgenden werden Embedding-basierte (s. u. sowie Abschnitt 2.7) Scoring-Modelle vorgestellt, die im Rahmen der Analysen zu Artikel 3 exploriert wurden. Solche Modelle sind interessant, da sie die Dauer des Trainings- bzw. Evaluierungsworkflows gegenüber dem Finetuning eines gesamten Sprachmodells um einen Faktor 5 bis ~100 verkürzen und für ihre Nutzung deutlich weniger Rechenleistung und Speicher benötigt wird. Darüber hinaus wird die Auswirkung bestimmter Vorverarbeitungs- und Workflowschritte anhand solcher Modelle evaluiert.

Sämtliche Modelle und Workflow-Alternativen sind mithilfe von CVs evaluiert. Dabei ist zu beachten, dass die Anzahl an verwendeten CV-Splits einen Einfluss auf die Schätzwerte der Performanz des erhaltenen Modells haben kann: Je mehr CV-Splits verwendet werden, umso höher und akkurater sind typischerweise die Schätzwerte für die Modellperformanz (Abbildung A4). In einer 10-fold-CV werden beispielsweise in jedem CV-Schritt ca. 90 % des Datensatzes für das Training und 10 % für die Evaluierung genutzt. Das Modell kann also einen großen Teil der Varianz im Datensatz potenziell erlernen. In einer 3-fold-CV werden in jedem CV-Schritt dagegen nur ca. 67 % des Datensatzes für das Training und 33 % für die Evaluierung genutzt. Das heißt, hier wird weniger Varianz des Datensatzes im Training abgedeckt. Die genauesten Performanzschätzwerte erhält man daher mit einer sog. „Leave-One-Out"-CV (z. B. Géron, 2019), bei der jeder einzelne Datenpunkt einmal als „Evaluierungsdatensatz" genutzt wird. Für die meisten Anwendungsfälle ist eine solche Evaluierung aber zu aufwändig; im Falle dieses Projekts müsste dafür jedes Modell im CV-Workflow 846-mal trainiert werden. Für das BERT-Modell aus Artikel 3 ergibt sich in der 10-fold-CV eine Accuracy von 75,1 % und ein Cohens κ von 0,560. In einer 3-fold-CV ergeben sich für dasselbe Modell eine Accuracy von 74,2 % und ein Cohens κ von 0,544. Die Unterschiede sind also gering, aber spürbar. Aus Zeit- und Effizienzgründen konnte nicht jedes Modell, welches hier exploriert wurde, im Rahmen einer 10-fold-CV evaluiert werden.

Insbesondere bei den finegetuneten Modellen wurde daher teilweise lediglich eine 3-fold-CV durchgeführt. Dabei ist zu beachten, dass primär Modelle mit der gleichen Anzahl an CV-Splits direkt miteinander verglichen werden[65]. Wenn im Folgenden keine explizite Anmerkung zu einem Modell vorhanden ist, kann davon ausgegangen werden, dass es sich bei finegetuneten Modellen um eine 3-fold-CV und bei Embedding-basierten Modellen um eine 10-fold-CV handelt.

Wie bereits eingangs in Abschnitten 2.6 und 2.7 dargestellt, ist ein wesentliches Element der automatisierten bzw. ML-basierten Sprachverarbeitung die Repräsentation von Worten und Texten als Zahlen. Im mathematischen Sinne handelt es sich bei diesen Zahlen um Vektoren, die im Kontext der automatisierten Sprachverarbeitung meist Embeddings genannt werden. Embeddings lassen sich aus Transformer-Sprachmodellen (siehe Abschnitt 2.7) über Embedding-Layer und Zwischenrepräsentationen extrahieren. Transformer-Sprachmodelle können auch explizit zur Generierung aussagekräftiger Embeddings trainiert werden (z. B. Reimers & Gurevych, 2019). Auf Basis dieser Embeddings können dann „klassische“ ML-Modelle wie beispielsweise logistische Regressionsmodelle, Decision Trees bzw. Entscheidungsbäume, Random Forests oder Support Vector Machines (SVM) trainiert werden (z. B. Géron, 2019; Rao & McMahan, 2019)[66]. Der Vorteil bei diesem Vorgehen ist, dass man die verfügbaren Texte nur ein einziges Mal mithilfe des Sprachmodells in Embeddings umwandeln muss. Das weitere Training ist dann je nach verwendetem ML-Modell deutlich schneller und effizienter als beim vollständigen Finetuning eines Sprachmodells.

Hier wurden zunächst Embeddings auf Basis dreier unterschiedlicher Sprachmodelle verwendet:

1. SBERT: Das Sentence-BERT Modell (Reimers & Gurevych, 2019) ist im Prinzip ein klassisches BERT-Modell. Der Unterschied liegt im Pretraining: Das SBERT-Modell wurde explizit so trainiert, dass ähnliche Texte ähnliche Embeddings ergeben. Die Embeddings sind 384-dimensional und das Modell hat ca. 110 Mio. Parameter.
2. LLaMA 3.2-1B: Die LLaMA-Modellfamilie (Touvron et al., 2023a; Touvron et al., 2023b) ist eine von Meta veröffentlichte und open-source-verfügbare Familie von Sprach- bzw. multimodalen[67] Modellen. Hier wurde das LLaMA 3.2-1B-Modell (Meta, 2024) mit einer Größe von ca. 1 Mrd. Parametern unter Nutzung der Open-Source-Software Ollama[68] (Ollama, 2024) verwendet. Die Embeddings sind 2048-dimensional.

---

[65] Abbildung A4 zeigt aber, dass die Unterschiede zwischen einer 3-fold- und einer 10-fold-CV eher in der Größenordnung von 1 bis 2 % bezüglich der Accuracy liegen, also überschaubar sind.

[66] Für detailliertere Informationen über die unterschiedlichen genutzten ML-Modelle sei hier auf entsprechende Literatur (z. B. Géron, 2019) verwiesen.

[67] Aufgrund des großen Erfolges von Transformer Modellen nicht nur bei der Text- sondern beispielsweise auch bei der Bildverarbeitung (z. B. Dosovitskiy et al., 2021; Esser et al., 2021) sind große Transformermodelle mittlerweile oft Multimodal und können insbesondere Text-, Ton- und Bilddaten verarbeiten.

[68] Ollama kann genutzt werden, um einen in den Programmiersprachen Go und C implementierten lokalen Server zu hosten, der eine Auswahl an Open-Source Modellen betreiben kann. Um die Modelle zu nutzen, können dann einfach Anfragen an diesen Server gesendet werden. Dadurch wird die hohe Performanz von Go und C

3. TE3s[49]: Das Modell „text-embedding-3-small“ ist die kleinere Version des aktuellen (Januar 2025) Embedding-Modells von OpenAI (o. D.-b)[69]. Das Modell kann über die OpenAI-API kostengünstig genutzt werden und liefert 1536-dimensionale Embeddings.

Im ersten Schritt werden unabhängig vom später genutzten ML-Modell die Embeddings generiert. Für den hier verwendeten Datensatz nimmt dies bei allen drei genutzten Embedding-Modellen mit den Implementierungen, die im digitalen Begleitmaterial hinterlegt sind, lediglich einige Minuten in Anspruch[70].

Um die Embedding-Modelle zu evaluieren, wurde zunächst einheitlich ein logistisches Regressionsmodell im Rahmen einer 10-fold-CV für das Scoring genutzt. Abbildung 6.16 stellt die Übereinstimmungswerte im automatisierten Scoring dar. Als Vergleichswert ist dort zudem die Performanz des finegetuneten BERT-Modells aus Artikel 3 dargestellt. Die Embedding-Modelle kombiniert mit logistischer Regression bleiben deutlich hinter dem finegetuneten Modell zurück. Insbesondere das Modell auf Basis der LLaMA-Embeddings zeigt eine niedrige Performanz. Die LLaMA-Embeddings werden daher aus Gründen der Übersichtlichkeit im Folgenden nicht mehr betrachtet.

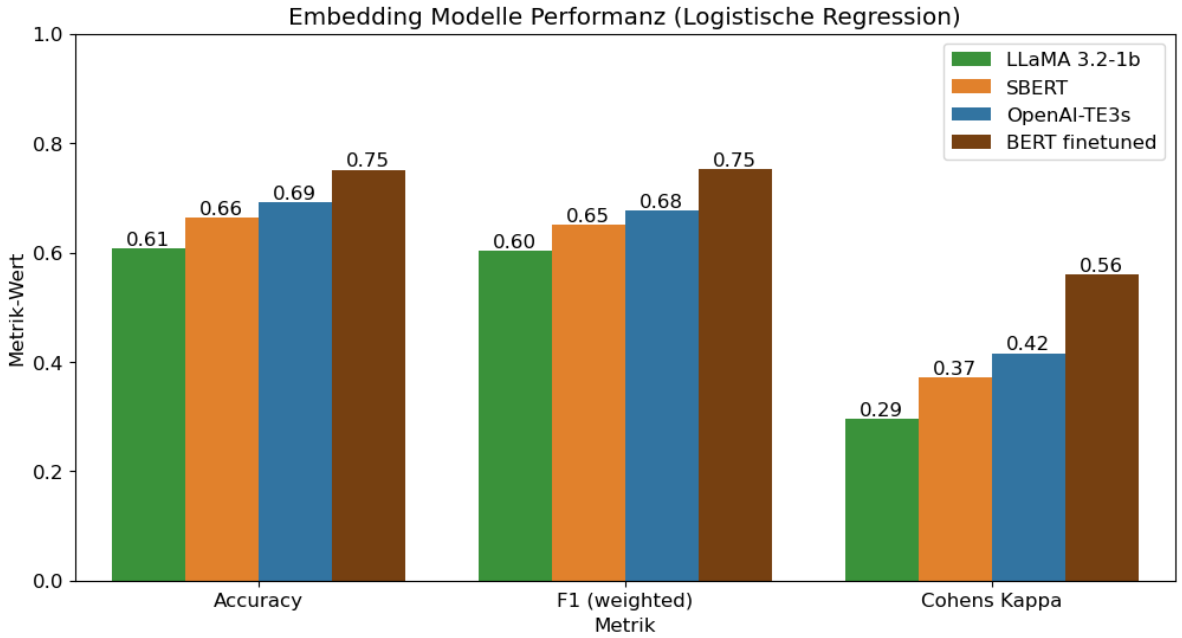


Abbildung 6.16 Vergleich der Performanz des automatisierten Scorings auf Basis dreier Embedding-Modelle mit logistischer Regression.

genutzt, aber ein einfaches Interface beispielsweise in Python oder Javascript bereitgestellt. Weitere Informationen unter https://github.com/ollama.

[69] OpenAI sind dabei weder bezüglich der zugrundeliegenden GPT-Version (z. B. GPT3.5 oder GPT4o) noch der Parameteranzahl des Modells transparent. Gemessen an der mit über 8000 Token recht großen Kontextlänge ist hier allerdings mit einigen 100 Mio. bis einigen Mrd. Parametern zu rechnen.

[70] Für die Analysen dieser Arbeit wurde ein Computer mit einem AMD Ryzen 3700X, 32 Gigabyte Arbeitsspeicher und einer Nvidia RTX 2080-GPU Super mit 8 Gigabyte Grafikspeicher genutzt.

6.7.7 Auswirkung von Vorverarbeitungsschritten und Modellwahl auf die Performanz des Assessment-Systems

Eine Möglichkeit, der ungleichmäßigen Verteilung der Score-Labels (Tabelle 6.4) zu begegnen ist das sog. Oversampling (Lemaître et al., 2017). Dabei werden die Datenpunkte mit den selteneren Labels randomisiert öfter im Training verwendet, sodass insgesamt eine gleichmäßige Abdeckung der Label-Verteilung im Training erzeugt wird. In Abbildung 6.17 sind für die SBERT- und OpenAI-Embeddings sowie das finegetunete BERT-Modell die Auswirkungen von Oversampling auf die Performanz dargestellt. Man erkennt insbesondere die positive Auswirkung auf Cohens κ, was für das automatisierte Scoring (insbesondere bei ungleichmäßiger Score-Verteilung) die wichtigste Metrik darstellt. Dabei sei erwähnt, dass die Evaluierungsdaten nicht dem Oversampling unterzogen werden (dürfen). Für das finegetunete BERT-Modell ist eine derartige positive Auswirkung des Oversamplings nicht zu beobachten (Abbildung 6.17, grau / schwarz). Beim Finetuning lohnt es sich hier also nicht, die (durch die hohe Schiefe der Score-Label-Verteilung deutlich) erhöhte Trainingsdauer in Kauf zu nehmen.

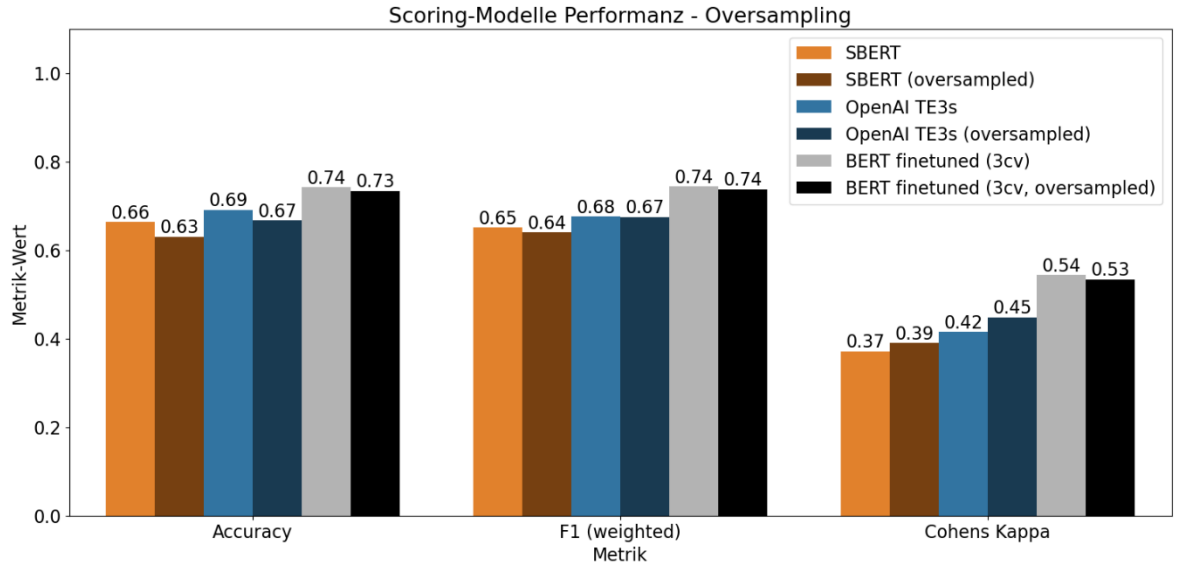


Abbildung 6.17 Auswirkung von Oversampling auf die Scoring-Modelle. Die Embedding-basierten Modelle (SBERT und OpenAI TE3s) bestehen aus den Embeddings gefolgt von einem logistischen Regressionsmodell und wurden mithilfe einer 10-fold-CV evaluiert. Das finegetunete BERT-Modell ist analog zu Artikel 3 trainiert, wurde allerdings hier zur Betrachtung der Auswirkung des Oversamplings nur im Rahmen einer 3-fold-CV evaluiert.

Analog zum Effekt des Oversamplings kann man auch den Effekt, den das Hinzufügen der Aufgabennamen in der Form „Aufgabe X: …“ zu den Antworttexten hat, quantifizieren. Dazu wurden die verwendeten Embeddings der Antworttexte einmal mit und einmal ohne die Aufgabennamen erstellt. Im Falle des BERT-Modells aus Artikel 3 kann direkt mit oder ohne hinzufügen der Aufgabennamen gearbeitet werden. Alle bisher berichteten Modelle sind *mit* Aufgabennamen trainiert worden, daher wird hier in Abbildung 6.18 explizit darauf hingewiesen, welche Modelle *ohne* Aufgabennamen trainiert wurden. Man erkennt einen leicht positiven Effekt der Nutzung der Aufgabennamen auf die Performanz, insbesondere auch für das finegetunete BERT-Modell. Hier kann zudem noch die Anzahl an „unmöglichen“

Punktzahlen betrachtet werden: Wie bereits in Abschnitt 6.7.3 erwähnt, können die Modelle grundsätzlich Punktzahlen vorhersagen, die das Testinstrument eigentlich nicht vorsieht, da nicht alle Aufgaben des Testinstruments die gesamte Spanne von 0 bis 2 Punkten abdecken. Für das SBERT-Modell sinkt die Anzahl an illegitimen Scores durch das Hinzufügen der Aufgabennamen von 49 auf 3, für das OpenAI-TE3s-Modell von 9 auf 0 und für das finegetunete BERT-Modell von 22 auf 0. Auch wenn dies im Vergleich zu den insgesamt 15.600 Evaluierungsvorhersagen kleine Werte sind, kann die Fehlerquote hier doch stark verringert werden, ohne, dass ein echter Mehraufwand in Kauf genommen werden muss.

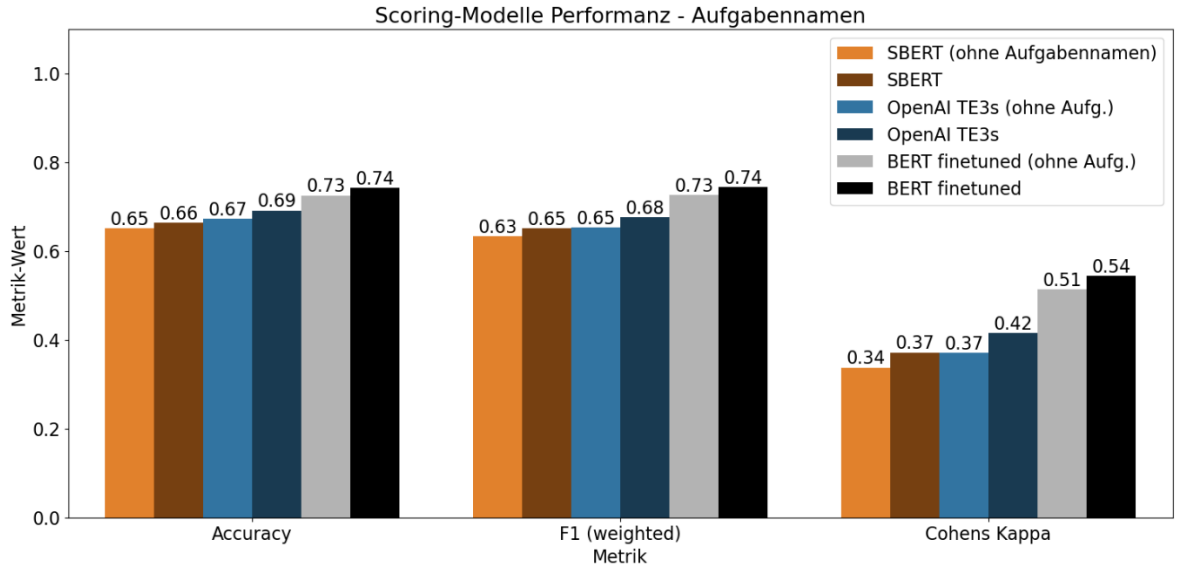


Abbildung 6.18 Auswirkung des Hinzufügens der Aufgabennamen zu den Aufgabentexten beim automatisierten Scoring.

Neben dem logistischen Regressionsmodell wurde für die Embedding-basierten Scoring-Modelle auch mit weiteren ML-Modellen experimentiert[71]. Einen nennenswerten Performanzzuwachs brachte dabei die Nutzung von SVMs: Eine SVM auf Basis der OpenAI-Embeddings erreicht hier ebenso hohe Übereinstimmungswerte, wie das finegetunete BERT-Modell aus Artikel 3 (Abbildung 6.19, links). Das Oversampling ermöglich dabei zudem eine deutlich bessere Vorhersage im Falle von 2 erreichten Punkten (Abbildung 6.19, rechts). Es muss allerdings beachtet werden, dass sich bei einer SVM (zumindest bei der Nutzung der gängigen Python-Software Scikit-Learn, Pedregosa et al., 2011) die Dauer des Workflows auch wieder auf ca. 1/3 der Dauer des Finetuning-Workflows erhöht. Die ~100-fache Beschleunigung des Workflows bei der Nutzung von Embeddings mit logistischer Regression kann mit SVMs leider nicht erreicht werden.

Auch, wenn das Modell auf Basis der OpenAI-TE3s-Embeddings eine größere Performanz erreicht, ist das SBERT-Modell trotzdem eine interessante Alternative. Wie auch das finegetunete BERT-Modell ist auch das SBERT-Modell vergleichsweise klein und zudem

[71] Dabei wurden ein Fully-Connected-Neural-Network, ein Random Forest, logistische Regressionsmodelle unter der Nutzung unterschiedlicher Optimierungsalgorithmen sowie SVMs verwendet (siehe digitales Begleitmaterial). Für Informationen zu diesen Modellklassen sei beispielsweise auf Géron (2019) verwiesen.

open-source verfügbar. Es kann also ohne Probleme vollständig lokal genutzt werden, ohne dass man auf die Nutzung von APIs Dritter angewiesen ist. Die Nutzung der SBERT-Embeddings ist vor allem interessant, wenn die verfügbare Hardware für ein vollständiges Finetuning nicht leistungsstark genug ist, denn das Erstellen der Embeddings erfordert deutlich weniger Rechenleistung und Arbeitsspeicher als ein vollständiges Finetuning[72].

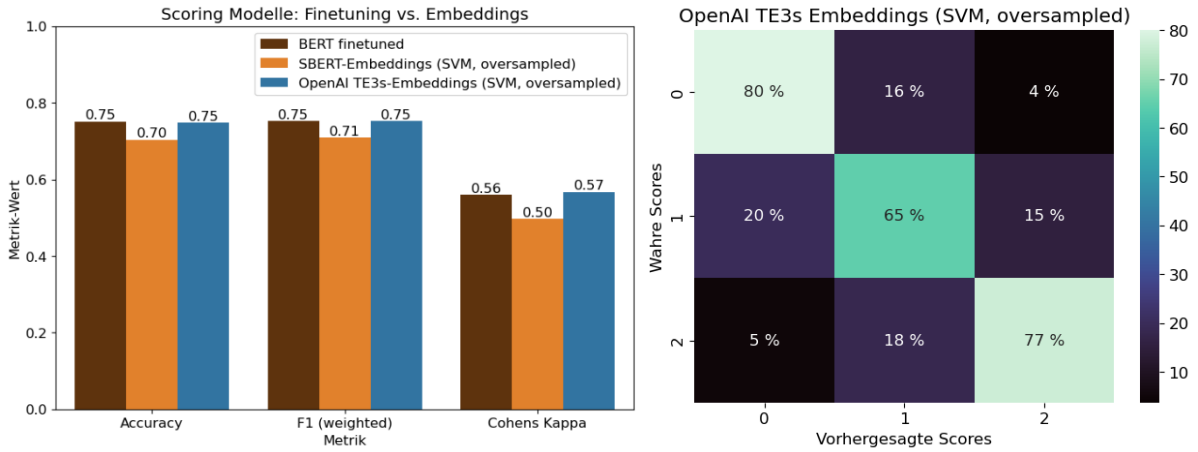


Abbildung 6.19 Performanz von SVMs auf Basis der Embedding-Modelle gegenüber der Performanz des finegetuneten BERT-Modell.

Auch die Score-Vorhersagen auf Basis der Embedding-Modelle können analog zum Vorgehen in Artikel 3 zur Vorhersage der Kompetenzprofile genutzt werden. Um auch hier die Vergleichbarkeit zur Mensch-Mensch-Übereinstimmung zu verbessern, wurde ebenfalls das „wahre“ GMM zur Zuordnung zu den Kompetenzprofilen auf Basis der Score-Vorhersagen genutzt. Abbildung 6.20 stellt die Performanz der unterschiedlichen Modelle bezüglich der Kompetenzprofilzuordnung dar. Zum Vergleich ist außerdem das BERT-Modell aus Artikel 3 enthalten. Darüber hinaus sind sowohl für die OpenAI-Embeddings als auch für die SBERT-Embeddings die Werte dargestellt, die man erhält, wenn ein logistisches Regressionsmodell zur Zuordnung der Cluster ausgehend von den Scores trainiert wird (Zusatz „Cluster-Fit“ in der Legende von Abbildung 6.20). Das SBERT-basierte Modell mit einer SVM für das Scoring und einem auf diesen Scores aufbauenden logistischen Regressionsmodells zur Vorhersage der Kompetenzprofile erreicht hierbei auffällig hohe Performanzwerte.

Analog zum Vorgehen in Abschnitt 6.7.4 kann auch für die anderen Scoring-Modelle die Performanz neben der Zuordnung zu den Kompetenzprofilen auch mit der Vorhersagegüte bezüglich der Subskalen-Summenscores evaluiert werden. Aus Platzgründen wird hier nicht für jedes Scoring-Modell eine ausführliche Betrachtung wie in Abschnitt 6.7.4 vorgestellt. Stattdessen werden hier die (Pearson-)Korrelationen zwischen den menschlichen und maschinellen Subskalen-Scores bezüglich der 10 Skalen (siehe Abschnitt 6.7.4) zusammengefasst als Boxplots dargestellt (Abbildung 6.21). Dabei wird sich ebenfalls analog zum vorherigen Vorgehen auf die schlichte Summierung der Score-Vorhersagen ohne ein

[72] Je nach genutztem Optimierungsverfahren sind die Speicheranforderungen beim Finetuning um einen Faktor 5 bis 10 gegenüber der sog. *Inference*, d. h. der Benutzung eines bestehenden Modells, erhöht.

zusätzliches Downstream-Modell bezogen, um eine Vergleichbarkeit zu den Mensch-Mensch-Werten herzustellen. Selbst das schlechteste Modell, d. h. das SBERT-Embedding-basierte Scoring-Modell mithilfe logistischer Regression, erreicht auch im schlechtesten Falle noch eine Korrelation von knapp 0,6. Die Modelle nähern sich in der Performanz der Mensch-Mensch-Übereinstimmung an. Um Subskalen-Scores vorherzusagen, scheinen also viele Modelle in Frage zu kommen, die teilweise innerhalb kurzer Zeit auch mit begrenzten Ressourcen trainiert bzw. evaluiert werden können.

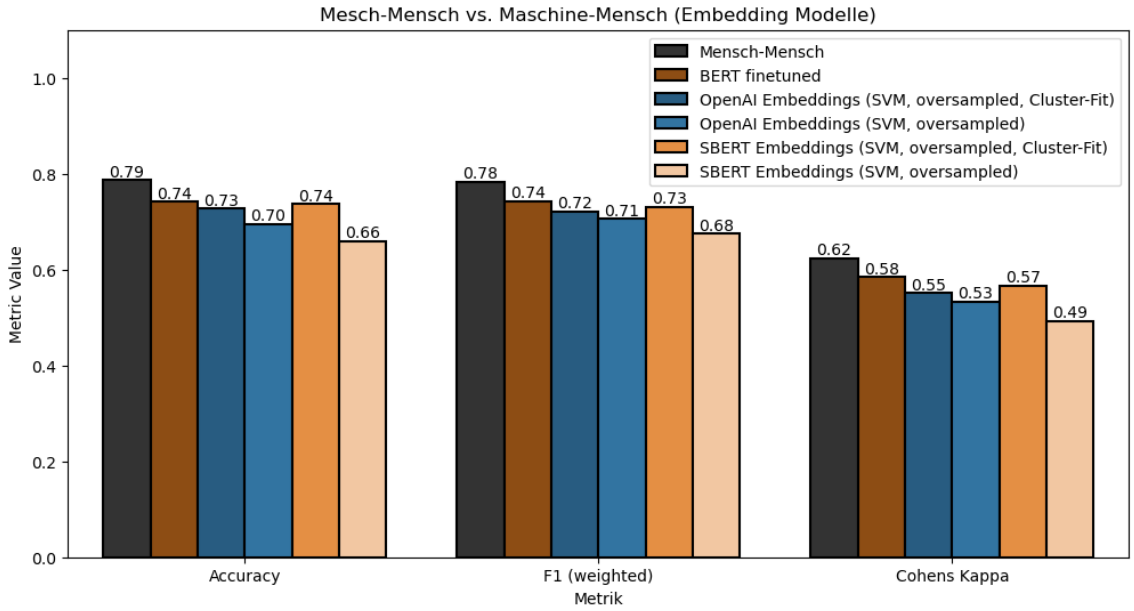


Abbildung 6.20 Kompetenzprofil-Vorhersagen aus Basis der Embedding-basierten Scoring-Modelle.

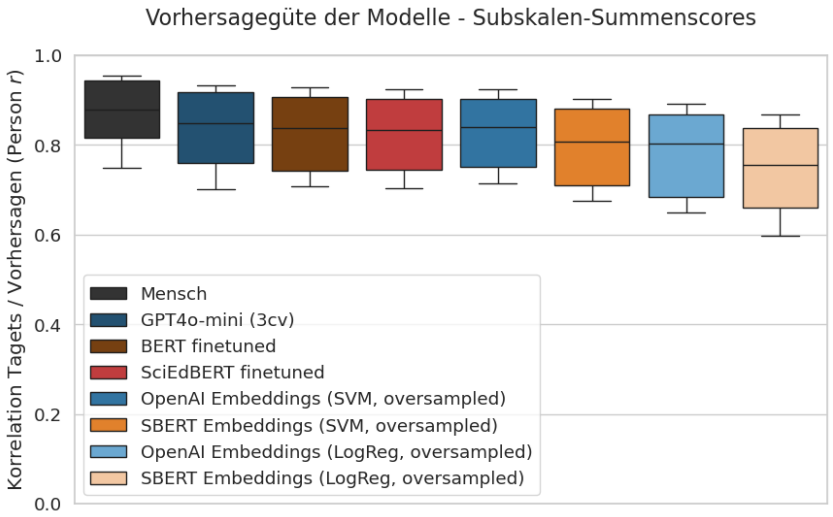


Abbildung 6.21 Korrelationen zwischen maschinellen und menschlichen Subskalen-Scores – unterschiedliche Modelle. In die einzelnen Boxplots gehen die Korrelationswerte bezüglich der 10 Skalen *Gesamtscore*, *Instruktionsstrategien*, *Schülervorstellungen*, *Experimente*, *Fachdidaktische Konzepte*, *Reproduzieren*, *Anwenden*, *Analysieren*, *Evaluieren* und *Kreieren* ein. Es werden hier auch bereits die Werte der in Abschnitt 6.7.8 betrachteten finegetuneten Modelle (GPT4o-mini & SciEdBERT) dargestellt. Das GPT4o-mini Modell wurde im Rahmen einer 3-fold-CV evaluiert, alle anderen Modelle im Rahmen einer 10-fold-CV. Das SciEdBERT[73]-Modell wird in Abschnitt 6.7.8 eingeführt.

Zuletzt sei hier noch kurz angemerkt, dass auch auf Basis von Embeddings versucht werden kann, die Kompetenzprofile direkt, d. h. ohne vorheriges Scoring der Aufgaben, vorherzusagen. Die dabei erreichten Übereinstimmungswerte sind jedoch ähnlich unzufriedenstellend, wie beim analogen Ansatz mit finegetuneten Modellen (Abschnitt 6.7.2). Lediglich die Gleichmäßigkeit der Falsch-Klassifikationen kann durch Oversampling etwas erhöht werden (siehe digitales Begleitmaterial).

6.7.8 Finegetunete Scoring-Modelle und ChatGPT als Scorer

Embedding-basierte Modelle stellen eine praktische Alternative für das automatisierte Scoring dar. Um einige Randbemerkungen aus Artikel 3 noch empirisch zu untermauern und darüber hinaus Vergleiche zu anderen bestehenden Ansätzen durchführen zu können, werden hier abschließend noch Informationen zu weiteren explorierten Modellen ergänzt, die wie das BERT-Modell in Artikel 3 vollständig finegetuned wurden.

In Artikel 3 wird erwähnt, dass sich weiteres Training über die dritte Epoch hinaus für das BERT-Modell nicht lohnt. In Abbildung 6.22 wird dies noch einmal verdeutlicht: Zusätzliches Training führt lediglich zu Overfitting, welches durch den steigenden Evaluierungsloss sichtbar wird. Die Performanz bezüglich der nicht-kontinuierlichen Metriken (Abbildung 6.22, rechts) bleibt allerdings fast identisch.

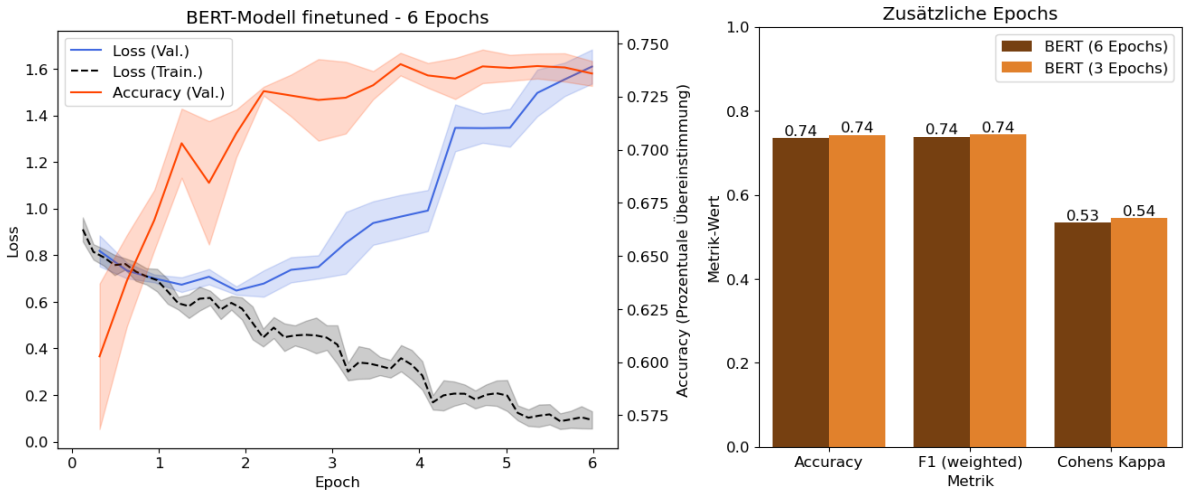


Abbildung 6.22 Auswirkung von zusätzlichen Trainingsepochs auf die Performanz des BERT-Scoring-Modells.

Als Alternative für das verwendete BERT-Modell wird in Artikel 3 das SciEdBERT[73]-Modell von Latif et al. (2024) erwähnt. Es ist begrüßenswert, dass in der deutschsprachigen naturwissenschaftsdidaktischen Forschung Ansätze verfolgt werden, flexibel einsetzbare Sprachmodelle zu entwickeln, die gegenüber den allgemein vortrainierten Modellen (wie BERT) ggf. einen Mehrwert bieten. Allerdings wurden für das domänenspezifische Training von SciEdBERT fachphysikalische Aufgaben auf Schulniveau aus den PISA-Studien von 2015 und 2018 verwendet, sodass hier nicht unbedingt zu erwarten ist, dass die Nutzung von SciEdBERT für die Bepunktung von fachdidaktischen Aufgaben auf universitärem Niveau

[73] **Sci**ence-**Ed**ucation-**BERT**

einen Vorteil hat. Dieser Verdacht bestätigte sich in einer 10-fold-CV, bei der in einem analogen Workflow zu Artikel 3 SciEdBERT eine Accuracy von 75,4 % und ein Cohens κ von 0,563 erreichte. Die Unterschiede zur Performanz des „klassischen" BERT-Modells aus Artikel 3 (75,1 %, $\kappa = 0{,}560$) sind also marginal. In Abbildung 6.21 sieht man zudem, dass SciEdBERT auch bezüglich der Vorhersage der Subskalen-Scores keine nennenswerten Vorteile bietet. In Abbildung 6.23 sind zudem die Performanzwerte aller finegetuneten Scoring-Modelle noch einmal im Rahmen einer 3-fold-CV dargestellt. Man erkennt, dass die durch die randomisierte Erstellung der CV-Splits entstehende statistische Schwankung der Performanzwerte den vermeintlichen minimalen Vorteil des SciEdBERT-Modells egalisiert (siehe auch Abbildung A4).

Neben dem Training von SciEdBERT stellen Latif et al. (2024) zudem einen Workflow vor, bei dem das Scoring-Modell neben dem gemeinsamen Training mit allen Aufgaben auch anschließend noch für jede einzelne Aufgabe finegetuned wird. Sie nutzen dabei einen Datensatz für das Finetuning eines BERT-Modells mithilfe von allen Aufgaben und einen zweiten Datensatz für das aufgabenspezifische Finetuning, sodass Data Leakage vermieden wird. Das somit zusätzlich finegetunete Modell erreicht dadurch im Durchschnitt aller Aufgaben einen Performanzzuwachs von ca. 10 %[74]. Da im hier vorgestellten Projekt kein zusätzlicher Datensatz für ein aufgabenweises Finetuning verfügbar ist, wird der bereits zuvor verwendete CV-Ansatz verfeinert. Der Evaluierungsworkflow lautet dann wie folgt:

1) Unterteile den Gesamtdatensatz randomisiert in k Segmente (CV-Splits).
2) Für jedes dieser Segmente:
   a) Trainiere ein BERT-Modell auf Basis aller Antwort-Score-Paare außer denen in diesem Segment (*Segment X*).
   b) *Optional:* Speichere die Vorhersagen des Modells aus *Schritt a)* bezüglich aller Antwort-Score Paare in *Segment X*.
   c) Für jede Aufgabe (*Aufgabe Y*) des Fragebogens (23 Aufgaben):
      i) Erstelle eine Kopie des Modells aus *Schritt a)* und trainiere es erneut bezüglich aller Antwort-Score-Paare der aktuellen *Aufgabe Y*, außer denen in *Segment X*. Dieses Modell wird *Modell X-Y* genannt.
      ii) Für alle Antwort-Score Paare zur aktuellen *Aufgabe Y* in *Segment X*: Speichere die Vorhersagen des *Modells X-Y* zu den Antworten für die spätere Evaluierung
3) Evaluiere die aufgabenweise finegetuneten Modelle mit den Vorhersagen aus *Schritt 2.c.ii)*.
4) *Optional:* Evaluiere das Aufgaben-agnostische Modell mit den Vorhersagen aus *Schritt 2.b)*.

Ohne die Schritte 2.c) und 3) ist dies der „normale" CV-Workflow, der auch in Artikel 3 verwendet wird. Die Schritte 2.c) und 3) ermöglichen eine Data-Leakage-freie Evaluierung der

[74] In den bisher (Stand Januar 2025) von Latif et al. (2024) zur Verfügung gestellten Reports und Code-Teilen wird der Workflow allerdings noch nicht ganz deutlich. Die Performanzzuwächse können beispielsweise auch dadurch entstehen, dass beim aufgabenweisen Finetuning noch einmal zusätzliche Daten verwendet werden, der Trainingsdatensatz sich also insgesamt vergrößert. Im ihren bisherigen Preprint bleiben Latif et al. (2024) hier recht vage.

aufgabenweise finegetuneten Modelle. Die Performanzwerte der aufgabenweisen finegetuneten Modelle sind in Abbildung 6.23 unter dem Label „BERT (aufgabenweise, 3cv)“ enthalten. Man erkennt nur kleine Zuwächse. Zusätzlich muss man nach einem aufgabenweisen Finetuning anstatt eines einzelnen Modells hier 23 Modelle nutzen. Anstatt ein einzelnes BERT-Modell von ca. 440 Megabyte in den Arbeitsspeicher des Servers zu laden, müssten dann insgesamt über 10 Gigabyte an BERT-Modellen geladen und wieder freigegeben werden, was einen großen Overhead an notwendiger Rechenleistung und -dauer erzeugt. Für die nur geringen Performanzzuwächse erscheint das nicht als lohnend[75].

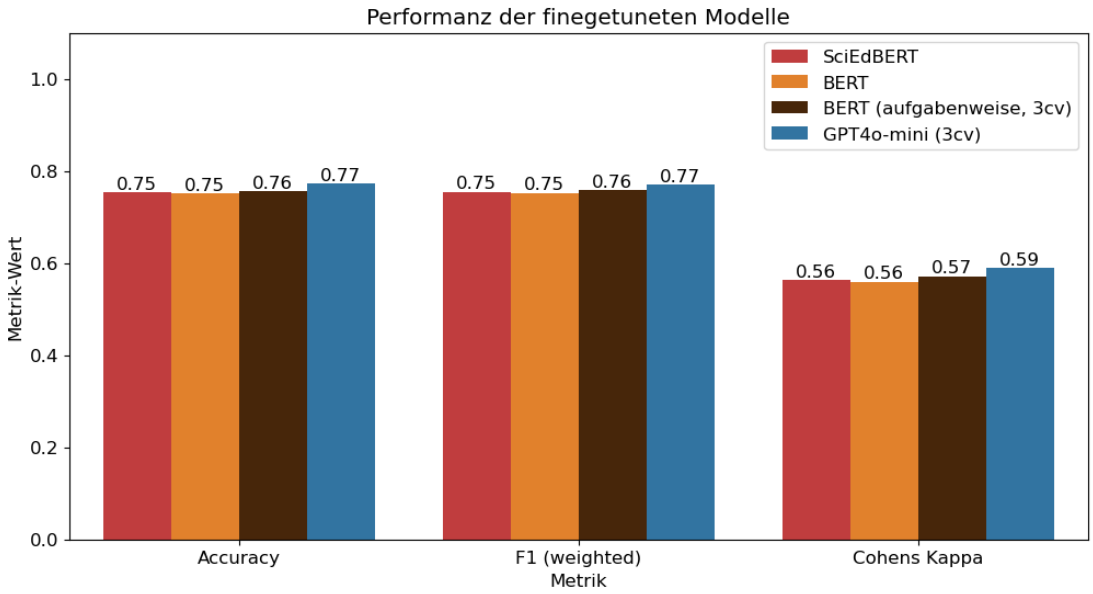


Abbildung 6.23 Performanz der explorierten Finetuning-Scoring-Modelle. Die Hinzunahme von GPT4o-mini wird unten noch diskutiert. Es ist zu beachten, dass SciEdBERT und BERT mit einer 10-fold-CV und das aufgabenweise BERT-Modell sowie GPT4o-mini lediglich mit einer 3-fold-CV evaluiert wurden.

Zuletzt wird am Beispiel von GPT4o-mini (OpenAI, 2022, 2024b) untersucht, ob die Nutzung bzw. das Finetuning von großen Sprachmodellen (LLMs) einen lohnenden Mehrwert gegenüber den anderen verwendeten Modellen bietet. Mittlerweile stellen die Anbieter von LLM-Tools wie ChatGPT vermehrt auch Funktionen zum Finetuning der Modelle für konkrete Anwendungszwecke bereit (z. B. OpenAI, o. D.-a)[49]. Latif und Zhai (2023) berichten in diesem Zusammenhang von Zuwächsen in der Performanz von bis zu 10 % bei der Nutzung eines finegetuneten GPT3.5-Modells gegenüber einem BERT-Modell für automatisiertes Scoring von Physikaufgaben auf Mittelstufenniveau. Inspiriert von diesen Ergebnissen wurde auch hier die OpenAI-API verwendet, um die aktuelle ChatGPT-Version (GPT4o-mini) zur Bepunktung der FDW-Aufgaben zu trainieren. Dies wurde im Rahmen einer 3-fold-CV evaluiert. Die

[75] In diesem Kontext könnte die Anwendung der sog. **Lo**w-**R**ank-**A**daptation Methode (LoRA, Hu et al., 2022) interessant sein. Dabei werden nur Teile eines Modells bei einem Finetuning verändert. Bei der erweiterten quantisierten LoRA-Methode (QLoRA, Dettmers et al., 2024) werden die Modellparameter zusätzlich in ein Speicherformat mit geringerer Speichernutzung überführt, sodass auch größere Sprachmodelle in den Blick genommen werden können. Da für den hier verfügbaren Datensatz aber sogar ein vollständiges Finetuning kaum Performanzzuwächse bringt, ist eine Evaluierung mit LoRA hier nicht zielführend, ggf. aber vielversprechend für inhaltlich verschiedene aber strukturell ähnlich Datensätze.

erhaltenen Performanzwerte bezüglich des automatisierten Scorings sind in Abbildung 6.23 enthalten. GPT4o-mini übertrifft das BERT-Modell dabei lediglich um ca. 3 % bzgl. der Accuracy und um ein Delta von 0,05 bzgl. Cohens κ. Dabei sei angemerkt, dass ein einzelner CV-Durchlauf beim hier vorliegenden Datensatz mit GPT4o-mini ca. 10 € an Kosten erzeugt und keine Zeitersparnis gegenüber dem Finetuning von BERT (bei Verfügbarkeit einer mittelstarken GPU[70]) erzielt wird. Abbildung 6.21 und Abbildung 6.24 zeigen entsprechende Zuwächse in den Übereinstimmungswerten auch für die auf dem Scoring basierende Vorhersage von Subskalen-Scores und Kompetenzprofilen.

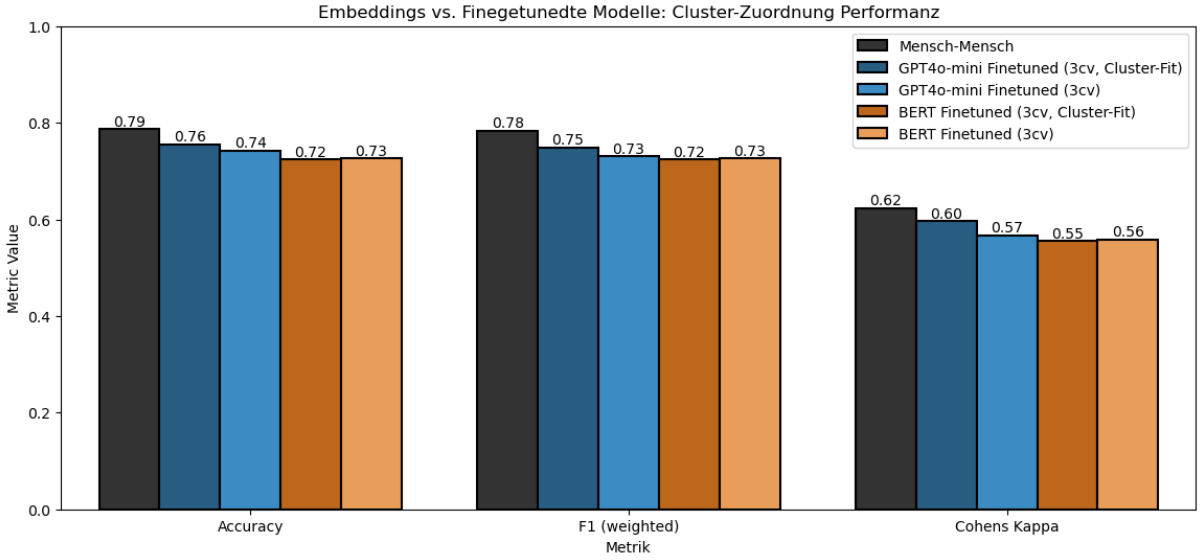


Abbildung 6.24 Cluster-Vorhersagen auf Basis von BERT und GPT4o-mini. Dargestellt sind sowohl die Vorhersageübereinstimmungen auf Basis des GMMs aus Artikel 3 zum direkten Vergleich mit der Mensch-Mensch-Baseline als auch auf Basis eines zusätzlichen logistischen Regressionsmodells zur Vorhersage der Cluster auf Basis der Scores („Cluster-Fit"-Zusatz in der Legende).

GPT4o-mini übertrifft die übrigen Modelle also leicht im automatisierten Scoring. Der Workflow für das Trainieren und Evaluieren des Modells ist aber ähnlich langwierig, wie das Finetuning von BERT und dabei deutlich kostspieliger. Wenn man trotzdem bei einer Implementation des Assessments auf GPT4o-mini setzen würde, wäre man zudem davon abhängig, dass OpenAI ihre API nicht verändert, sich also die Struktur der Anfragen und Antworten an den Server nicht verändert. Solche Änderungen würden entsprechende Anpassungen an den Code und Workflow eines Assessment-Tools notwendig machen.

Im Kontext von LLMs hat sich in den letzten Jahren ein Paradigmenwechsel vollzogen (Liu et al., 2023). Bei kleineren Sprachmodellen wie BERT besteht der typische Workflow aus einem Pre-Training mithilfe großer, allgemeiner Datensätze (z. B. Wikipedia-Texten, siehe Devlin et al., 2019) gefolgt von einem anwendungsspezifischen Finetuning, wie auch eingangs (Abschnitt 2.7) bereits beschrieben wurde. Große Sprachmodelle wie GPT3.5 oder GPT4o werden allerdings häufig auch auf eine andere Weise verwendet. Die Menge an Pre-Training-Daten und ihre Größe ermöglicht es solchen Modellen, in einem „Chat"-Setting teilweise ohne weiteres Finetuning bestimmte Aufgaben zu erfüllen. In diesem Setting ist dann die Entwicklung eines geeigneten *Prompts*, d. h. einer Aufforderung, die dem Modell übergeben

wird, zentral. Der Paradigmenwechsel vollzieht sich also von einem „Pretrain-Finetune-Predict-Workflow“ hin zu einem „Pretrain-Prompt-Predict-Workflow“ (Liu et al., 2023). Mit wachsender Kontextlänge der angebotenen Sprachmodelle bzw. ChatBots haben sich hierbei auch Methoden ausgebildet, bei denen größere Mengen an Informationen den Modellen mit dem Prompt mit übergeben werden. Beispielsweise hat GPT4o-mini eine Kontextlänge von 128.000 Token, was der Größenordnung der vorliegenden Arbeit entspricht[76]. Bei Prompting-Strategien wird unter anderem zwischen dem sog. „Zero-Shot-Prompting“ (Sanh et al., 2022; Wei et al., 2022) und „Few-Shot-Prompting“ (Brown et al., 2020) unterschieden. Beim Zero-Shot-Prompting wird der Prompt ohne Rückgriff auf den Datensatz erstellt, während beim Few-Shot-Prompting Beispieldaten genutzt werden. Ein Few-Shot-Prompt für das automatisierte Scoring einer Aufgabe könnte Beispiele für Antworten unterschiedlicher Punktzahlen aus dem tatsächlichen Datensatz enthalten.

Dai et al. (2023) berichten, dass ChatGPT (bei ihnen in der Version basierend auf GPT3.5) in einem Zero-Shot-Ansatz in der Lage sei, Feedback zu Data-Science-Projektbeschreibungen durch Studierende zu erstellen. Betrachtet man aber die begrenzte tatsächliche Übereinstimmung mit menschlichen Bewertungen (Dai et al., 2023, Tabelle 1) muss man zu dem Schluss kommen, dass das durch ChatGPT erstellte Feedback zwar offenbar augenscheinlich als valide aufgefasst wurde, für ein tatsächliches automatisiertes Assessment hier aber keine ausreichende Übereinstimmung erreicht wird, um menschliche Rater zu ersetzen.

Für den hier vorliegenden Datensatz liefert ein einfacher Zero-Shot-Prompt in der Form

```
Classify the following German response to the questionnaire-task ("Aufgabe")
on teacher knowledge into one of the three score-levels:
  <scores>
  [0, 1, 2]
  </scores>
Here is the German response:
  <responses>
  {response (hier wird die Antwort eingefügt)}
  </responses>
Respond using this format:
  <score>
  The score goes here as an integer.
  </score>
```

mit ChatGPT (GPT4o-mini[49]) keine Übereinstimmung mit dem menschlichen Bepunktungen (Cohens $\kappa = 0{,}054$), was nicht überrascht, da das Modell, wie eigentlich auch bei Dai et al.

[76] Die großen Kontextlängen haben zur Ausbildung einer Vielzahl an teilweise komplexen Workflows geführt, die auch unter dem Oberbegriff Retrieval-Augmented Generation (Gao et al., 2024) zusammengefasst werden. Gemein ist diesen Workflows, dass die Sprachgenerierung des Modells bzw. ChatBots durch zusätzlich herangezogene Quellen („Retrieval“) verbessert bzw. erweitert wird.

(2023), keine Anhaltspunkte zur Vergabe der Scores hat. In einem weiteren Experiment wurde daher stattdessen ein Prompt aus dem Erwartungshorizont des Testinstruments (Gramzow, 2015) exemplarisch zu zwei Aufgaben erstellt. Dazu wurden die beiden Aufgaben 1a) und 3) ausgewählt, da sie häufig bearbeitet wurden und zwei unterschiedliche Aufgabentypen darstellen: Aufgabe 1a) erfordert die Analyse einer beschriebenen Unterrichtssituation und Aufgabe 3) erfordert die Reproduktion von fachdidaktischem Wissen. Die Prompts bzw. Prompt-Templates, in die die Antworten der Proband:innen im Workflow an entsprechender Stelle automatisiert eingefügt wurden, sind in Anhang F enthalten. Die erhaltenen Übereinstimmungswerte zu den menschlichen Bepunktungen sind in Abbildung 6.25 den entsprechenden Werten des BERT-Scoring-Modells aus Artikel 3 gegenübergestellt. Für Aufgabe 1a) erreicht das GPT4o-mini-Modell mit dem Prompt in Anhang F deutlich höhere Übereinstimmungswerte, während es für Aufgabe 3) leicht hinter der Performanz des BERT-Modells zurückbleibt. Eine Evaluierung für alle Aufgaben liegt außerhalb der Zielsetzung dieses Projekts; diese Ergebnisse deuten aber darauf hin, dass es sich, insbesondere bei modernen Modellen wie GPT4, lohnen kann, für ein automatisiertes Scoring Zero-Shot-Prompting auf Basis von Kodiermanualen bzw. Erwartungshorizonten in den Blick zu nehmen. Die Nutzung von solchen Zero-Shot-Ansätzen ist besonders interessant, wenn nicht genügend Daten für das Finetuning eines Modells vorliegen. Weitere Evaluierungen von Zero-Shot-Prompting-Methoden auch auf Basis anderer Aufgaben und Testinstrumente werden hier nachdrücklich empfohlen.

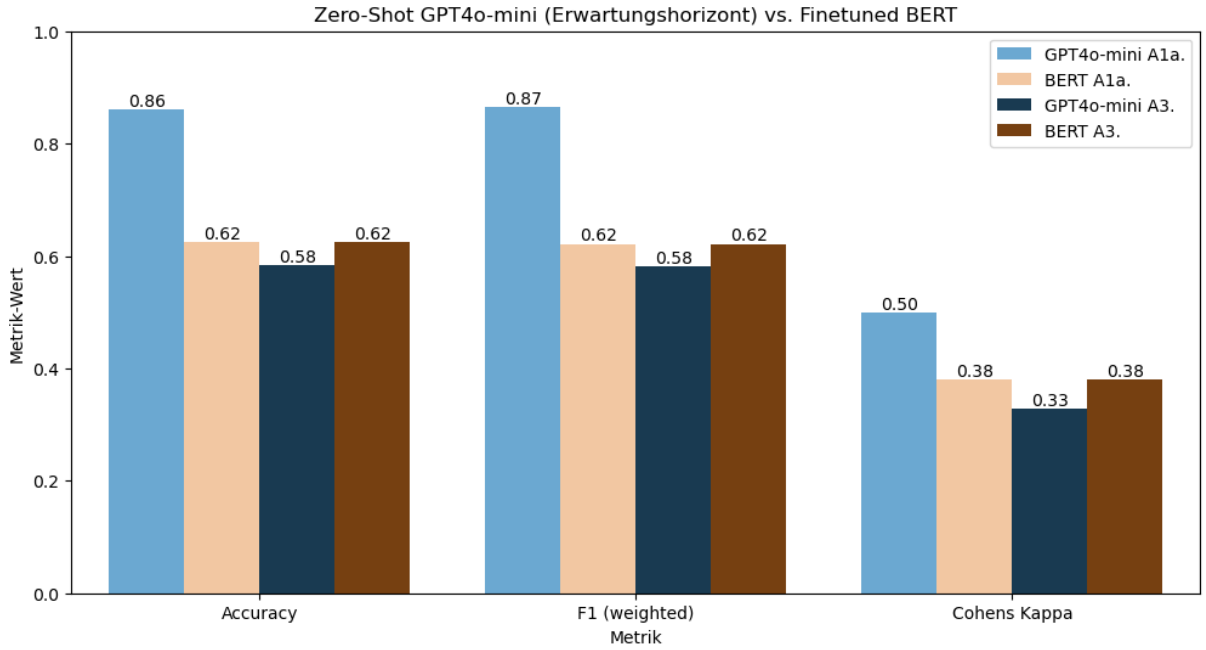


Abbildung 6.25 Zero-Shot Performanz von GPT4o-mini beim Scoring der Aufgaben 1a) und 3).

7. Zusammenfassende Diskussion

Im hier vorgestellten kumulativen Dissertationsprojekt wurde das fachdidaktische Wissen (FDW) von (angehenden) Physiklehrkräften einer detaillierten Untersuchung unterzogen. Aufbauend auf theoretischen (Baumert & Kunter, 2006; Blömeke et al., 2015; Gramzow, 2015; Hume et al., 2019; Riese, 2009) und empirischen sowie methodischen (Schiering et al., 2023; Woitkowski & Riese, 2017; Zeller et al., 2022) Grundlagen und Vorarbeiten wurden dazu zunächst datengetriebene exploratorische Analysen durchgeführt, um die innere Struktur des FDW detaillierter als bislang zu beschreiben. Dabei konnten empirisch basiert sowohl hierarchische Kompetenzniveaus als auch nicht-hierarchische Kompetenzprofile insbesondere mit Fokus auf kognitiven Anforderungen (Anderson & Krathwohl, 2001) identifiziert und beschrieben werden. Anschließend wurden sowohl bestehende Subskalen als auch die identifizierten Kompetenzprofile genutzt, um ein vollständig automatisiertes FDW-Assessment-System auf Basis von Machine Learning (ML) und Natural Language Processing (NLP) zu entwickeln und zu evaluieren. Hierzu wurden unterschiedliche Möglichkeiten und Workflows vorgestellt, von denen mehrere eine (verglichen mit der Interrater-Übereinstimmung) gute Performanz erreichten und somit für einen realen Einsatz zum Assessment in Frage kommen.

Im Folgenden werden nun noch einmal zusammenfassend die Ergebnisse und Beiträge der einzelnen Artikel des Projekts in den Kontext der jeweiligen Zielformulierung (Abschnitt 3.1) eingeordnet und mit einem Fokus auf der potenziellen Generalisierbarkeit der Ergebnisse sowie entsprechender Limitationen diskutiert (Abschnitt 7.1). Anschließend werden die Beiträge der Analysen zusammenfassend aus theoretischer und methodischer Sicht in den Forschungsstand eingeordnet (Abschnitt 7.2). Zuletzt werden im Ausblick (Abschnitt 7.3) offene Enden und Anknüpfungspunkte für mögliche Folgeprojekte aufgezeigt und die Beiträge des Projekts noch einmal übersichtsartig dargestellt (Abschnitt 7.4).

7.1. Beiträge und Limitationen der einzelnen Artikel

In Artikel 1 wurden mithilfe des Scale-Anchoring-Verfahrens (Mullis & Fishbein, 2020) auf Basis von item-response-theoretischen (IRT) Modellen Niveaustufen des FDW für zwei unterschiedliche Projekte und FDW-Testinstrumente identifiziert. Dabei zeigten sich projektübergreifende Parallelen bezüglich des Auftretens von Operatoren, die im Kontext lernpsychologischer Taxonomien (z. B. Anderson & Krathwohl, 2001) interpretierbar sind: In niedrigen Niveaus ist FDW auf reproduktive Aspekte beschränkt, während in höheren Niveaus bewertende und kreative Elemente hinzukommen. Bezüglich der Forschungsfrage 1.1 (FF1.1) konnten also projektunabhängige Niveaus gefunden werden, die allerdings lediglich recht grob gefasst werden können. Da sich die verwendeten Testinstrumente hinsichtlich fachlicher und fachdidaktischer Inhalte unterscheiden, konnten hier keine Aussagen über mögliche Ähnlichkeiten von Niveaustufen bezüglich dieser Dimensionen getroffen werden. Das ist allerdings nicht unerwartet, da allgemein angenommen wird, dass FDW vom betrachteten Fachinhalt abhängig ist (z. B. Hume et al., 2019) und zudem in einzelnen Testinstrumenten meist lediglich eine Auswahl möglicher fachdidaktischer Inhalte bzw. Facetten betrachtet wird

(z. B. Schmelzing, 2010). Ein größerer Detailgrad der Niveaubeschreibungen würde also immer mit einer Einschränkung der Generalisierbarkeit einhergehen.

Mit einem regressionsanalytischen Ansatz (FF1.2) konnten in Artikel 1 trotz des Versuchs der Erstellung eines projektunabhängigen Modells hierarchischer Komplexität (Commons et al., 1998) zur Beschreibung schwierigkeitserzeugender Merkmale des FDW keine projektübergreifenden Strukturen gefunden werden. Das für das FDW adaptierte Modell hierarchischer Komplexität hat sich dabei durch eine vermeintlich zu große Nähe zum FDW-Testinstrument des ProfiLe-P(+)[2]-Projekts als limitiert erwiesen und lies sich nicht gewinnbringend auf das Testinstrument des KiL[25]/KeiLa[26]-Projekts übertragen. Vor dem Hintergrund der Ergebnisse zum zweiten Zielpaket (nicht-hierarchische Strukturen) kann allerdings auch grundsätzlich angezweifelt werden, ob die innere Struktur des FDW überhaupt mithilfe eines derart strikt hierarchischen Ansatzes beschrieben werden kann bzw. sollte. Es besteht die Möglichkeit, dass das adaptierte Modell hierarchischer Komplexität eher die Aufgabenstruktur des ProfiLe-P-Tests widerspiegelt, als dass es tatsächlich die Qualität des FDW (beispielsweise gemessen über den Grad der Vernetztheit, z. B. Schnotz, 1994) abbildet. Interessant wäre hier der Versuch der Übertragung des Modells hierarchischer Komplexität auf andere Testinstrumente, die ggf. eine ähnlichere Aufgabenstruktur zum ProfiLe-P-Test aufweisen. Auch eine weitere Nutzung des hier entwickelten Modells hierarchischer Komplexität im Rahmen der Entwicklung zukünftiger Testinstrumente erscheint nützlich.

Die Scale-Anchoring-Analysen sind zunächst methodisch dadurch limitiert, dass keine Möglichkeit für eine direkte Verknüpfung der beiden verwendeten Datensätze (Anker-Items oder Linking-Stichprobe) vorlag. Daher wurden die entsprechenden Niveaumodelle zunächst getrennt voneinander entwickelt und anschließend durch einen inhaltlichen Vergleich zusammengebracht, anstatt ein Gesamtmodell auf Basis beider Datensätze zu erstellen. Das ProfiLe-P-Testinstrument ist zudem auf den Fachinhalt Mechanik beschränkt, während in KiL/KeiLa mehrere Fachinhalte abgedeckt wurden. Es ist somit auch im Kontext der Niveaumodelle noch nicht abschließend geklärt, inwieweit sich die gefundenen Strukturen auch auf andere Fachinhalte übertragen lassen, auch wenn die Nutzung der KiL/KeiLa-Daten hier eine Generalisierbarkeit vermuten lässt. Auch die eher grobe Beschreibung der Niveaus erhöht die Wahrscheinlichkeit ihrer Generalisierbarkeit, ist aber für die Bearbeitung des Desiderats der detaillierteren inhaltlichen Beschreibung des FDW problematisch. Obwohl das Scale-Anchoring-Verfahren relativ robust gegenüber Verschiebungen der manuell wählbaren Parameter des Workflows ist (Mullis & Fishbein, 2020), hat sich in den Arbeiten zu Artikel 1 gezeigt, dass der hohe Schwierigkeitsgrad des ProfiLe-P-Testinstruments eine Hürde für die Anwendbarkeit des Verfahrens ist: Nur unter Ausschluss der Personen, die das Testinstrument nur in Teilen bearbeitet haben, war eine sinnvolle Niveaubildung möglich. Wenn man diese wahrscheinlichen Test-Abbrecher in die Analyse einschließt, wird die Personengruppe mit niedrigem Fähigkeitsparameter so groß, dass sich die Aufgabengruppen des Scale-Anchoring-Verfahrens (siehe Abschnitt 4.4.3) sehr weit bezüglich der Schwierigkeitsparameter nach oben verschieben und eine sinnvolle Niveaubeschreibung nicht mehr möglich ist. Um hier eine Vergleichbarkeit zu den anderen Analysen (Zielpaket 2) beizubehalten, wurden auch dort die wahrscheinlichen Test-Abbrecher aus den explorativen Analysen ausgeschlossen.

Insgesamt konnten in Artikel 1 entsprechend dem projektübergreifenden Ansatz zwar Kompetenzniveaubeschreibungen mit einem hohen Grad an Generalisierbarkeit ermittelt werden, diese gehen aber kaum über eher grobe, allgemeinpsychologische Beschreibungen hinaus. Es deutete sich dabei bereits an, dass das FDW bezüglich lernpsychologischer Operatoren nicht-hierarchische Strukturen aufweist. Insofern waren diese Ergebnisse hilfreich, um die nicht-hierarchischen Analysen des zweiten und dritten Zielpakets zu leiten: Um eine (projektübergreifende) Generalisierbarkeit der Ergebnisse weiterer Analysen wahrscheinlicher zu machen, lohnt sich demnach insbesondere der Fokus auf die kognitiven Anforderungen anstelle beispielsweise der fachdidaktischen Facetten. Die verwendete Methodik der Zusammenführung von Ergebnissen aus unterschiedlichen Projekten ohne eine gemeinsame Stichprobe oder ein gemeinsames Testinstrument über den strukturierten Vergleich der Scale-Anchoring-Niveaus kann zudem auch auf andere Forschungsinhalte übertragen werden.

Aufbauend auf den Ergebnissen zur Generalisierbarkeit von Aussagen auf Basis von kognitiven Anforderungen aus Artikel 1 wurden in Artikel 2 nicht-hierarchische Cluster-Analysen mit Fokus auf den Anforderungskategorien *Reproduzieren*, *Anwenden*, *Analysieren*, *Evaluieren* und *Kreieren* durchgeführt. Aufgrund der aus Gründen der Testökonomie begrenzten Anzahl an Aufgaben pro Anforderungskategorie (Gramzow, 2015), konnte hier nur der K-Means Algorithmus verwendet werden. Daher sollten die vier gefundenen Cluster bzw. Kompetenzprofile *Low Achievers*, *Applying Creatives*, *Analytic Evaluators* und *High Achievers* (FF2.1) eher als „fundiertere Leistungsquantile“ denn als tatsächlich latente Personengruppen aufgefasst werden. Um die Ergebnisse stärker empirisch zu untermauern wurde daher ein Workflow auf Basis der Computational Grounded Theory (CGT, Nelson, 2020) entwickelt und die Ergebnisse auf Basis der quantitativen Score-Daten zusätzlich mit den authentischen Sprachproduktionen in Beziehung gesetzt (FF2.2). Dabei zeigte ein Structural Topic Model (STM, Roberts et al., 2019) erwartungskonform, dass die Applying Creatives einen Fokus auf die Beschreibung und Begründung von Beispielen zum Einsatz in Unterrichtssituationen legen, während die Analytic Evaluators und insbesondere die High Achievers eher Schülervorstellungen thematisieren. Die Interpretation der Ergebnisse von STMs ist allerdings ein qualitativer Analyseprozess, weshalb hier die Objektivierbarkeit teilweise kritisiert wird (Chang et al., 2009). Um dieser Limitation zu begegnen, wurde (1) ein strukturierter Prozess zur Beschreibung der Topics (Abschnitt 5.4.3 & 5.5.2) durchgeführt und (2) alle Zwischenergebnisse transparent berichtet (siehe auch digitales Begleitmaterial).

Auch wenn die Kompetenzprofile der K-Means Analyse aus methodischen Limitationen heraus für sich genommen wenig generalisierbar sind, legen die erwartungskonformen und informativen Systematiken bzgl. des prototypischen Sprachgebrauchs der Personencluster eine Generalisierbarkeit der Ergebnisse über das konkrete Analysesetting hinaus nahe. Eine CGT-Pattern-Confirmation-Analyse, bei der die Vorhersagbarkeit der Kompetenzprofile auf Basis der Scores gezeigt wurde, unterstreicht zusätzlich die Robustheit der identifizierten Kompetenzprofile (FF2.4). Nichtsdestotrotz ist der Workflow zur Analyse der nicht-hierarchischen FDW-Strukturen komplex und mit vielen Design-Entscheidungen wie der Wahl der Cluster- und Topicanzahl, Datenvorverarbeitungsschritten etc. verbunden. Da das verwendete K-Means-Clustermodell zudem keine latenten Strukturen abbildet, kann das Ergebnis eher als eine von unterschiedlichen denkbaren validen Möglichkeiten aufgefasst

werden, nicht-hierarchische Strukturen des FDW zu beschreiben.

Der genutzte Daten-Mix aus einerseits Bepunktungen und andererseits den Textproduktionen zu offenen Testaufgaben ist ein prototypisches Setting (nicht nur) für die Kompetenzmessung in der Bildungsforschung. Die genutzte Methodik bzw. der genutzte Workflow mit der Cluster-Analyse als CGT-Pattern-Detection, der Sprachanalyse mithilfe von STMs als CGT-Pattern-Refinement und auch der Vorhersage der Cluster auf Basis der Scores als Pattern-Confirmation sind in diesem Setting für die Analyse von prototypischen Personengruppen auch auf andere Datensätze und Projekte generalisierbar und übertragbar. Der zur Analyse verwendete Programmcode ist daher einerseits im digitalen Begleitmaterial dieser Arbeit enthalten und andererseits in weiten Teilen im Rahmen eines Open-Source-Projekts angelegt[77]. Dadurch wird auch eine Überprüfung der Generalisierbarkeit vorbereitet, da ein analoger Workflow unter Nutzung des bestehenden Codes nun mit minimalem Aufwand auch auf andere FDW-Testinstrumente bzw. FDW-Datensätze angewendet werden kann.

Die Ergebnisse aus Artikel 2 sind in zweierlei Hinsicht limitiert. Erstens sind die Kompetenzprofile aus der K-Means-Analyse aus methodischen Gründen nur begrenzt als tatsächlich latente bzw. prototypische Personengruppen zu verstehen. Auch, wenn diese Problemstelle durch die Anwendung des CGT-Workflows etwas abgemildert wird, bleibt die grundsätzliche Einschränkung des Algorithmus bestehen. Zweitens ist die ML-basierte Vorhersage der Kompetenzprofil-Zugehörigkeit auf Basis der Scores zwar für die Pattern Confirmation geeignet (siehe z. B. Tschisgale et al., 2023), allerdings hat ein solches ML-Modell nur wenig praktische Relevanz, denn der hohe Aufwand der manuellen Bepunktung bleibt für ein etwaiges Assessment so trotzdem notwendig. Beiden Limitationen wird in Artikel 3 durch eine erweiterte Cluster-Analyse und einen erweiterten Pattern-Confirmation-Workflow begegnet.

In Artikel 3 wurde auf Basis der Beobachtung, dass Kompetenzen im Analysieren und Evaluieren bzw. Anwenden und Kreieren tendenziell zusammenhängen (Artikel 2), zunächst der Datenverarbeitungs-Workflow der Clusteranalyse angepasst. Anstelle der fünf kognitiven Anforderungskategorien aus Artikel 2 wurde sich nun auf die drei Kategorien *Reproduzieren*, *Anwenden-Kreieren* und *Analysieren-Evaluieren* fokussiert. Dadurch wurden die Abstufungen der einzelnen betrachteten Subskalen feiner („mehr Punkte pro Kategorie“), sodass nun eine latente Profilanalyse (LPA, Spurk et al., 2020) angewendet werden konnte (FF2.4). Diese bestätigte im Wesentlichen bezüglich der Score-Cluster die Beobachtungen aus Artikel 2, weshalb auch die Kompetenzprofil-Bezeichnungen beibehalten wurden. Ein zusätzliches Pattern-Refinement bezüglich dieser latenten Kompetenzprofile wurde erneut mit einem STM durchgeführt und konsolidierte die beobachteten Parallelen zwischen den Cluster-Ergebnissen aus Artikel 2 und 3 weiter (Abschnitt 6.7.4).

Die latenten Kompetenzprofile weisen aus methodischer Sicht eine gegenüber den K-Means-Clustern aus Artikel 2 deutlich erhöhte Generalisierbarkeit auf. Trotzdem sind auch hier die Aussagen bisher weiterhin auf das konkret verwendete Testinstrument und den zugehörigen Datensatz limitiert. Dabei sei hier erneut erwähnt, dass das verwendete Testinstrument

[77] https://github.com/JannisZeller/questionnaire-tools, siehe auch Anhang G.

(Gramzow, 2015) – wie auch das Projekt ProfiLe-P (Riese et al., 2015) als Ganzes – auf den Fachinhalt Mechanik fokussiert ist. Auch hier wäre weitere Forschung zur Reproduzierbarkeit mithilfe weiterer Testinstrumente und Datensätze wünschenswert[78]. Neben der Erweiterung der Cluster-Analyse selbst wurde in Artikel 3 zusätzlich vor allem der Pattern-Confirmation-Schritt der Analyse erweitert. Anstatt die Kompetenzprofile lediglich auf Basis der Scores vorherzusagen, wurde nun ein zweistufiger Workflow entwickelt, bei dem zunächst die Aufgaben automatisiert bepunktet werden (FF3.1) und anschließend eine Zuordnung zu den Kompetenzprofilen auf Basis der Score-Vorhersagen vorgenommen wird (FF3.2). Somit wird hier eine Zuordnung von Bearbeitungen des Tests zu den Kompetenzprofilen vollständig ohne manuellen Aufwand durch Kodierung o. Ä. vorgenommen, was im Sinne der CGT ein verstärktes Argument für die Robustheit und Generalisierbarkeit der Kompetenzprofile darstellt (Nelson, 2020). Zusätzlich wurde im Rahmen dieser Pattern-Confirmation der latenten Kompetenzprofile der zweistufige Workflow für die Cross-Validierung (CV) der verwendeten Modelle mit einem personenweisen CV-Splitting systematisiert. So kann das automatisierte Assessment System neben dem reinen Scoring auch anhand der Kompetenzprofil-Zuordnung und der Vorhersage von Subskalen-Scores evaluiert werden. Insbesondere dieser Workflow ist in einer Testinstrument-unabhängigen Version im digitalen Begleitmaterial in Form von entsprechendem Code enthalten.

Wie die explorativen Analysen aus Artikel 2 basieren auch die explorativen latenten Profilanalysen in Artikel 3 auf einem komplexen Workflow. Obwohl die latenten Kompetenzprofile aus methodischen Gründen (latentes Cluster-Modell und erweiterte Pattern Confirmation) eine höhere Robustheit und höheres Generalisierungspotenzial aufweisen, sind demnach auch hier weitere (ggf. konfirmatorische) Analysen zur möglichen Übertragbarkeit und praktischen Relevanz ratsam. Die automatisierte Auswertung als Pattern Confirmation deutet hier allerdings darauf hin, dass es sich bei den gefundenen Personengruppen zumindest um eine Systematik mit hoher Validität handelt, auch wenn sie keine Aussagen über die praktische Relevanz ermöglicht. Das automatische Assessment als Anwendungszweck hat sich als insbesondere verglichen mit der Mensch-Mensch-Baseline sehr performant erwiesen, allerdings sind vor allem für das Training entsprechender Modelle als auch für die spätere Nutzung von trainierten Modellen einige Hardwareanforderungen[70] zu erfüllen. Darüber hinaus sollten für einen realen Einsatz des Assessments die bisher eher nüchternen und teilweise eher quantitativen Aussagen, die durch die Modelle geliefert werden, noch in ein prosaisches Format transformiert werden, welches im Sinne eines formativen Feedbacks (Hattie & Timperley, 2007) auch Hinweise auf mögliche Verbesserungspotenziale gibt und nächste Schritte explizit macht.

Insgesamt kann man zusammenfassen, dass alle Zielpakete der Arbeit mit Erfolg bearbeitet wurden und die entsprechenden Fragestellungen als beantwortet aufgefasst werden können. Zu Beginn des Projekts waren auch längsschnittliche Betrachtungen geplant, diese sind allerdings zugunsten des ersten Zielpakets zurückgestellt worden. Erste explorative Betrachtungen

[78] Gerade hier wird die Weiternutzung des oben angesprochenen Analysecodes interessant: Wie im digitalen Begleitmaterial zu sehen ist, müssen für die Überführung der K-Means-Analyse in die LPA-Analyse im Wesentlichen nur einige wenige Zeilen Code verändert werden.

zeigten hier, dass der Datensatz – trotz seiner für den deutschsprachigen Raum im Kontext der Professionswissensforschung von Physiklehrkräften einzigartigen Größe – offenbar nicht umfangreich genug für aussagekräftige längsschnittliche Betrachtungen bezüglich der FDW-Kompetenzprofile ist (siehe Anhang D).

7.2. Beitrag des Dissertationsprojekts als Ganzes

Den Ausgangpunkt dieses Projekts stellt ein umfangreicher verfügbarer Datensatz zum Professionswissen von (angehenden) Physiklehrkräften dar. Ohne diese bereits vorhandene umfangreiche Datenbasis, wären die methodisch aufwändigen Analysen in diesem Dissertationsprojekt nicht möglich gewesen. Bezüglich des Fachwissens (FW) und Pädagogischen Wissens (PW) existieren bereits Ansätze zur detaillierteren Beschreibung der inneren Struktur (Kaiser et al., 2020; König, 2009; Woitkowski & Riese, 2017). Zum FDW liegen zudem hierarchische Niveaubeschreibungen einzelner Projekte isoliert voneinander vor (Schiering et al., 2023; Schiering et al., 2019; Zeller et al., 2022). Darüber hinaus wurde sich im Kontext des FDW in der Naturwissenschaftsdidaktik zuletzt eher auf die Untersuchung von handlungsnäheren Kompetenzen im Sinne eines *enacted* Pedacogical Content Knowledge (ePCK) im Rahmen des Refined Consensus Model (RCM, Carlson et al., 2019) fokussiert, wie beispielsweise dem Planen von Unterricht (Behling et al., 2022b; Schröder et al., 2020), dem Erklären physikalischer Phänomene (Kulgemeyer et al., 2020; Kulgemeyer & Tomczyszyn, 2015), dem Reflektieren über Unterricht (Kulgemeyer et al., 2021; Reimer & Tepner, 2022) oder konkretem Handeln im Klassenzimmer (Förtsch et al., 2016; Förtsch et al., 2018; She et al., 2024). Dabei stellt aber auch die empirisch basierte Beschreibung innerer Strukturen des FDW im Sinne eines *personal* PCK (pPCK) des RCM bzw. einer Disposition des Kontinuumsmodells (MoC, Blömeke et al., 2015) nach wie vor ein Forschungsdesiderat dar (Kaiser et al., 2020; Riese et al., 2017). Darüber hinaus gestaltet sich ein authentisches und valides Assessment des FDW insbesondere mit Aufgaben in offenem Antwortformat (Kulgemeyer et al., 2023) als sehr aufwändig, da bislang hierzu meist händische Kodierungen vorgenommen werden müssen (Gramzow, 2015; Kröger, 2019).

Im hier vorgestellten Projekt wurde diesen Desideraten entsprechend das FDW einer empirisch basierten, inhaltlichen Detailanalyse unterzogen. Dabei wurden zunächst IRT-Niveaumodelle in den Blick genommen und projektübergreifend betrachtet. Mithilfe des Scale-Anchoring-Verfahrens konnten hier generalisierbare aber eher allgemeine Aussagen über das FDW abgeleitet werden (s. o.). Basierend auf diesen Ergebnissen wurden anschließend auch nicht-hierarchische Analysen durchgeführt, sodass nun latente prototypische Kompetenzprofile des FDW beschrieben werden können, die ein hohes Maß an Validität und Robustheit aufweisen und für die die Wahrscheinlichkeit einer Reproduzierbarkeit im Kontext anderer FDW-Datensätze und Projekte entsprechend hoch ist. Diese Kompetenzprofile können zudem genutzt werden, um weitere Forschung zum FDW bzw. PCK mit einer größeren Auflösung von Personengruppen und Subskalen vorzunehmen. Besonders interessant erscheint hierbei die Untersuchung des Zusammenhangs zwischen den latenten (pPCK-)Kompetenzprofilen und den Komponenten des ePCK, die im Rahmen des Plan-Teach-Reflect-Cycles (PTR-Cylce, Alonzo et al., 2019) identifiziert bzw. charakterisiert werden. Die latenten Kompetenzprofile

zeigen in diesem Kontext, dass auch für das FDW bzw. pPCK eine Unterteilung unterschiedlicher Teilkompetenzen wie beim ePCK oder den handlungsnahen Aspekten des MoC sinnvoll sein kann. Die dargestellten Sprachanalysen auf Basis von Topic Modelling unterstreichen dies auch auf Basis konkreter verwendeter Begrifflichkeiten und fokussierter Themen. Insgesamt wird hier also eine systematische Möglichkeit zur Beschreibung von Teilkompetenzen des FDW angebahnt. Weitere potenziell konfirmatorische Analysen zur Trennbarkeit der durch die Kompetenzprofile suggerierten potenziellen Teilkompetenzen „FDW-Reproduzieren“, „FDW-Anwenden-Kreieren“ und „FDW-Analysieren-Evaluieren“ nach dem Vorbild des ePCK-plan, ePCK-teach und ePCK-reflect des PTR-Cycles wären hier ein nächster Anknüpfungspunkt. Auch Betrachtungen zum Zusammenhang dieser Teilkomponenten mit anderen Professionswissensdomänen (z. B. Fachwissen oder Pädagogisches Wissen) sowie mit handlungsnahen Kompetenzen (z. B. Planen von Unterricht) werden hierdurch ermöglicht und nahegelegt. Erste solche Analysen auf Basis basaler Gruppenvergleiche und Korrelationsbetrachtungen erweisen sich allerdings bisher nicht als informativ (siehe Abbildung A3).

Neben diesem inhaltlichen Beitrag wurden zudem Workflows entwickelt und erprobt, die mehr oder weniger nahtlos auch auf ähnliche strukturierte Datensätze anderer Projekte angewandt werden können. Dabei sind zunächst die explorativen IRT-Niveauanalysen mit einer projektübergreifenden Niveaubetrachtung auf Basis von lernpsychologischen Operatoren, sowie die Cluster-Analyse auf Basis von Subskalen des Testinstruments zu nennen. Auch die STM-Anwendung zur Ausschärfung der Beschreibung der identifizierten Personengruppen kann bei Bedarf auf die entsprechenden Sprachdaten angewandt werden. Insbesondere der zweistufige Workflow für die CGT-Pattern-Confirmation der explorativen Analysen, bestehend aus einem Scoring-Modell und darauf aufbauenden weiteren Downstream-Modellen zur Vorhersage von Gruppenzugehörigkeiten oder Subskalenscores, ist ggf. auch für andere Projekte von Interesse. Um die Nutzbarkeit dieser methodischen Beiträge für das Forschungsfeld zu erleichtern, wurden sämtliche Code-Elemente im Rahmen des digitalen Begleitmaterials dieser Arbeit festgehalten und insbesondere der zweistufige Workflow wurde bereits in einer Testinstrument-unabhängigen Weise in Code umgesetzt (und genutzt). Somit können vergleichbare Analysen mit geringem Aufwand durchgeführt werden.

Im Rahmen des zweistufigen Workflows zur Pattern Confirmation wurde zudem ein vollständig automatisiertes Assessment-System des FDW entwickelt, welches im Vergleich zur Mensch-Mensch-Übereinstimmung und auch absolut betrachtet gute bis sehr gute Übereinstimmungswerte zu den menschlichen Assessment-Ergebnissen wie (Summen-)Scores und Kompetenzprofilen aufweist. Im Sinne des DEFT-Frameworks (Kubsch et al., 2022, siehe auch Abschnitt 2.5) wurden dabei im zweiten Zielpaket des Projekts eher Grounded-High-Inference-Untersuchungen zur Ermittlung der Kompetenzprofile durchgeführt (Kapitel 5 & 6), während die Arbeiten zum dritten Zielpaket darauf aufbauend eher als Supervised Settings mit sowohl Low(er)-Inference- (Bepunktung des Testinstruments) als auch High-Inference-Elementen (FDW-Kompetenzprofile und FDW-Subskalen) eingeordnet werden können (Kapitel 6).

Orientiert an der Nutzung von BERT-Modellen (Devlin et al., 2019) in anderen

naturwissenschaftsdidaktischen Forschungsvorhaben (z. B. Mientus et al., 2021; Tschisgale et al., 2023; Wulff et al., 2023) wurde auch hier primär ein BERT-Modell verwendet, der Workflow ist aber nicht auf dieses beschränkt (siehe Abschnitt 6.7.6 & 6.7.8). So konnten zudem punktuelle Beiträge zum Forschungsstand über den Zusammenhang von Aufgabencharakteristika und automatischer „Scorebarkeit“ von Aufgaben (Abschnitt 6.7.5), zur Anwendbarkeit von rein Embedding-basierten Modellen für automatisiertes Assessment (Abschnitt 6.7.6 & 6.7.7), zur Auswirkung bestimmter Vorverarbeitungsschritte auf die Assessment-Modellperformanz (Abschnitt 6.7.7) sowie zur Nutzbarkeit von großen Sprachmodell-Tools wie ChatGPT (Abschnitt 6.7.8) geleistet werden. Um die hierzu berichteten Ergebnisse weiter zu konsolidieren, ist aber noch weitere Forschung über das verwendete Testinstrument bzw. den verwendeten Datensatz hinaus notwendig.

7.3. Ausblick

Wie bereits mehrfach beschrieben, ist der erste und wichtigste Anknüpfungspunkt für weitere Forschung die Evaluierung der Übertragbarkeit der Ergebnisse auf weitere FDW-Datensätze. Da die inhaltlichen Aussagen (Kompetenzniveaus und -profile) primär auf allgemeingültige kognitive Anforderungskategorien bezogen sind, ist man hierfür nicht auf das Fach Physik oder die Naturwissenschaften beschränkt. Auch eine Übertragbarkeitsbetrachtung für andere Fächer erscheint sinnvoll. Die für die explorativen Analysen entwickelten und erprobten Workflows können zudem leicht im Kontext anderer Subskalen wie fachdidaktischer Facetten angewendet werden (siehe auch Abschnitt 5.7.2). Auf der inhaltlichen Ebene wäre zudem die systematische Untersuchung des Zusammenhangs von Kompetenzprofilen oder Kompetenzausprägungen bzgl. der kognitiven Anforderungen mit anderen Professionswissensdomänen oder eher handlungsnahen Kompetenzen (im Sinne eines ePCK) interessant. Hier liegen aus dem Projekt ProfiLe-P+ auch bereits Daten vor (z. B. Kulgemeyer et al., 2020; Schröder et al., 2020), deren systematische Betrachtung hier aber nicht mehr zu den Projektzielen gehörte.

Um Analysen zur Übertragbarkeit etc. durchzuführen oder die Workflows für das automatisierte Assessment auch für andere Testinstrumente und Datensätze durchzuführen, kann der als Open-Source-Projekt angelegte Python-Code dieser Arbeit auch für andere Projekte interessant sein. In diesem Projekt ist daher auch eine Dokumentation (siehe Anhang G) enthalten, die bereits viele Erläuterungen und Beispiele enthält. Besonders ist an diesem Code vor allem, dass er Testinstrument-unabhängig gestaltet ist. Die Konfiguration des Testinstruments kann verändert werden, sodass die Analysen bei minimalen Code-Anpassungen auch für unterschiedliche Kombinationen von offenen und geschlossenen Aufgaben mit unterschiedlichen Bepunktungsverfahren, Maximalscores, Subskalen etc. durchgeführt werden können.

Darüber hinaus wären auf methodischer Seite auch weitere, detailliertere Betrachtungen der Interpretierbarkeit bzw. Erklärbarkeit und Modellfairness des BERT-Scoring-Modells aus Artikel 3 interessant. Die ersten Analysen in Artikel 3 (Abschnitt 6.6.2) sind eher als Machbarkeitsnachweis gedacht und haben noch nicht den Anspruch hier bereits einen größeren Beitrag zu leisten. Die bestehenden Ansätze insbesondere zur Erklärbarkeit des Modells, d. h. der Zurückführung von Score-Entscheidungen auf konkreten Sprachgebrauch und Wortwahl

(Gombert et al., 2023; Sundararajan et al., 2017), können aber praktikabel genutzt werden, um die Betrachtungen zu intensivieren. Ähnliches gilt auch für die aufgabenweisen Analysen zu Einflussfaktoren auf die Scoring-Performanz (Zesch et al., 2023; Zhai, 2021 Abschnitt 6.7.5), wobei der Code des Projekts hierfür bereits Methoden enthält, die auch auf andere Testinstrumente und Datensätze direkt angewandt werden können.

Ein weiterer Anknüpfungspunkt sind auch die explorierten alternativen Scoring-Modelle. Dabei sind Embedding-basierte Ansätze (Abschnitt 6.7.6) aufgrund des deutlich verringerten Aufwands beim Training gerade für sehr große Datensätze interessant. Die exemplarische Untersuchung der Fähigkeit von GPT4o-mini Aufgaben auf Basis eines Prompts, der aus dem Kodiermanual abgeleitet wurde, zuverlässig zu bepunkten (Abschnitt 6.7.8) ist dagegen gerade für kleine Datensätze interessant, da hierfür überhaupt keine Trainingsdaten benötigt werden. Zuletzt wurden auch Methoden zur Anreicherung von Datensätzen (insbesondere im Kontext der ungleichmäßigen Verteilung von Labels) mithilfe generativer KI in den Blick genommen (Kieser et al., 2023; Martin & Graulich, 2024). Dies könnte einen Ansatzpunkt darstellen, um die Performanz der hier bereits explorierten Modelle weiter zu erhöhen.

Die entwickelten Modelle dienen zwar in den Artikeln primär als Mittel zum Zweck, um im Rahmen des CGT-Frameworks die Validität der gefundenen inhaltlichen Strukturen zu unterstreichen, ihr potenzieller Nutzen für ein vollautomatisiertes FDW-Assessment liegt jedoch auf der Hand und wurde bereits mehrfach in dieser Arbeit erwähnt. Um diese Modelle in der Praxis tatsächlich für ein Assessment einzusetzen, müssten sie allerdings nutzbar gemacht werden. Einerseits können hierfür Datensätze über klassische Umfragetools gesammelt werden und im Anschluss manuell mithilfe des bestehenden Modells über den Programmcode dieses Projekts bepunktet und ausgewertet werden. Für einen solchen Einsatz sind alle notwendigen Voraussetzungen somit bereits erfüllt. In diesem Setting ist zwar der manuelle Aufwand der Kodierung und somit der größte Arbeitsanteil automatisiert, allerdings wäre dabei immer noch ein Zwischenschritt mit menschlicher Beteiligung notwendig. Praktikabler wäre die direkte Anbindung des Scoring-Modells an entsprechende Umfragesoftware, was sich allerdings bei den üblichen Anbietern entsprechender Tools (z. B. LimeSurvey[79]) meist als schwierig bis unmöglich erweist. Als mögliche Alternative wurde daher in diesem Projekt auch ein Proof of Concept für ein vollständiges Open-Source-Webtool erstellt, welches die Bearbeitung des Testinstruments ermöglicht und auch das BERT-Modell für die automatisierte Auswertung umfasst[80]. Einige Impressionen der „Beta-Version“ dieses Webtools sind in Anhang H festgehalten; hier ist allerdings noch weitere Entwicklungsarbeit und im Anschluss auch eine Evaluierung des Tools notwendig.

[79] https://www.limesurvey.org/de, Zugegriffen 17. Januar 2025

[80] Dieses Webtool wurde ausschließlich mit frei verfügbaren Python- und Javascript-Bibliotheken erstellt und die Datenorganisation orientiert sich an den von Buschhüter et al. (2023) vorgeschlagenen Strukturen für eine flexible Datenverwaltung im Forschungskontext. Auch der Code für das Webtool ist in einer Testinstrument-unabhängigen Form denkbar. Wie der Analysecode ist auch der Code für das Webtool als Open-Source-Projekt öffentlich verfügbar (https://github.com/JannisZeller/questionnaire-webtool), um für eine etwaige Weiterentwicklung und Nachnutzung bereit zu stehen. Siehe auch Anhang H.

7.4. Beiträge des Dissertationsprojekts als Übersicht

Beiträge zur Theoriebildung im Kontext des FDW

- Entwicklung projektübergreifend gültiger (aber recht allgemeiner) FDW-Kompetenzniveaus auf Basis von Item-Response-Modellierungen zweier FDW-Datensätze aus unterschiedlichen Projekten (Kapitel 4 / Artikel 1).
- Entwicklung eines Modells hierarchischer Komplexität für das FDW in Physik durch Adaption eines bestehenden Modells hierarchischer Komplexität für physikalisches Fachwissen (Kapitel 4 / Artikel 1), allerdings mit eingeschränkter Übertragbarkeit auf FDW-Testinstrumente außerhalb des Projekts ProfiLe-P.
- Theoretisch fundierte Analyse von FDW-Testaufgaben mit Fokus auf lernpsychologische Operatoren angelehnt an die Taxonomie von Anderson und Krathwohl (2001) inklusive der Betrachtung von Interrater-Übereinstimmungswerten dreier Expert:innen (Kapitel 5 / Artikel 2).
- Ermittlung und Bestätigung (im Sinne der CGT) von K-Means-Clustern (Kapitel 5 / Artikel 2) und latenten Kompetenzprofilen (Kapitel 6 / Artikel 3) des FDW mit prototypischen Stärken und Schwächen bezüglich der Anforderungsbereiche „Anwenden-Kreieren“ und „Analysieren-Evaluieren“ sowie Tendenzen zu prototypischem Sprachgebrauch in den Antworten auf die offenen Fragen des Testinstruments. Insbesondere diese Ergebnisse deuten auf die Trennbarkeit von entsprechenden einzelnen FDW-Komponenten hin, die im Rahmen der Weiterentwicklung von FDW-Rahmenmodellen wie dem RCM of PCK oder dem Kontinuumsmodells von Interesse sind.

Methodische und Praktische Beiträge

- Entwicklung eines Verfahrens zur gemeinsamen Niveaubetrachtung eines Konstrukts auf Basis von unterschiedlichen Datensätze und IRT-Modellen (Kapitel 4 / Artikel 1).
- Entwicklung eines CGT-orientierten Workflows für interpretierbare und informative Cluster-Analysen von Score-Datensätzen durch die Betrachtung theoretisch fundierter Subskalen (Kapitel 5 & 6 / Artikel 2 & 3) → Bereitstellung von entsprechendem Testinstrument-unabhängigen Python-Analysecode für zukünftige Analysen.
- Entwicklung eines zweistufigen Workflows zur Evaluierung eines automatisierten Scoring-Systems mit der Betrachtung zusätzlicher „Downstream-Tasks“ (Kompetenzprofile & Subskalenscores) über die reine Bepunktung hinaus (Kapitel 6 / Artikel 3). Dabei wird insbesondere ein Data-Leakage freies Cross-Validation-Verfahren beschrieben (siehe auch Abschnitt 6.7.3) → Bereitstellung von entsprechendem Testinstrument-unabhängigen Python-Analysecode für zukünftige Analysen.
- Exploration der Nutzung unterschiedlicher Modelle und Tools für ein automatisiertes Scoring der Freitextaufgaben mit Kurzantworten (Abschnitt 6.7.5, 6.7.6 & 6.7.8).
- Entwicklung eines Proof-of-Concept für ein vollautomatisiertes Machine-Learning-basiertes Assessment-Webtool unter der Nutzung von Open-Source-Software (Anhang H).

Literaturverzeichnis

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., . . . Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. In K. Keeton & T. Roscoe (Hrsg.), *ACM Other conferences, OSDI'16: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation* (S. 265–283). USENIX Association. https://dl.acm.org/doi/10.5555/3026877.3026899

Alonzo, A. C., Berry, A. & Nilsson, P. (2019). Unpacking the Complexity of Science Teachers' PCK in Action: Enacted and Personal PCK. In A. Hume, R. Cooper & A. Borowski (Hrsg.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (S. 273–288). Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2_12

Amari, S.-I. (1972). Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements. *IEEE Transactions on Computers*, *C-21*(11), 1197–1206. https://doi.org/10.1109/T-C.1972.223477

Andersen, N. & Zehner, F. (2021). shinyReCoR: A Shiny Application for Automatically Coding Text Responses Using R. *Psych*, *3*(3), 422–446. https://doi.org/10.3390/psych3030030

Anderson, L. W. & Krathwohl, D. R. (Hrsg.). (2001). *A taxonomy for learning, teaching, and assessing A revision of Bloom's taxonomy of educational objectives* (4. Aufl.). Longman.

Anthropic (Hrsg.). (2024). *Raising the bar on SWE-bench Verified with Claude 3.5 Sonnet*. Abgerufen am 17.01.2025, von https://www.anthropic.com/research/swe-bench-sonnet.

Bahdanau, D., Cho, K. & Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv:1409.0473v7

Ball, D. L., Lubienski, S. T. & Mewborn, D. S. (2001). Research on Teaching Mathematics: The Unsolved Problem of Teachers' Mathematical Knowledge. In V. Richardson (Hrsg.), *Handbook of Research on Teaching* (4. Aufl., S. 433–456). American Educational Research Association.

Barocas, S., Hardt, M. & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://fairmlbook.org/

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. (2011). Das Kompetenzmodell von COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 29–53). Waxmann. http://www.ciando.com/ebook/bid-229459/intRefererID/241664

Beaton, A. E. & Allen, N. L. (1992). Interpreting Scales Through Scale Anchoring. *Journal of Educational Statistics*, *17*(2), 191–204. https://doi.org/10.2307/1165169

Behling, F., Förtsch, C. & Neuhaus, B. J. (2022a). The Refined Consensus Model of Pedagogical Content Knowledge (PCK): Detecting Filters between the Realms of PCK. *Education Sciences*, *12*(9). https://doi.org/10.3390/educsci12090592

Behling, F., Förtsch, C. & Neuhaus, B. J. (2022b). Using the Plan-Teach-Reflect Cycle of the Refined Consensus Model of PCK to Improve Pre-Service Biology Teachers' Personal PCK as Well as Their Motivational Orientations. *Education Sciences*, *12*(10). https://doi.org/10.3390/educsci12100654

Bernholt, S. (2010). *Kompetenzmodellierung in der Chemie. Theoretische und empirische Reflexion am Beispiel des Modells hierarchischer Komplexität. Studien zum Physik- und Chemielernen*. Logos-Verlag.

Berry, A., Friedrichsen, P. & Loughran, J. (Hrsg.). (2015). *Re-Examining Pedagogical Content Knowledge In Science Education*. Routledge. https://doi.org/10.4324/9781315735665

Bird, S., Loper, E. & Klein, E. (2009). *Natural Language Processingwith Python*. O'Reilly Media, Inc.

Bishop, C. M. & Lasserre, J. (2007). Generative or Discriminative? Getting the Best of Both Worlds. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith & M. West (Hrsg.), *Bayesian Statistics* (8. Aufl., S. 3–24). Oxford University Press.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M. & Lafferty, J. D. (2005). Correlated Topic Models. *Advances in Neural Information Processing Systems*, *18*. https://proceedings.neurips.cc/paper/2005/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Blömeke, S., Felbrich, A., Müller, C., Kaiser, G. & Lehmann, R. (2008a). Effectiveness of teacher education. *ZDM Mathematics Education*, *40*, 719–734. https://doi.org/10.1007/s11858-008-0096-x

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., Jentsch, A., Ross, N., Kaiser, G. & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progress. *Learning and Instruction*, *79*, 101600. https://doi.org/10.1016/j.learninstruc.2022.101600

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.). (2008b). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer - Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare - Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung*. Waxmann Verlag.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin (Hrsg.), *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (S. 1877–1901). Curran Associates, Inc. https://dl.acm.org/doi/abs/10.5555/3495724.3495883

Buehl, M. M. & Beck, J. S. (2014). The Relationship between Teachers' Beliefs and Teachers' Practices. In H. Fives & M. G. Gill (Hrsg.), *International Handbook of Research on Teachers' Beliefs.* Routledge.

Buschhüter, D., Zeller, J., Oltmanns, S., Borowski, A., Kulgemeyer, C., Riese, J. & Vogelsang, C. (2023). Forschungsdatenmanagement erleichtern durch relationale Datenbanken: Ein Datenmodell für naturwissenschaftsdidaktische Forschung. In H. van Vorst (Hrsg.), *Lernen, Lehren und Forschen in einer digital geprägten Welt, Tagungsband der GDCP Jahrestagung 2022* (S. 869–872). Gesellschaft für Didaktik der Chemie und Physik.

Campello, R. J. G. B., Moulavi, D. & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda & G. Xu (Hrsg.), *Advances in Knowledge Discovery and Data Mining. 17th Pacific-Asia Conference, PAKDD 2013, Proceedings, Part II* (S. 160–172). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Camus, L. & Filighera, A. (2020). Investigating Transformers for Automatic Short Answer Grading. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin & E. Millán (Hrsg.), *Artificial Intelligence in Education 21 (AIED 2020)* (S. 43–48). Springer Cham. https://doi.org/10.1007/978-3-030-52240-7_8

Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Kam Ho Chan, K., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., . . . Wilson, C. D. (2019). The

Refined Consensus Model of Pedagogical Content Knowledge in Science Education. In A. Hume, R. Cooper & A. Borowski (Hrsg.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (S. 77–94). Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2_2

Cauet, E., Liepertz, S., Borowski, A. & Fischer, H. E. (2015). Does it matter what we measure? Domain-specific professional knowledge of physics. *Revue suisse des sciences de l'éducation*, *37*(3), 462–479. https://doi.org/10.25656/01:12746

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams & A. Culotta (Hrsg.), *NIPS'09: Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Vol. 22, S. 288–296). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

Chen, Y.-H., Sarokin, R., Lee, J., Tang, J., Chang, C.-L., Kulik, A. & Grundmann, M. (2023). *Speed Is All You Need: On-Device Acceleration of Large Diffusion Models via GPU-Aware Optimizations*. arXiv:2304.11267v2

Chen, Z. & Liu, B. (2017). Topic Models for NLP Applications. In C. Sammut & G. I. Webb (Hrsg.), *Encyclopedia of Machine Learning and Data Mining* (S. 1276–1280). Springer. https://doi.org/10.1007/978-1-4899-7687-1_906

Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G. & LeCun, Y. (2015). The Loss Surfaces of Multilayer Networks. In G. Lebanon & S. V. N. Vishwanathan (Hrsg.), *Proceedings of Machine Learning Research, Proceedings of the 18th International Conference on Artificial Intelligence and Statistics* (S. 192–204). PMLR. https://proceedings.mlr.press/v38/choromanska15.html

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S. & Amodei, D. (2017). Deep reinforcement learning from human preferences. In U. von Luxburg, I. Guyon, S. Bengio, H. M. Wallach & R. Fergus (Hrsg.), *NIPS'17, NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (S. 4302–4310). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3294996.3295184

Chung, H. W., Le Hou, Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., . . . Wei, J. (2024). Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, *25*, 1–53. https://jmlr.org/papers/v25/23-0870.html

Commons, M. L., Crone-Todd, D. & Chen, S. J. (2014). Using SAFMEDS and direct instruction to teach the model of hierarchical complexity. *The Behavior Analyst Today*, *14*(1-2), 31–45. https://doi.org/10.1037/h0101284

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A. & Krause, S. R. (1998). Hierarchical Complexity of Tasks Shows the Existence of Developmental Stages. *Developmental Review*, *18*(3), 237–278. https://doi.org/10.1006/drev.1998.0467

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D. & Chen, G. (2023). Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In M. Chang, N.-S. Chen, R. Kuo, G. Rudolph, D. G. Sampson & A. Tlili (Hrsg.), *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (S. 323–325). IEEE Computer Society. https://doi.org/10.1109/ICALT58122.2023.00100

Dao, T., Fu, D., Ermon, S., Rudra, A. & Ré, C. (2024). FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh (Hrsg.), *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems* (S. 16344–16359). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3600270.3601459

Davier, A. A. von, Carstensen, C. H. & Davier, M. von. (2008). Linking competencies in horizontal, vertical, and longitudinal settings and measuring growth. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 121–149). Hogrefe.

Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. (2024). QLoRA: Efficient Finetuning of Quantized LLMs. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Hrsg.), *NIPS '23, NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems* (S. 10088–10115). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3666122.3666563

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran & T. Solorio (Hrsg.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (S. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence*, *2*. https://doi.org/10.1038/s42256-020-0219-9

Döring, N. (2023). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (6. Aufl.). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-64762-2

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR 2021)*. https://openreview.net/forum?id=YicbFdNTTy

Duda, R. O., Hart, P. E. & Stork, D. G. (2001). *Pattern Classification* (2. Aufl.). Wiley.

Enkrott, P. (2021). *Entwicklung des fachlichen Wissens angehender Physiklehrkräfte* [Dissertation]. Universität Potsdam. https://doi.org/10.25932/publishup-50040

Esser, P., Rombach, R. & Ommer, B. (2021). Taming Transformers for High-Resolution Image Synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Estrellado, R. A., Freer, Emily, A., Rosenberg, J. M. & Velásquez, I. C. (2020). *Data Science in Education Using R*. Routledge. https://doi.org/10.4324/9780367822842

Fleiss, J. L. & Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. https://doi.org/10.1177/001316447303300309

Förtsch, C., Werner, S., Kotzebue, L. von & Neuhaus, B. J. (2016). Effects of biology teachers' professional knowledge and cognitive activation on students' achievement. *International Journal of Science Education*, *38*(17), 2642–2666. https://doi.org/10.1080/09500693.2016.1257170

Förtsch, S., Förtsch, C., Kotzebue, L. von & Neuhaus, B. J. (2018). Effects of Teachers' Professional Knowledge and Their Use of Three-Dimensional Physical Models in Biology Lessons on Students' Achievement. *2227-7102*, *8*(3). https://doi.org/10.3390/educsci8030118

Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M. & Gerjets, P. (2023). ChatGPT in Education: Global Reactions to AI Innovations. *Scientific Reports*, *13*, Artikel 15310. https://doi.org/10.1038/s41598-023-42227-6

Gamieldien, Y., McCord, R. & Katz, A. (2023). Utilizing Natural Language Processing to Examine Self-Reflections in Self-Regulated Learning. *SSRN pre-print.* https://doi.org/10.2139/ssrn.4487795

Gan, J. & Qi, Y. (2021). Selection of the Optimal Number of Topics for LDA Topic Model - Taking Patent Policy Analysis as an Example. *Entropy*, *23*(10). https://doi.org/10.3390/e23101301

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M. & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv:2312.10997v5

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2. Aufl.). O'Reilly Media, Inc.

Gess-Newsome, J. (1999). Pedagogical Content Knowledge: An Introduction and Orientation. In J. Gess-Newsome & N. G. Lederman (Hrsg.), *Examining pedagogical content knowledge* (S. 3–17). Springer. https://doi.org/10.1007/0-306-47217-1_1

Gess-Newsome, J. & Lederman, N. G. (Hrsg.). (1999). *Examining pedagogical content knowledge*. Springer. https://doi.org/10.1007/0-306-47217-1

Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K. & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning*, *39*(3), 767–786. https://doi.org/10.1111/jcal.12767

Gramzow, Y. (2015). Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik: Modellierung und Testkonstruktion. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 181). Logos Verlag.

Gramzow, Y., Riese, J. & Reinhold, P. (2013). Modellierung fachdidaktischen Wissens angehender Physiklehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 7–30. https://doi.org/10.25656/01:31713

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794

Großmann, L. & Krüger, D. (2022). Welche Rolle spielt das fachdidaktische Wissen von Biologie-Referendar*innen für die Qualität ihrer Unterrichtsentwürfe? *Zeitschrift für Didaktik der Naturwissenschaften*, *28*(4), 53–72. https://doi.org/10.1007/s40573-022-00141-w

Großschedl, J., Harms, U., Kleickmann, T. & Glowinski, I. (2015). Preservice biology teachers' professional knowledge: Structure and learning opportunities. *Journal of Science Teacher Education*, *26*(3), 291–318. https://doi.org/10.1007/s10972-015-9423-6

Gu, A. & Dao, T. (2023). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. arXiv:2312.00752v2

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. Edward Arnold.

Harms, U. & Riese, J. (2018). Professionelle Kompetenz und Professionswissen. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 283–298). Springer Spektrum. https://doi.org/10.1007/978-3-662-56320-5_17

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung* (S. 83–99). Beltz. https://doi.org/10.25656/01:3143

Hattie, J. (2003). Teachers Make a Difference, What is the research evidence. In M. Meiers (Hrsg.), *Proceedings of the Building Teacher Quality: What does the research tell us ACER Research Conference.* ACER. http://research.acer.edu.au/research_conference_2003/4/

Hattie, J. (2009). *Visible Learning. A synthesis of over 800 metaanalyses relating to achievement*. Routledge.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Hattie, J. & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hennig, P., Stern, D., Herbrich, R. & Graepel, T. (2012). Kernel Topic Models. In N. D. Lawrence & M. Girolami (Hrsg.), *Proceedings of Machine Learning Research. Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (Bd. 22, S. 511–519). PMLR. https://proceedings.mlr.press/v22/hennig12.html

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D. & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, *9*(3), e3310. https://doi.org/10.1002/rev3.3310

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Las Casas, D. de, Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., . . . Sifre, L. (2024). Training compute-optimal large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh (Hrsg.), *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems* (S. 30016–30030). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3600270.3602446

Hornik, K., Stinchcombe, M. & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR 2022)*. https://openreview.net/forum?id=nZeVKeeFYf9

Hume, A., Cooper, R. & Borowski, A. (Hrsg.). (2019). *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science*. Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2

Jolliffe, I. T. (2002). *Principal Component Analysis* (2. Aufl.). Springer New York. https://doi.org/10.1007/b98835

Jordans, M., Zeller, J., Große-Heilmann, R. I. & Riese, J. (2022). Weiterentwicklung eines physikdidaktischen Tests zum Online-Assessment. In S. Habig & H. van Vorst (Hrsg.), *Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen, Tagungsband der GDCP Jahrestagung 2021* (S. 764–767). Gesellschaft für Didaktik der Chemie und Physik.

Jurafsky, D. & Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Stanford University. Abgerufen am 17.01.2025, von https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Kaiser, G., Bremerich-Vos, A. & König, J. (2020). Professionswissen. In C. Cramer, J. König, M. Rothland & S. Blömeke (Hrsg.), *Handbuch Lehrerinnen- und Lehrerbildung* (S. 811–818). Klinkhardt. https://doi.org/10.35468/hblb2020-100

Kapoor, S. & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kaufman, S., Rosset, S. & Perlich, C. (2012). Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), Artikel 15, 1–21. https://doi.org/10.1145/2382577.2382579

Keller, M. M., Neumann, K. & Fischer, H. E. (2017). The Impact of Physics Teachers' Pedagogical Content Knowledge and Motivation on Students' Achievement and Interest. *Journal of Research in Science Teaching*, *54*(5), 586–614. https://doi.org/10.1002/tea.21378

Kieser, F., Wulff, P., Kuhn, J. & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. *Physical Review Physics Education Research*, *19*(2). https://doi.org/10.1103/PhysRevPhysEducRes.19.020150

Kingma, D. P. & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. arXiv:1412.6980v9

Kirschner, S. (2013). Modellierung und Analyse des Professionswissens von Physiklehrkräften. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 161). Logos Verlag. urn:nbn:de:hbz:464-20131210-150745-4

Kirschner, S., Sczudlek, M., Tepner, O., Borowski, A., Fischer, H. E., Lenske, G., Leutner, D., Neuhaus, B. J., Sumfleth, E., Thillmann, H. & Wirth, J. (2017). Professionswissen in den Naturwissenschaften (ProwiN). In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals* (S. 113–130). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-07274-2_7

Kleickmann, T., Großschedl, J., Harms, U., Heinze, A., Herzog, S., Hohenstein, F., Köller, O., Kröger, J., Lindmeier, A., Loch, C., Mahler, D., Möller, J., Neumann, K., Parchmann, I., Steffensky, M., Taskin, V.

& Zimmermann, F. (2014). Professionswissen von Lehramtsstudierenden der mathematisch-naturwissenschaftlichen Fächer-Testentwicklung im Rahmen des Projekts KiL. *Unterrichtswissenschaft*, *42*(3), 280–288.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. *Bildungsreform: Bd. 1*. BMBF. https://doi.org/10.25656/01:20901

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., . . . Liang, P. (2021). WILDS: A Benchmark of in-the-Wild Distribution Shifts. In M. Meila & T. Zhang (Hrsg.), *Proceedings of Machine Learning Research, Proceedings of the 38th International Conference on Machine Learning* (S. 5637–5664). PMLR. https://proceedings.mlr.press/v139/koh21a

Kokoska, S. & Zwillinger, D. (2000). *CRC Standard Probability and Statistics Tables and Formulae, Student Edition*. CRC Press. https://doi.org/10.1201/b16923

König, J. (2009). Zur Bildung von Kompetenzniveaus im Pädagogischen Wissen von Lehramtsstudierenden: Terminologie und Komplexität kognitiver Bearbeitungsprozesse als Anforderungsmerkmale von Testaufgaben? *Lehrerbildung auf dem Prüfstand*, *2*(2), 244–262. https://doi.org/10.25656/01:14703

König, J. (Hrsg.). (2012). *Teachers' pedagogical beliefs. Definition and operationalisation, connections to knowledge and performance, development and change*. Waxmann. https://doi.org/10.25656/01:21030

König, J. & Seifert, A. (Hrsg.). (2012). *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung*. Waxmann. https://doi.org/10.25656/01:21029

Köpf, A., Kilcher, Y., Rütte, D. von, Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H. & Mattick, A. (2024). OpenAssistant conversations - democratizing large language model alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Hrsg.), *NIPS '23, NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems* (S. 47669–47681). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3666122.3668186

Kramer, M., Förtsch, C., Boone, W. J., Seidel, T. & Neuhaus, B. J. (2021). Investigating Pre-Service Biology Teachers' Diagnostic Competences: Relationships between Professional Knowledge, Diagnostic Activities, and Diagnostic Accuracy. *2227-7102*, *11*(3). https://doi.org/10.3390/educsci11030089

Krauss, S., Neubrand, M., Blum, W., Baumert, J., Brunner, M., Kunter, M. & Jordan, A. (2008). Die Untersuchung des professionellen Wissens deutscher Mathematik-Lehrerinnen und -Lehrer im Rahmen der COACTIV-Studie. *Journal für Mathematik-Didaktik*, *29*(3-4), 233–258. https://doi.org/10.1007/BF03339063

Krebs, R. (1997). The Swiss Way to Score Multiple True-False Items: Theoretical and Empirical Evidence. In Scherpbier, Albert J. J. A., C. P. M. van der Vleuten, J.-J. Rethans & A. F. W. van der Steeg (Hrsg.), *Advances in Medical Education* (S. 158–161). Springer Netherlands. https://doi.org/10.1007/978-94-011-4886-3_46

Kröger, J. (2019). *Struktur und Entwicklung des Professionswissens angehender Physiklehrkräfte* [Dissertation]. Christian-Albrechts-Universität Kiel.

Krüger, D. & Krell, M. (2020). Maschinelles Lernen mit Aussagen zur Modellkompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, *26*, 157–172. https://doi.org/10.1007/s40573-020-00118-7

Kubsch, M., Krist, C. & Rosenberg, J. M. (2022). Distributing epistemic functions and tasks - A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*, *60*(2), 423–447. https://doi.org/10.1002/tea.21803

Kubsch, M., Neumann, K., Rochnia, M. & Gräsel, C. (2023). 50 Jahre Unterrichtswissenschaft - Themen der Forschung über das Lehren und Lernen im Wandel. *Unterrichtswissenschaft*, *51*, 15–37. https://doi.org/10.1007/s42010-023-00164-3

Kubsch, M., Rosenberg, J. M. & Krist, C. (2021a). Beyond Supervision: Human / Machine Distributed Learning in Learning Sciences Research. In E. de Vries, Y. Hod & Ahn June (Hrsg.), *15th International Conference of the Learning Sciences (ICLS) - Proceedings* (S. 897–898). International Society of the Learning Sciences.

Kubsch, M., Stamer, I., Steiner, M., Neumann, K. & Parchmann, I. (2021b). Beyond p-values: Using Bayesian Data Analysis in Science Education Research. *Practical Assessment, Research, and Evaluation*, *26*, Article 4. https://doi.org/10.7275/vzpw-ng13

Kulgemeyer, C., Borowski, A., Buschhüter, D., Enkrott, P., Kempin, M., Reinhold, P., Riese, J., Schecker, H., Schröder, J. & Vogelsang, C. (2020). Professional knowledge affects action-related skills: The development of preservice physics teachers' explaining skills during a field experience. *Journal of Research in Science Teaching*, *52*(10), 1554–1582. https://doi.org/10.1002/tea.21632

Kulgemeyer, C., Borowski, A., Fischer, H. E., Gramzow, Y., Reinhold, P., Riese, J., Schecker, H., Tomczyszyn, E. & Walzer, M. (2012). ProfiLe-P - Professionswissen in der Lehramtsausbildung Physik. Vorstellung eines Forschungsverbundes. *PhyDid B - Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung*. http://phydid.de/index.php/phydid-b/article/view/380

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 Erklärensfähigkeit angehender Physiklehrkräfte in einer simulierten Unterrichtssituation. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*, 111–126. https://doi.org/10.1007/s40573-015-0029-5

Kultusministerkonferenz. (2024). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. (Beschluss der Kultusministerkonferenz vom 16.10.2008 i. d. F. vom 08.02.2024)*. Abgerufen am 17.01.2025, von https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T. & Hachfeld, A. (2013). Professional Competence of Teachers: Effects on Instructional Quality and Student Development. *Journal of Educational Psychology*, *105*, 805–820. https://doi.org/10.1037/a0032583

Latif, E., Lee, G.-G., Neumann, K., Kastorff, T. & Zhai, X. (2024). *G-SciEdBERT: A Contextualized LLM for Science Assessment Tasks in German*. arXiv:2402.06584v2

Latif, E. & Zhai, X. (2023). *Fine-tuning ChatGPT for Automatic Scoring*. arXiv:2310.10072v3

Leacock, C. & Chodorow, M. (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities*, *37*, 389–405. https://doi.org/10.1023/A:1025779619903

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Lee, H.-S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M. & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590–622. https://doi.org/10.1002/sce.21504

Lee, W.-C. & Lee, G. (2018). Linking and Equating. In Paul Irwing, Tom Booth & David J. Hughes (Hrsg.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (S. 639–673). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118489772.ch21

Lemaître, G., Nogueira, F. & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*(17), 1–5. http://jmlr.org/papers/v18/16-365.html

Liepertz, S. & Borowski, A. (2019). Testing the Consensus Model: relationships among physics teachers' professional knowledge, interconnectedness of content structure and student achievement. *International Journal of Science Education*, *41*(7), 890–910. https://doi.org/10.1080/09500693.2018.1478165

Linacre, J. M. (1998). Thurstone Thresholds and the Rasch Model. *Rasch Measurement Transactions*, *12*(2), 634–635. https://www.rasch.org/rmt/rmt122j.htm

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G. (2023). Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9). https://doi.org/10.1145/3560815

Liu, Z., Lin, Y. & Sun, M. (2020). *Representation Learning for Natural Language Processing*. Springer. https://doi.org/10.1007/978-981-15-5573-2

Lok, B., McNaught, C. & Young, K. (2016). Criterion-referenced and norm-referenced assessments: compatibility and complementarity. *Assessment & Evaluation in Higher Education*, *41*(3), 450–465. https://doi.org/10.1080/02602938.2015.1022136

Ludwig, S., Mayer, C., Hansen, C., Eilers, K. & Brandt, S. (2021). Automated Essay Scoring Using Transformer Models. *Psych*, *3*(4), 897–915. https://doi.org/10.3390/psych3040056

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Hrsg.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (S. 281–297). University of California Press.

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B. & Krajcik, J. (2021). Using Machine Learning to Score Multi-Dimensional Assessments of Chemistry and Physics. *Journal of the Learning Sciences*, *30*, 239–254. https://doi.org/10.1007/s10956-020-09895-9

Magnusson, S., Krajcik, J. & Borko, H. (1999). Nature, Sources, and Development of Pedagogical Content Knowledge for Science Teaching. In J. Gess-Newsome & N. G. Lederman (Hrsg.), *Examining pedagogical content knowledge* (S. 95–132). Springer. https://doi.org/10.1007/0-306-47217-1_4

Martin, P. P. & Graulich, N. (2024). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence*, *7*, 100265. https://doi.org/10.1016/j.caeai.2024.100265

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mayfield, E. & Rosé, C. P. (2012). *LightSIDE Text Mining and Machine Learning User's Manual*. Carnegie Mellon University. Abgerufen am 17.01.2025, von https://www.cs.cmu.edu/~emayfiel/LightSIDE.pdf.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (2. Aufl.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

McInnes, L., Healy, J. & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J. & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv:1802.03426v3

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. https://doi.org/10.1037/0003-066X.50.9.741

Meta (Hrsg.). (2024). *Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.* https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Mientus, L., Hume, A., Wulff, P., Meiners, A. & Borowski, A. (2022). Modelling STEM Teachers' Pedagogical Content Knowledge in the Framework of the Refined Consensus Model: A Systematic Literature Review. *Education Sciences*, *12*(6). https://doi.org/10.3390/educsci12060385

Mientus, L., Wulff, P., Nowak, A. & Borowski, A. (2021). ReFeed: computerunterstütztes Feedback zu Reflexionstexten. In M. Kubsch, S. Sorge, J. Arnold & N. Graulich (Hrsg.), *Lehrkräftebildung neu gedacht Ein Praxishandbuch für die Lehre in den Naturwissenschaften und deren Didaktiken* (S. 160–165). Waxmann. https://doi.org/10.25656/01:22414

Mientus, L., Wulff, P., Nowak, A. & Borowski, A. (2023). Fast-and-frugal means to assess reflection-related reasoning processes in teacher training—Development and evaluation of a scalable machine learning-based metric. *Zeitschrift für Erziehungswissenschaft*, *26*, 677–702. https://doi.org/10.1007/s11618-023-01166-8

Mikolov, T., Chen, K., Corrade, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations (ICLR 2013)*. https://api.semanticscholar.org/CorpusID:5959482

Miller, L. A. & Lovler, R. L. (2018). *Foundations of Psychological Testing: A Practical Approach* (6. Aufl.). Sage Publications.

Mistral AI (Hrsg.). (o. D.). *Tokenization*. Abgerufen am 17.01.2025, von https://docs.mistral.ai/guides/tokenization/.

Mitchell, T. (1997). *Machine learning*. McGraw Hill.

Mladenić, D., Brank, J. & Grobelnik, M. (2016). Document Classification. In C. Sammut & G. I. Webb (Hrsg.), *Encyclopedia of Machine Learning and Data Mining.* Springer US. https://doi.org/10.1007/978-1-4899-7502-7_75-1

Moosbrugger, H. & Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion* (3. Aufl.). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-61532-4

Mullis, I. V. S., Cotter, K. E., Centurino, V. A. S., Fishbein, B. G. & Liu, J. (2016). Using scale anchoring to interpret the TIMSS 2015 achievement scales. In M. O. Martin, I. V. S. Mullis & M. Hooper (Hrsg.), *Methods and Procedures in TIMSS 2015* (14.1-14.47). Lynch School of Education. https://timssandpirls.bc.edu/publications/timss/2015-methods/chapter-1.html

Mullis, I. V. S. & Fishbein, B. G. (2020). Using scale anchoring to interpret the TIMSS 2019 achievement scales. In M. O. Martin, M. von Davier & I. V. S. Mullis (Hrsg.), *Methods and procedures: TIMSS 2019 technical report* (15.1-15.60). TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-15.html

Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press. https://probml.github.io/pml-book/book1.html

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. & Mian, A. (2024). *A Comprehensive Overview of Large Language Models*. arXiv:2307.06435v10

Nehring, A., Buschhüter, D., Kubsch, M., Ludwig, T., Wulff, P. & Neumann, K. (2025). Künstliche Intelligenz in den Naturwissenschaftsdidaktiken – gekommen, um zu bleiben: Potenziale, Desiderata,

Herausforderungen. *Zeitschrift für Didaktik der Naturwissenschaften*, *31*, Artikel 2. https://doi.org/10.1007/s40573-025-00177-8

Nelson, L. K. (2020). Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Neumann, K. (2014). Rasch-Analyse naturwissenschaftsbezogener Leistungstests. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 355–369). Springer Spektrum. https://doi.org/10.1007/978-3-642-37827-0_28

Neumann, K., Kind, V. & Harms, U. (2019). Probing the amalgam: the relationship between science teachers' content, pedagogical and pedagogical content knowledge. *International Journal of Science Education*, *41*(7), 847–861. https://doi.org/10.1080/09500693.2018.1497217

Ng, A. & Jordan, M. (2001). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In Dietterich, T., G., S. Becker & Z. Ghahramani (Hrsg.), *NIPS'01: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic.* MIT Press. https://proceedings.neurips.cc/paper/2001/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf

Nold, G., Rossa, H. & Hartig, J. (2008). Proficiency scaling in DESI listening and reading EFL tests: Task characteristics, item difficulty and cut-off points. In L. Taylor & C. J. Weir (Hrsg.), *Multilingualism and assessment. Achieving transparency, assuring quality, sustaining diversity. proceedings of the ALTE Berlin Conference, May 2005.* (S. 94–116). Cambridge University Press.

Ollama (Hrsg.). (2024). *Llama 3.2 goes small and multimodal*. Abgerufen am 17.01.2025, von https://ollama.com/blog/llama3.2.

OpenAI (Hrsg.). (o. D.-a). *Fine-tuning*. Abgerufen am 17.01.2025, von https://platform.openai.com/docs/guides/fine-tuning.

OpenAI (Hrsg.). (o. D.-b). *Vector embeddings*. Abgerufen am 17.01.2025, von https://platform.openai.com/docs/guides/embeddings.

OpenAI (Hrsg.). (2022). *ChatGPT: Optimizing Language Models for Dialogue*. Abgerufen am 17.01.2025, von https://openai.com/blog/chatgpt/.

OpenAI (Hrsg.). (2024a). *Enterprise privacy at OpenAI*. Abgerufen am 17.01.2025, von https://openai.com/enterprise-privacy/.

OpenAI (Hrsg.). (2024b). *GPT-4o mini: advancing cost-efficient intelligence*. Abgerufen am 17.01.2025, von https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI (Hrsg.). (2024c). *Hello GPT-4o*. Abgerufen am 17.01.2025, von https://openai.com/index/hello-gpt-4o/.

Organisation for Economic Cooperation and Development (Hrsg.). (2018). *PISA 2018 Technical Report*. Abgerufen am 17.01.2025, von https://www.oecd.org/en/about/programmes/pisa/pisa-data.html.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. & Lowe, R. (2024). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh (Hrsg.), *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems* (S. 27730–27744). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3600270.3602281

Park, S. & Oliver, J. S. (2008). National Board Certification (NBC) as a catalyst for teachers' learning about teaching: The effects of the NBC process on candidate teachers' PCK development. *Journal of Research in Science Teaching*, *45*(7), 812–834. https://doi.org/10.1002/tea.20234

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning

Library. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc & E. B. Fox (Hrsg.), *NIPS '19: Proceedings of the 33th International Conference on Neural Information Processing Systems* (S. 8026–8037). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3454287.3455008

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Derczynski, L., Du, X., Grella, M., Gv, K., He, X., Hou, H., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., . . . Zhu, R.-J. (2023). RWKV: Reinventing RNNs for the Transformer Era. In H. Bouamor, J. Pino & K. Bali (Hrsg.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (S. 14048–14077). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.936

Pennington, J., Socher, R. & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang & W. Daelemans (Hrsg.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (S. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In M. Walker, H. Ji & A. Stent (Hrsg.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (S. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

Python Software Foundation (Hrsg.). (o. D.). *Python Language Reference, version 3.12*. Abgerufen am 17.01.2025, von http://www.python.org.

R Core Team (Hrsg.). (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Abgerufen am 17.01.2025, von https://www.R-project.org.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Rao, D. & McMahan, B. (2019). *Natural Language Processing with PyTorch Build Intelligent Language Applications Using Deep Learning*. O'Reilly Media, Inc.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Reimer, S. & Tepner, O. (2022). Reflexion videobasierter Erkläreinheiten. In S. Habig & H. van Vorst (Hrsg.), *Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen, Tagungsband der GDCP Jahrestagung 2021* (S. 688–691). Gesellschaft für Didaktik der Chemie und Physik.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In S. Padó & R. Huang (Hrsg.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (S. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Reinhold, P., Riese, J. & Gramzow, Y. (2017). Fachdidaktisches Wissen im Lehramtsstudium Physik. In H. Fischler & E. Sumfleth (Hrsg.), *Professionelle Kompetenz von Lehrkräften der Chemie und Physik* (S. 39–56). Logos Verlag.

Riese, J. (2009). Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 97). Logos Verlag.

Riese, J., Gramzow, Y. & Reinhold, P. (2017). Die Messung fachdidaktischen Wissens bei Anfängern und Fortgeschrittenen im Lehramtsstudiengang Physik. *Zeitschrift für Didaktik der Naturwissenschaften*, *23*, 99–112. https://doi.org/10.1007/s40573-017-0059-2

Riese, J., Kulgemeyer, C., Zander, S., Borowski, A., Fischer, H. E., Gramzow, Y., Reinhold, P., Schecker, H. & Tomczyszyn, E. (2015). Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik. *Zeitschrift für Pädagogik*, *61*, 55–79.

Riese, J. & Reinhold, P. (2012). Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen. *Zeitschrift für Erziehungswissenschaften*, *15*, 111–143. https://doi.org/10.1007/s11618-012-0259-y

Riese, J., Schröder, J. & Vogelsang, C. (2022a). Die Entwicklung physikdidaktischen Wissens im Längsschnitt. In S. Habig & H. van Vorst (Hrsg.), *Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen, Tagungsband der GDCP Jahrestagung 2021.* Gesellschaft für Didaktik der Chemie und Physik.

Riese, J., Vogelsang, C., Schröder, J., Borowski, A., Kulgemeyer, C., Reinhold, P. & Schecker, H. (2022b). Entwicklung von Unterrichtsplanungsfähigkeit im Fach Physik: Welchen Einfluss hat Professionswissen? *Zeitschrift für Erziehungswissenschaft*, *25*, 843–867. https://doi.org/10.1007/s11618-022-01112-0

Roberts, M. E., Stewart, B. M. & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684

Roberts, M. E., Stewart, B. M. & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(1), 1–40. https://doi.org/10.18637/jss.v091.i02

Roberts, M. E., Stewart, B. M., Tingley, D. & Benoit Kenneth. (2023). *Package 'stm'. Estimation of the Structural Topic Model (v1.3.6.1)*. Abgerufen am 17.01.2025, von https://doi.org/10.32614/CRAN.package.stm.

Robitzsch, A., Kiefer, T. & Wu, M. (2024). *TAM: Test Analysis Modules*. Abgerufen am 17.01.2025, von https://doi.org/10.32614/CRAN.package.TAM.

Rosenberg, J. M. & Krist, C. (2021). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology*, *30*(2), 255–267. https://doi.org/10.1007/s10956-020-09862-4

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536. https://doi.org/10.1038/323533a0

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., . . . Rush, A. M. (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations (ICLR 2022)*. https://openreview.net/forum?id=9Vrb9D0WI4

Sawatzki, J., Schlippe, T. & Benner-Wickner, M. (2022). Deep Learning Techniques for Automatic Short Answer Grading: Predicting Scores for English and German Answers. In E. C. K. Cheng, R. B. Koul, T. Wang & X. Yu (Hrsg.), *Artificial Intelligence in Education: Emerging Technologies, Models and Applications. Proceedings of 2021 2nd International Conference on Artificial Intelligence in Education Technology* (S. 65–75). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-7527-0_5

Schaper, N. (2014). Validitätsaspekte von Kompetenzmodellen und -tests für hochschulische Kompetenzdomänen. In F. Musekamp & G. Spöttl (Hrsg.), *Berufliche Bildung in Forschung, Schule und Arbeitswelt: Bd. 12. Kompetenz im Studium und in der Arbeitswelt. Nationale und internationale Ansätze zur Erfassung von Ingenieurkompetenzen.* (S. 21–48). Lang. https://doi.org/10.25656/01:12883

Schecker, H., Wilhelm, T., Hopf, M. & Duit, R. (Hrsg.). (2018). *Schülervorstellungen und Physikunterricht*. Springer Spektrum. https://doi.org/10.1007/978-3-662-57270-2

Schiering, D. (2021). *Was wirkt in der universitären Physik-Lehramtsausbildung? Determinanten in der Entwicklung des Professionswissens angehender Physiklehrkräfte*. Christian-Albrechts-Universität zu Kiel. urn:nbn:de:gbv:8:3-2022-00229-5

Schiering, D., Sorge, S., Keller, M. M. & Neumann, K. (2023). A proficiency model for pre-service physics teachers' pedagogical content knowledge (PCK)—What constitutes high-level PCK? *Journal of Research in Science Teaching*, *60*(1), 136–163. https://doi.org/10.1002/tea.21793

Schiering, D., Sorge, S. & Neumann, K. (2021). Hilft viel viel? Der Einfluss von Studienstrukturen auf das Professionswissen angehender Physiklehrkräfte. *Zeitschrift für Erziehungswissenschaft*, *24*, 545–570. https://doi.org/10.1007/s11618-021-01003-w

Schiering, D., Sorge, S., Petersen, S. & Neumann, K. (2019). Konstruktion eines qualitativen Niveaumodells im fachdidaktischen Wissen von angehenden Physiklehrkräften. *Zeitschrift für Didaktik der Naturwissenschaften*, *25*, 211–229. https://doi.org/10.1007/s40573-019-00100-y

Schmelzing, S. (2010). *Das fachdidaktische Wissen von Biologielehrkräften: Konzeptionalisierung, Diagnostik, Struktur und Entwicklung im Rahmen der Biologielehrerbildung*. Logos Verlag Berlin.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen. Untersuchungen zur Kohärenzbildung beim Wissenserwerb mit Texten*. *Fortschritte der psychologischen Forschung. 20*. Beltz.

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*, 103–122. https://doi.org/10.1007/s40573-020-00115-w

She, J., Chan, K. K. H., Wang, J., Hu, X. & Liu, E. (2024). Effect of Science Teachers' Pedagogical Content Knowledge on Student Achievement: Evidence From Both Text- and Video-Based Pedagogical Content Knowledge Tests. *American Educational Research Journal.* https://doi.org/10.3102/00028312241278627

Sherin, B. (2013). A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. *Journal of the Learning Sciences*, *22*(4), 600–638. https://doi.org/10.1080/10508406.2013.836654

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), 1–23. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Sorge, S., Keller, M. M., Petersen, S. & Neumann, K. (2018). Die Entwicklung des Professionswissens angehender Physiklehrkräfte. In C. Maurer (Hrsg.), *Qualitätsvoller Chemie- und Physikunterricht - normative und empirische Dimensionen, Tagungsband der GDCP Jahrestagung 2017* (S. 114–117). Gesellschaft für Didaktik der Chemie und Physik.

Sorge, S., Kröger, J., Petersen, S. & Neuman, K. (2019). Structure and development of pre-service physics teachers' professional knowledge. *International Journal of Science Education*, *41*(7), 862–889. https://doi.org/10.1080/09500693.2017.1346326

Spurk, D., Hirschi, A., Wang, M., Valero, D. & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445

Steinke, I. (1999). *Kriterien qualitativer Forschung. Ansätze zur Bewertung qualitativ-empirischer Sozialforschung.* Juventa.

Strübe, M. (2020). *Modelle und Experimente im Chemieunterricht: Eine Videostudie zum fachspezifischen Lehrerwissen und -handeln* [Dissertation]. Universität Duisburg-Essen.

Sundararajan, M., Taly, A. & Yan, Q. (2017). Axiomatic attribution for deep networks. In D. Precup & Y. W. Teh (Hrsg.), *ICML'17, ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70* (S. 3319–3328). JMLR.org. https://dl.acm.org/doi/10.5555/3305890.3306024

Tepner, O., Borowski, A., Dollny, S., Fischer, H. E., Jüttner, M., Kirschner, S., Leutner, D., Neuhaus, B. J., Sandmann, A., Sumfleth, E., Thillmann, H. & Wirth, J. (2012). Modell zur Entwicklung von Testitems zur Erfassung des Professionswissens von Lehrkräften in den Naturwissenschaften. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 7–28. https://doi.org/10.25656/01:31747

Tepner, O. & Dollny, S. (2014). Measuring Chemistry Teachers' Content Knowledge: Is It Correlated to Pedagogical Content Knowledge? In C. Bruguière, A. Tiberghien & P. Clément (Hrsg.), *Contributions from Science Education Research. Topics and Trends in Current Science Education* (Bd. 1, S. 243–254). Springer Netherlands. https://doi.org/10.1007/978-94-007-7281-6_15

Terhart, E. (2012). Wie wirkt Lehrerbildung? Forschungsprobleme und Gestaltungsfragen. *Zeitschrift für Bildungsforschung*, *2*(1), 3–21. https://doi.org/10.1007/s35834-012-0027-3

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. & Lample, G. (2023a). *LLaMA: Open and Efficient Foundation Language Models*. arXiv:2302.13971v1

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., . . . Scialom, T. (2023b). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv:2307.09288v2

Tröger, H., Sumfleth, E. & Tepner, O. (2017). Chemistry Teachers' Professional Knowledge, Classroom Action, and Students' Learning: The Relevance of Technical Language. In K. Hahl, K. Juuti, J. Lampiselkä, A. Uitto & J. Lavonen (Hrsg.), *Contributions from Science Education Research. Cognitive and Affective Aspects in Science Education Research* (Bd. 3, S. 207–218). Springer International Publishing. https://doi.org/10.1007/978-3-319-58685-4_16

Tschisgale, P., Wulff, P. & Kubsch, M. (2023). Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review in Physics Education Research*, *19*, 20123. https://doi.org/10.1103/PhysRevPhysEducRes.19.020123

Tschisgale, P., Wulff, P. & Kubsch, M. (2025). Erratum: Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory [Phys. Rev. Phys. Educ. Res. 19 020123 (2023)]. *Physical Review Physics Education Research*, *21*, 19901. https://doi.org/10.1103/PhysRevPhysEducRes.21.019901

van Dusen, B., Vogelsang, C., Taylor, J. & Cauet, E. (2021). How to Teach a Teacher: Challenges and Opportunities in Physics Teacher Education in Germany and the USA. In H. E. Fischer & R. Girwidz (Hrsg.), *Physics Education* (S. 55–81). Springer International Publishing. https://doi.org/10.1007/978-3-030-87391-2_3

Varshney, K. R. (2019). Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, *25*(3), 26–29. https://doi.org/10.1145/3313109

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vogelsang, C., Borowski, A., Buschhüter, D., Enkrott, P., Kempin, M., Kulgemeyer, C., Reinhold, P., Riese, J., Schecker, H. & Schröder, J. (2019). Entwicklung von Professionswissen und Unterrichtsperformanz im Lehramtsstudium Physik-Analysen zu valider Testwertinterpretation. *Zeitschrift für Pädagogik*, *65*(4), 473–491. https://doi.org/10.25656/01:23990

Vogelsang, C., Borowski, A., Kulgemeyer, C. & Riese, J. (2018). Profile-P+ - Entwicklung von Kompetenz und Performanz im Physiklehramt. In C. Maurer (Hrsg.), *Qualitätsvoller Chemie- und Physikunterricht - normative und empirische Dimensionen, Tagungsband der GDCP Jahrestagung 2017* (Vol. 38, S. 867–870). Gesellschaft für Didaktik der Chemie und Physik.

Vogelsang, C., Kulgemeyer, C. & Riese, J. (2022). Learning to Plan by Learning to Reflect? - Exploring Relations between Professional Knowledge, Reflection Skills, and Planning Skills of Preservice Physics Teachers in a One-Semester Field Experience. *Education Sciences*, *12*(7). https://doi.org/10.3390/educsci12070479

Vollmer, H. J. & Klette, K. (2023). Pedagogical Content Knowledge and Subject Didactics - An Intercontinental Dialogue? In F. Ligozat, K. Klette & J. Almqvist (Hrsg.), *Didactics in a Changing World: European Perspectives on Teaching, Learning and the Curriculum* (S. 17–33). Springer International Publishing. https://doi.org/10.1007/978-3-031-20810-2_2

Voss, T., Kunina-Habenicht, O., Hoehne, V. & Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften: Empirische Zugänge und Befunde. *Zeitschrift für Erziehungswissenschaft*, *60*(2), 184–201. https://doi.org/10.1007/s11618-015-0626-6

Webb, G. I., Lee, L. K., Goethals, B. & Petitjean, F. (2018). Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, *32*(5), 1179–1199. https://doi.org/10.1007/s10618-018-0554-1

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. & Le, Q. V. (2022). Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations (ICLR 2022)*. https://openreview.net/forum?id=gEZrGCozdqR

Woitkowski, D. (2015). Fachliches Wissen Physik in der Hochschulausbildung: Konzeptualisierung, Messung, Niveaubildung. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 185). Logos Verlag.

Woitkowski, D. (2019). Erfolgreicher Wissenserwerb im ersten Semester Physik. *Zeitschrift für Didaktik der Naturwissenschaften*, *25*, 97–114. https://doi.org/10.1007/s40573-019-00094-7

Woitkowski, D. (2020). Tracing physics content knowledge gains using content complexity levels. *International Journal of Science Education*, *42*(10), 1585–1608. https://doi.org/10.1080/09500693.2020.1772520

Woitkowski, D. & Riese, J. (2017). Kriterienorientierte Konstruktion eines Kompetenzniveaumodells im physikalischen Fachwissen. *Zeitschrift für Didaktik der Naturwissenschaften*, *23*, 39–52. https://doi.org/10.1007/s40573-016-0054-z

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. von, Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., . . . Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Q. Liu & D. Schlangen (Hrsg.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (S. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wright, B. D. (2003). Rack and Stack: Time 1 vs. Time 2 or Pre-Test vs. Post-Test. *Rasch Measurement Transactions*, *17*(1), 905–906. https://www.rasch.org/rmt/rmt171a.htm

Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A. & Borowski, A. (2022). Bridging the Gap Between Qualitative and Quantitative Assessment in Science Education Research with Machine Learning - A Case for Pretrained Language Models-Based Clustering. *Journal of Science Education and Technology*, *31*, 490–513. https://doi.org/10.1007/s10956-022-09969-w

Wulff, P., Buschhüter, D., Westphal, A., Nowak, A., Becker, L., Robalino, H., Stede, M. & Borowski, A. (2021). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. *Journal of Science Education and Technology*, *30*, 1–15. https://doi.org/10.1007/s10956-020-09865-1

Wulff, P., Mientus, L., Nowak, A. & Borowski, A. (2023). Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections. *International Journal of Artificial Intelligence in Education*, *33*, 439–466. https://doi.org/10.1007/s40593-022-00290-6

Yan, D., Rupp, A. A. & Foltz, P. W. (2020). *Handbook of Automated Scoring: Theory into Practice*. CRC Press.

Zehner, F., Sälzer, C. & Goldhammer, F. (2016). Automatic Coding of Short Text Responses via Clustering in Educational Assessment. *Educational and Psychological Measurement*, *76*(2), 280–303. https://doi.org/10.1177/0013164415590022

Zeller, J., Jordans, M. & Riese, J. (2022). Ansätze zur Ermittlung von Kompetenzniveaus im Fachdidaktischen Wissen. In S. Habig & H. van Vorst (Hrsg.), *Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen, Tagungsband der GDCP Jahrestagung 2021* (S. 768–771). Gesellschaft für Didaktik der Chemie und Physik.

Zeller, J. & Riese, J. (2023). Datenbasierte Fähigkeitsprofile im Physikdidaktischen Wissen. In H. van Vorst (Hrsg.), *Lernen, Lehren und Forschen in einer digital geprägten Welt, Tagungsband der GDCP Jahrestagung 2022* (S. 794–797). Gesellschaft für Didaktik der Chemie und Physik.

Zeller, J. & Riese, J. (2024). Fähigkeitsprofile im Physikdidaktischen Wissen mithilfe von Machine Learning. In H. van Vorst (Hrsg.), *Frühe naturwissenschaftliche Bildung, Tagungsband der GDCP Jahrestagung 2023* (S. 122–125). Gesellschaft für Didaktik der Chemie und Physik.

Zeller, J. & Riese, J. (2025). Competency profiles of PCK using unsupervised learning: What implications for the structures of pPCK emerge from non-hierarchical analyses? *Journal of Research in Science Teaching.* https://doi.org/10.1002/tea.70001

Zeller, J., Schiering, D., Kulgemeyer, C., Neumann, K., Riese, J. & Sorge, S. (2024). Empirisch-kriterienorientierte Analyse des fachdidaktischen Wissens angehender Physiklehrkräfte. Welche inhaltlichen Strukturen zeigen sich über unterschiedliche Projekte hinweg? *Unterrichtswissenschaft.* https://doi.org/10.1007/s42010-024-00200-w

Zesch, T., Horbach, A. & Zehner, F. (2023). To Score or Not to Score: Factors Influencing Performance and Feasibility of Automatic Content Scoring of Text Responses. *Educational Measurement: Issues and Practice*, *42*(1), 44–58. https://doi.org/10.1111/emip.12544

Zhai, S., Talbott, W., Srivastava, N., Huang, C., Goh, H., Zhang, R. & Susskind, J. (2021a). *An Attention Free Transformer*. arXiv:2105.14103v2

Zhai, X. (2021). Practices and Theories: How Can Machine Learning Assist in Innovative Assessment Practices in Science Education. *Journal of Science Education and Technology*, *30*, 139–149. https://doi.org/10.1007/s10956-021-09901-8

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H. & Urban-Lurain, M. (2020a). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, *57*, 1430–1459. https://doi.org/10.1002/tea.21658

Zhai, X., Neumann, K. & Krajcik, J. (2023). Editorial: AI for tackling STEM education challenges. *Frontiers in Education*, *8*, 1183030. https://doi.org/10.3389/feduc.2023.1183030

Zhai, X., Shi, L. & Nehm, R. H. (2021b). A Meta-Analysis of Machine Learning-Based Science Assessments: Factors Impacting Machine-Human Score Agreements. *Journal of Science Education and Technology*, *30*, 361–379. https://doi.org/10.1007/s10956-020-09875-z

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C. & Shi, L. (2020b). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, *56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

Zhu, R.-J., Zhang, Y., Sifferman, E., Sheaves, T., Wang, Y., Richmond, D., Zhou, P. & Eshraghian, J. K. (2024). *Scalable MatMul-free Language Modeling*. arXiv:2406.02528v5

Anhang

Abbildungen

Tabellen

A. Beispielhafte Herleitung einer Loss Funktion

Die folgende an den Ausführungen von Murphy (2022) orientierte Betrachtung ist vereinfacht dargestellt. Es liege ein Datensatz $(x_i, y_i), i = 1 \dots N$ vor, bei dem angenommen wird, dass die abhängige Variable $Y = (Y_1, \dots Y_N)$ der Gauß- bzw. Normalverteilung

$$Y \mid W \sim \mathcal{N}(f_W(x), \sigma^2 \mathbf{1})$$

mit einer (stetigen) Funktion f_W folgt, wobei die Funktionsparameter $W = (W_1, \dots, W_M)$ selbst einer (noch nicht näher bestimmten) Wahrscheinlichkeitsverteilung folgen. Hier stellen zudem x den Vektor $(x_1, \dots, x_N)$ und $\mathbf{1}$ die Einheitsmatrix dar. Die Wahrscheinlichkeitsverteilung ist dann für $y = (y_1, \dots, y_N)$ gegeben durch

$$P(Y = y \mid W = w) = \frac{1}{(2\pi)^{N/2}\sigma} \exp\left(-\frac{\sum_{i=1}^{N}\left(y_i - f_w(x_i)\right)^2}{2\sigma^2}\right).$$

Man nennt in diesem Wahrscheinlichkeitsverteilung $P(Y \mid W)$ auch *Likelihood*. Um nun aus den vorhandenen (x_i, y_i)-Daten die *wahrscheinlichsten Funktionsparameter* $\widehat{w}$, die zu diesen Daten passen, muss nach dem Satz von Bayes

$$P(W = w \mid Y = y) = \frac{P(\,Y = y \mid W = w\,) \cdot P(W = w)}{P(Y = y)}$$

in w maximiert werden. Diese Wahrscheinlichkeit wird auch die *A-Posteriori*-Wahrscheinlichkeit (engl. „*Posterior"*) genannt. Den Ansatz nennt man dementsprechend auch die *Maximum-a-posteriori* Schätzung (MAP). Es gilt also:

$$\widehat{w} = \underset{w}{\operatorname{argmax}}\, P(W = w \mid Y = y) = \underset{w}{\operatorname{argmax}}\, P(\,Y = y \mid W = w\,) \cdot P(W = w)\,,$$

wobei der Nenner aus der vorherigen Gleichung nicht relevant ist, da er nicht von w abhängt. Mit dem üblichen Trick (Logarithmus ist strikt monoton wachsend),

$$\underset{w}{\operatorname{argmax}}\, f(w) = \underset{w}{\operatorname{argmin}}\, (-f(w)) = \underset{w}{\operatorname{argmin}}\, (-\log f(w))\,,$$

kann man dies umformen zu:

$$\widehat{w} = \underset{w}{\operatorname{argmin}}\, (-\log P(\,Y = y \mid W = w\,) - \log\, P(W = w))\,.$$

Für den zweiten Summanden ist nun eine Annahme über die *A-Priori*-Verteilung (engl. „*Prior"*) der Funktionsparameter W notwendig. Exemplarisch wird hier angenommen, dass W ebenfalls normalverteilt ist mit

$$W \sim\, \mathcal{N}(0, \tau^2 \mathbf{1})\,.$$

Durch Einsetzen der Wahrscheinlichkeitsverteilung der Gaußverteilung ergibt sich mit den Regeln für den Logarithmus und nach Wegstreichen von bezüglich w konstanter Terme:

$$\widehat{w} = \underset{w}{\operatorname{argmin}} \left(\sum_{i=1}^{N}\left(y_i - f_w(x_i)\right)^2 + \frac{\sigma^2}{\tau^2}\sum_{j=1}^{M} w_j^2\right).$$

Der erste Term ist leicht zu erkennen als die klassische (vermeintlich heuristische) Least-Squares Loss-Funktion, die auch in der linearen Regression, d. h. für $f_w(x_i) = w \cdot x_i$, zum Einsatz kommt. Der Zweite Term entspricht dem sog. Ridge-Regularisierungsterm (Murphy, 2022; Géron, 2019), der dafür sorgt, dass die Parameter w in der Optimierung eine „Tendenz in Richtung 0“ erhalten. Der Regularisierungsterm fällt weg, wenn angenommen wird, dass W gleichverteilt ist. Alternativ kann auch direkt $P(Y \mid W)$ optimiert werden, dann ändert sich aber die Interpretation der gefundenen $\hat{w}$: Anstelle der bei gegebenem Datensatz wahrscheinlichsten Parameter findet man dann die Parameter, unter deren Annahme die Wahrscheinlichkeit, die erhaltenen Daten zu observieren, maximal ist. Diese Option nennt man auch die *Maximum Likelihood Schätzung* (MLE), die vielen algorithmischen Modellen zugrunde liegt, bei denen keine Annahmen über die Verteilung der Parameter W mit in die Modellierung einfließen (Murphy, 2022). An den zwei Formulierungen zur Interpretation der $\hat{w}$ im MAP- bzw. MLE-Ansatz erkennt man zudem leicht die Vorteile der Interpretierbarkeit des probabilistischen MAP-Ansatzes, die in Abschnitt 2.4 schon angedeutet wurden.

B. Handreichung zur Einordnung der FDW-Testaufgaben

Handreichung zur Zuordnung der FDW-Testaufgaben zu Stufen hierarchischer Komplexität

Schließen sich gegenseitig aus!

Fakten (I)

- Reproduktion einzelner, unverbundener Informationen
- Keine oder kaum Bezugnahme auf Situation oder sonstige Beschreibung
- Keine oder kaum Verknüpfung der genannten Informationen
- *Beispiel:* Nennen von Fakten zu einem Fachdidaktischen Konzept

Einstufige Kausalität (II)

- Verknüpfung von zwei oder mehr Fakten, Informationen oder Äußerungen zu einem Produkt (z. B. Schlussfolgerungen, Argumentationen)
- Begründungen, Analysen und Argumentationen mir nur einer Argumentations- / Analysestufe
- *Beispiel:* (einstufige) Analyse oder Evaluation einer Situation

Mehrstufige Kausalität (III)

- Begründungen, Argumentationen, Evaluationen mit mehr als einer Argumentations- / Analysestufe
- Alle Anforderungen, die komplexere Analysen / Argumentation verlangen als II
- *Beispiel:* Analyse und Evaluation einer Situation

Handreichung zur Zuordnung der FDW-Testaufgaben zu Anforderungskategorien

Kognitive Prozesse nach Anderson und Krathwohl (2001)

Mehrfachnennung möglich!

Erinnern:

- Etwas wiederzuerkennen oder abzurufen und dies nennen bzw. wiederzugeben, ist Kernbestandteil der Aufgabe.
- Weite Teile der Aufgabe sollten allein durch Erinnern an Fachdidaktische Inhalte lösbar sein.
- Es wird nach „typischen“ Aspekten (z. B. Schülervorstellungen) gefragt, was impliziert, dass es um konsens-Wissen geht, welches explizit in Lehrveranstaltungen erworben werden kann.
- *Beispiel:* Fakten zu bestimmten Fachdidaktischen Konzepten nennen

- *Gegenbeispiel:* Eine Schüleräußerung wird betrachtet.

Verstehen:

- Ein Element Fachdidaktischen Wissens verstanden zu haben, bedeutet, dieses Element beschreiben, klassifizieren, vergleichen und erklären zu können, bzw. es in ein Begriffsnetz einordnen zu können.
- Eine Aufgabe wird der Dimension „Verstehen“ zugeordnet, wenn diese Fähigkeiten / Kompetenzen die Bearbeitung der Aufgabe vereinfachen.
- Weite Teile der Aufgabe sollten allein durch das Verstehen Fachdidaktischer Inhalte lösbar sein, insbesondere ohne die Konzepte bereits auf Situationen übertragen zu müssen.
- *Beispiel:* Die Funktionen von Unterrichtselementen (z. B. Einleitung, Sicherung, Experimentieren) erleichtert deren Auflistung.
- *Gegenbeispiel:* Eine Situation oder ein konkreter Gegenstand wird betrachtet.

Anwenden:

- Fachdidaktisches Wissen, ein Verfahren oder eine Prozedur anzuwenden oder zu ermitteln, wann die Anwendung einer Prozedur legitim ist, ist Kernbestandteil der Aufgabe.
- Konstruktion / geeignete Auswahl von physikalischen Beispielen zu gegebenen Fragestellungen.
- *Beispiel:* Prognostizieren von typischen Fehlern mithilfe von Wissen über Schülervorstellungen
- *Gegenbeispiel:* Analyse eines exemplarischen Unterrichtsmaterials

Analysieren:

- Einen Aspekt, eine Situation, eine Äußerung zu analysieren, ist Kernbestandteil der Aufgabe und / oder eine Analyse wird explizit in der Aufgabenstellung eingefordert.
- *Beispiel:* Rekonstruktion von Schülervorstellungen aus Äußerungen
- *Gegenbeispiel:* Auswahl eines geeigneten Beispiels zur Vermittlung eines Fachinhalts

Evaluieren:

- Qualitätsurteile über fachdidaktisch relevante Elemente (z. B. Handlungen, Material, etc.) auf Basis von Kriterien und Standards bzw. des Wissens treffen, d. h. zu überprüfen und kritisieren, ist Kernbestandteil der Aufgabe.
- Auch die Begründung eines (möglicherweise vorgegebenen) Qualitätsurteil fällt unter diese Kategorie.
- Dabei liegt der Fokus auf der Evaluation von fachdidaktisch relevanten Elementen und nicht der Evaluation von Fachwissen beispielsweise in Schüler:innenäußerungen.
- *Beispiel:* Ein beschriebenes Vorgehen einer Lehrkraft bewerten / kommentieren
- *Gegenbeispiel:* Eigenes Vorgehen wird begründet

Kreieren:

- Selbst auf Basis einer Situation oder Beschreibung Elemente fachdidaktisch relevanter Handlungen oder vollständige fachdidaktisch relevante Handlungsketten zu kreieren, ist Kernbestandteil der Aufgabe.
- *Beispiel:* Selbst eine Lösungsstrategie entwickeln oder Alltagsbeispiele unter konkreten Zielsetzungen begründet auswählen

Zusätzliche Dimensionen

Mehrfachnennung möglich!

Notwendigkeit des Einbezugs von Fachwissen:

- Zur Lösung der Aufgabe ist verstärkt explizites physikalisches Fachwissen notwendig.

Bezug auf ein Beispiel:

- Ein physikalisches Beispiel kann ein Alltagsbeispiel, ein Beispielexperiment etc. sein
- Die Aufgabe beinhaltet die Beschreibung oder Betrachtung eines Beispiels entweder durch den / die Probandin selbst oder die Betrachtung eines Beispiels (z. B. in einer Unterrichtsvignette) ist wesentlicher Teil der Aufgabe

Bezug auf Unterrichtssituation:

- Die Aufgabe bezieht sich auf Elemente konkreter Unterrichtssituationen, die in Stamm der Aufgabe beschrieben wird.
- Die Aufgabe muss sich mindestens auf konkrete Handlungen / Äußerungen von Schüler:innen und / oder Lehrkräften beziehen.

Kommentare

- Bei Erinnern und Verstehen genügt es nicht, dass eines der beiden notwendige Voraussetzung zur Bearbeitung der Aufgabe ist
- Betrachtet werden die Kategorien unter der Annahme, dass ein:r durchschnittliche:r Studierende:r die Aufgabe bearbeitet und entsprechende Lehrveranstaltungen bereits besucht hat.

Literatur

Anderson, L. W. & Krathwohl, D. R. (Hrsg.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (4. Aufl.). New York: Longman.

C. Material zur Bepunktung des Testinstruments

In den folgenden zwei Tabellen sind zusätzliche Informationen zur Bepunktung des Testinstruments exemplarisch als Supplement zu Artikel 2 dargestellt.

Table A1 Scoring rubric for task 15 of the pPCK-test instrument used for the analysis. This the scoring rubric for the task presented in Figure 5.3; translated from German to English. The task is scored dichotomous, i.e., zero or one point. The example responses stem from real samples or expert interviews of the test instrument's piloting phase (Gramzow, 2015).

Expectation correct	Expectation incorrect
The channel no longer hinders the ball, and the centrifugal force can therefore continue to act outwards. Only the centrifugal force exists, only the channel hinders it, e.g. *"Student imagines that the centrifugal force drives the ball away from the center."* The centrifugal force must be named or described. The resulting movement of the ball must be made clear. Edge cases: *"Ball seeks compensation for the constraint of the trajectory curve"* *"The ball contains a twist in the trajectory and therefore rolls outwards"* *"The student believes that a force is acting outwards on the ball at point R. If the channel is no longer there to hold the ball in place, the ball must move to the right."*	Answers that do not refer to a content matter concept, e.g.: *"Student thinks there is a repulsion from the center"* *"Student has not understood the principle of centripetal force"* *"The mass has no energy directed eastwards, as it exits R vertically. But the student assumes that it does, because the ball moves in an easterly direction from the start to point R."* All kinds of responses that do not describe what the student does not understand.

Table A2 Further exemplary responses to task 15 of the used pPCK test instrument. Further exemplary responses from the dataset for the task presented in Figure 5.3; translated from German to English.

Correct responses (1 point)	Incorrect responses (0 points)
"The path is not left tangentially / the ball is "pushed" outwards"	*"That the ball does not follow the curve and that the ball can simply change direction"*
"The assumption that there is a force driving the ball outwards (as seen on the circular path)."	*"- Objects always have a rightward velocity - Objects always "fall" downwards in a curve"*
"This is based on the idea that a so-called centrifugal force exists. The student assumes that the ball must therefore move away from the center."	*"The centripetal force causes the ball to fly outwards. Inertia is not applied."*

D. Entwicklungsverläufe bezüglich der LPA-Kompetenzprofile

Die folgenden beiden Abbildungen visualisieren Entwicklungsverläufe bezüglich der in der LPA (Artikel 3 / Kapitel 6) ermittelten Kompetenzprofile. In den Blöcken zu den jeweiligen Messzeitpunkten ist die Anzahl der Proband:innen im entsprechenden Kompetenzprofil abgebildet. Auf Basis dieser Datenlage sind verlässliche Aussagen über Systematiken nicht gerechtfertigt. Interaktive Versionen dieser Abbildungen sind auch im digitalen Ergänzungsmaterial enthalten.

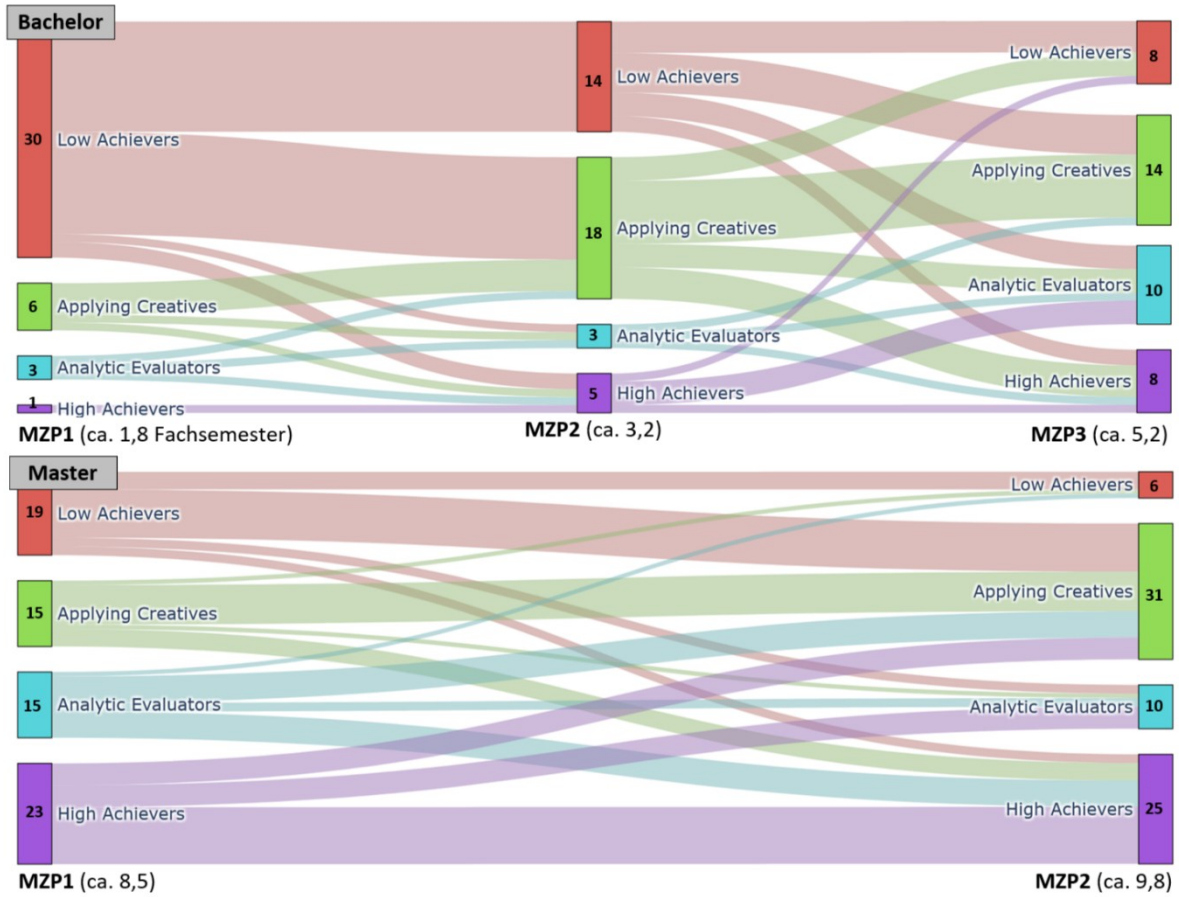


Abbildung A1 Sankey Plots der Bachelor- und Master-Proband:innen ohne Dropout. Bemerkenswert ist insbesondere, dass die Master-MZPs vor und nach dem Praxissemester lagen. Damit erscheint der Zuwachs an Applying Creatives plausibel. Gleichzeitig ist die Instabilität insbesondere des „High Achievers" Profils für einen nachhaltigen Kompetenzerwerb sicherlich suboptimal.

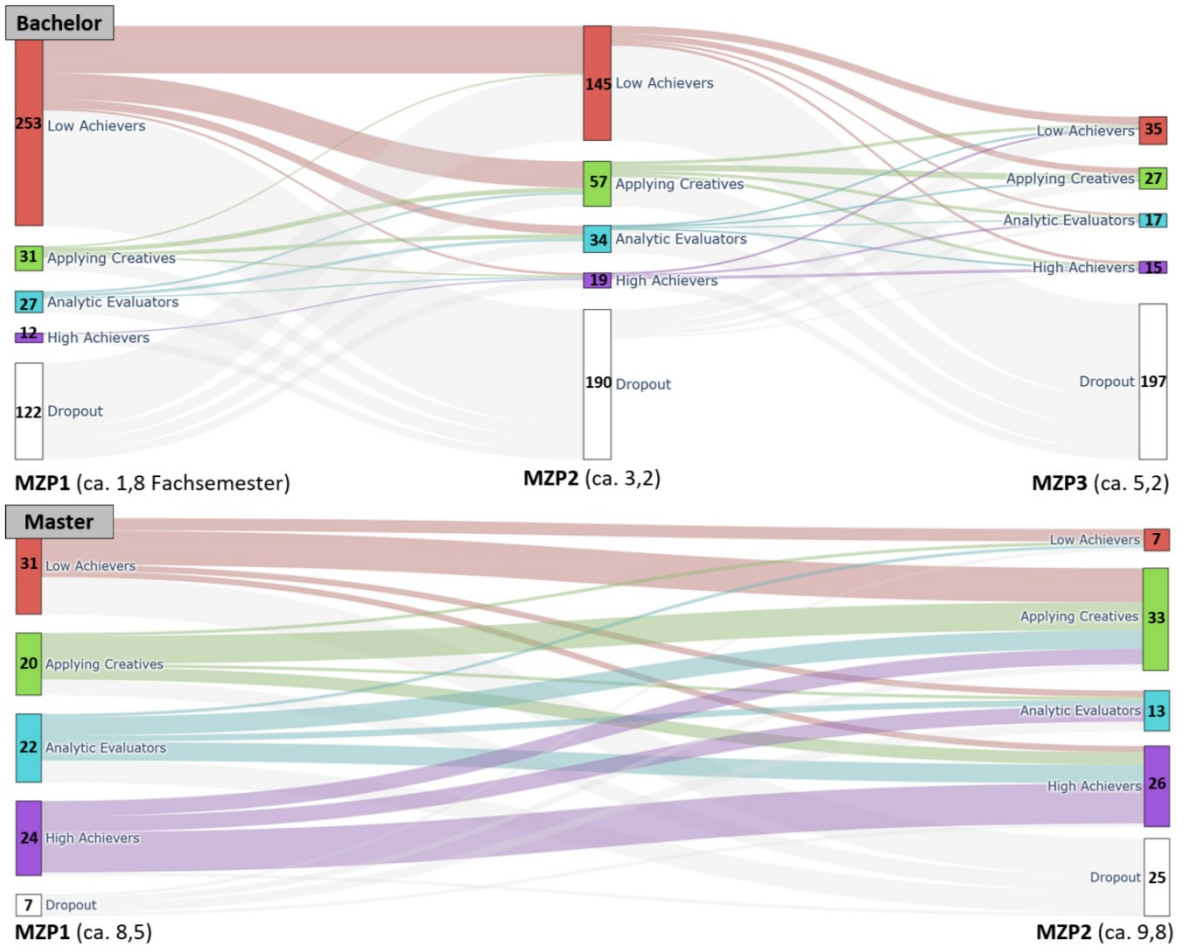


Abbildung A2 Sankey Plots der Bachelor- und Master-Proband:innen mit Dropout.

E. Zusätzliche Analysen zu den LPA-Kompetenzprofilen

Die folgenden Tabellen zeigt zusätzliche Analysen zur Beschreibung der latenten Kompetenzprofile aus Artikel 3.

Tabelle A3 Vergleich der latenten Kompetenzprofile in Hinsicht auf Fachsemester, FDW-Gesamtscore, Demographischer Daten, Umfang und Durchschnittscores in den kognitiven Anforderungsdimensionen. In den „T-Test"- und „Cohens d"-Zeilen werden entsprechende Vergleiche zwischen dem FDW-Profil und dem nächst „niedrigerem" Kompetenzprofil berichtet. N_{tot} bezeichnet die Anzahl an Personen im Cluster bzgl. des Gesamtdatensatzes, d. h., wenn vormals ausgeschlossene Proband:innen (siehe Abschnitt 6.4.4) nachträglich zugeordnet werden. Im Sinne einer besseren Übersicht, werden hier alle anderen Werte als Relativwerte bezogen auf die in der ursprünglichen LPA einbezogenen 785 Proband:innen dargestellt.

		Low-Achievers	**Applying Creatives**	**Analytic Evaluators**	**High-Achievers**
$\boldsymbol{N / N_{\mathrm{tot}}}$		411 / 470	166 / 167	112 / 113	96 / 96
T-Test-Freiheitsgrade		-	575	276	206
Anteil weiblich		34 %	34 %	25 %	43 %
Schulabschlussnote		2,33	2,07	2,07	1,83
Fachsemester Physik	*M*	2,87	5,31	5,28	6,96
	SD	2,56	3,59	3,92	3,71
	T-Test	-	$T = 9{,}2$ $p < 0{,}001$	$T = 0{,}1$ $p = 0.94$	$T = 3{,}2$ $p = 0.002$
	Cohens d	-	0,84	-	0,44
FDW-Gesamt	*M*	0,22	0,41	0,43	0,58
	SD	3,27	3,28	3,67	3,765
	T-Test	-	$T = 25{,}6$ $p < 0{,}001$	$T = 1{,}34$ $p = 0.18$	$T = 12{,}9$ $p < 0{,}001$
	Cohens d	-	2,36	-	1,79
Reproduzieren	*M*	0,31	0,53	0,52	0,69
	SD	0,14	0,15	0,18	0,16
	T-Test	-	$T = 17{,}1$ $p < 0{,}001$	$T = 0{,}4$ $p = 0{,}67$	$T = 7{,}1$ $p < 0{,}001$
	Cohens d	-	1,58	-	0,99
FDW: Anwenden-Kreieren	*M*	0,18	0,51	0,32	0,67
	SD	0,12	0,11	0,10	0,13
	T-Test	-	$T = 30{,}9$ $p < 0{,}001$	$T = -14{,}4$ $p < 0{,}001$	$T = 21{,}8$ $p < 0{,}001$
	Cohens d	-	2,84	$-1{,}77$	3,03
FDW: Analysieren-Evaluieren	*M*	0,21	0,32	0,57	0,64
	SD	0,12	0,10	0,10	0,13
	T-Test	-	$T = 11{,}2$ $p < 0{,}001$	$T = 20{,}5$ $p < 0{,}001$	$T = 4{,}5$ $p < 0{,}001$
	Cohens d	-	1,03	2,50	0,62

Tabelle A4 Durchschnittliche Scores der latenten Kompetenzprofile bezüglich anderer ProfiLe-P+ - Tests. Im Sinne einer besseren Übersicht, werden hier alle Werte als Relativwerte bezogen auf die in der ursprünglichen LPA einbezogenen 785 Proband:innen dargestellt (siehe Abschnitt 6.4.2). Weitere Informationen zu den Testinstrumenten sind aus den entsprechenden Quellen zu entnehmen (für eine Übersicht siehe auch Vogelsang et al., 2018): Mathematisches Wissen (MaW) bei Riese et al. (2015) – Fachwissen (FW) bei Enkrott (2021) – Pädagogisches Wissen (PW) bei Riese (2009)

		Low Achievers	Applying Creatives	Analytic Evaluators	High Achievers
N		411	166	112	96
PW	*M*	0,83	0,87	0,88	0,94
	SD	0,13	0,16	0,15	0,21
	T-Test	-	$T = 3{,}0$ $p = 0{,}003$	$T = 0{,}8$ $p = 0.43$	$T = 2{,}0$ $p = 0.044$
	Freiheitsgrade	-	575	276	206
	Cohens d	-	0,42	-	0,28
FW	*M*	0,55	0,59	0,59	0,61
	SD	0,09	0,06	0,06	0,05
	T-Test	-	$T = 4{,}5$ $p < 0{,}001$	$T = 0{,}2$ $p = 0.85$	$T = 1{,}9$ $p = 0.07$
	Freiheitsgrade	-	449	221	153
	Cohens d	-	0,63	-	-
MaW	*M*	0,48	0,66	0,63	0,70
	SD	0,23	0,21	0,24	0,22
	T-Test	-	$T = 7{,}7$ $p < 0{,}001$	$T = 1{,}0$ $p = 0.33$	$T = 1{,}9$ $p = 0.06$
	Freiheitsgrade	-	449	221	153
	Cohens d	-	1,07	-	-

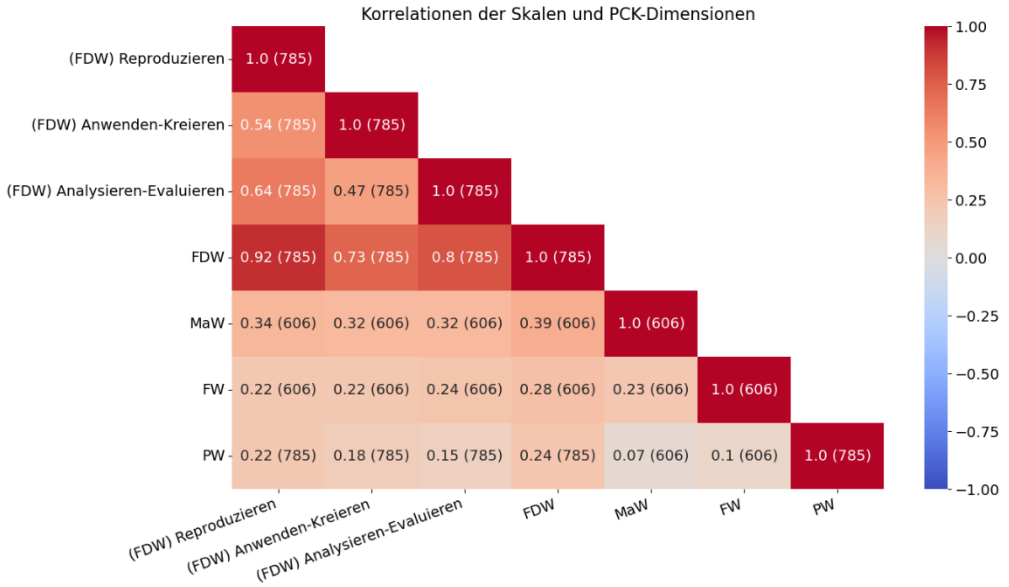


Abbildung A3 Heatmap der Korrelationen der FDW-Skalen und weiterer Professionswissensdimensionen. Die Abkürzungen zu den einzelnen Konstrukten werden in Tabelle A4 erläutert.

Tabelle A5 Genese der LPA-Topics aus den Wortlisten des STMs

Topic	Fachkonzepte	Wissensch. Arbeitsweisen	Schüler-vorstellungen	Symbolische Sprache	Beispiele	Begründung von Beispielen	*Kurzinterpretation / Topic Titel*
1	bewegung, kräfte, impuls, vakuum, system	vernachlässigung, idealfall, auswertungen	vorstellung		beispielsweise, schaukeln, kreisel	medial	*Allgemeine Konzepte I*
2	geschwindigkeit				auto, berg, fährt, kinder, fahrend	anschauungsmaterial	*Nutzung von Beispielen*
3	bergriff, gerundet, grundkenntniss, unterrichtsstoff	eigenständig, ergbenis	vorstellung, verstanden		beispiel, zug, rampe		*Allgemeine Konzepte II*
4	actio	aufbereiten		*v, f, ~, 0, t2, s, p, a*	apfel, bergig		*Symbolische Beschreibung*
5	Bewegung, kräfte	arbeitsweisen, erkenntnisgewinnung	schülervorstellungen, kognitiven, alltagserfahrungen, begriff, konflikt, alltagsbezüge, begriffwechsel, sachstruktur		körper, beispiel, lampe, kabel		*Schülervorstellungen*
6	actio, reactio, wirkt				Auto, z. B., fahrer, billiardkugeln	veranschaulichung, denkanstöße, vollziehen	*Begründung von Beispielen*

F. Zusätzliche Analysen zum Automatisierten Assessment

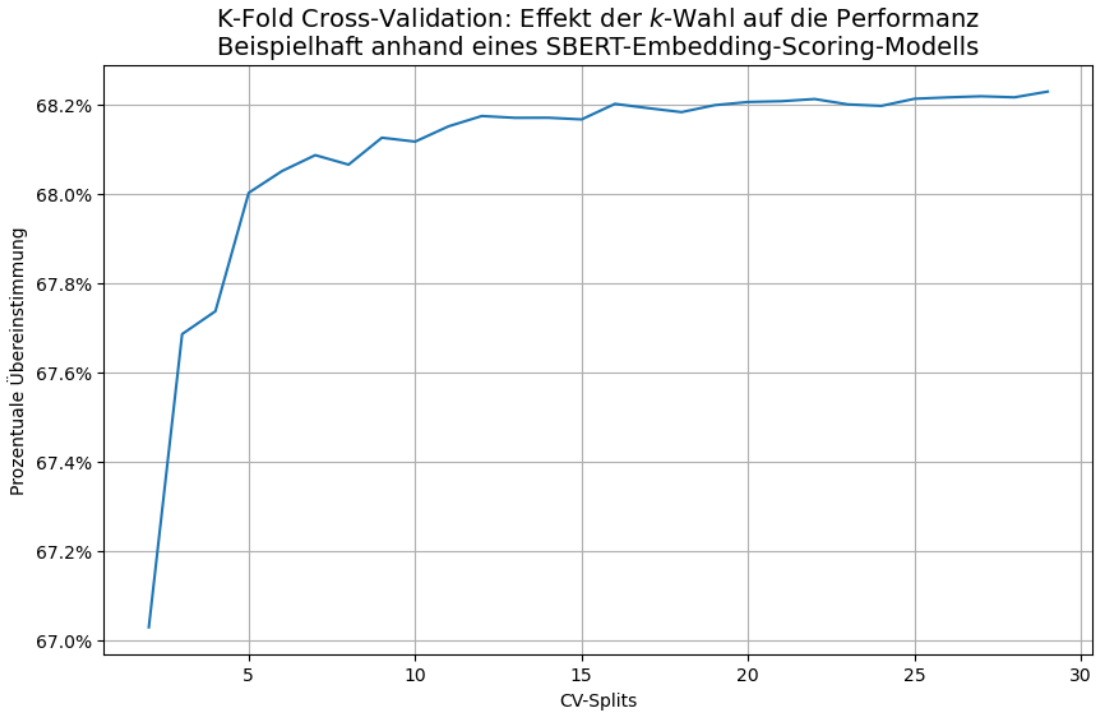


Abbildung A4 Auswirkungen der Anzahl an CV-Splits auf die Performanz-Schätzung. Hier wurde exemplarisch ein Scoring-Modell auf Basis der SBERT-Embeddings (siehe Abschnitt 6.7.6) trainiert. Man erkennt deutlich, wie die erhaltene Schätzung für die prozentuale Übereinstimmung zwischen den Score-Vorhersagen (Maschine) und Score-Labels (Mensch) mit zunehmender Anzahl an CV-Splits zunimmt und sich (unter statistischem Rauschen) asymptotisch einem Maximum annähert. Der Unterschied in der Performanz-Schätzung zwischen nur 2 und 30 CV-Splits beträgt allerdings gerade einmal 1,2 %. Das Kosten-Nutzen-Verhältnis bei der Nutzung sehr vieler CV-Splits ist also begrenzt.

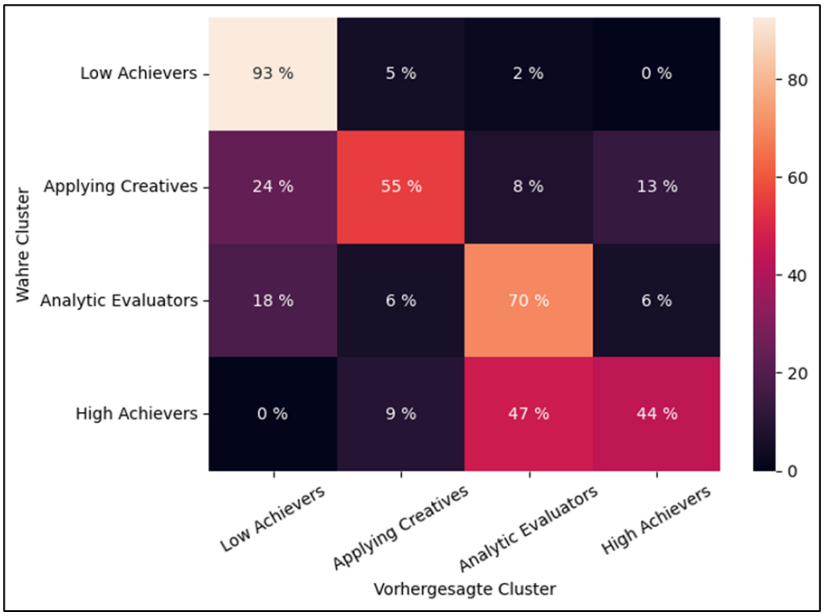


Abbildung A5 Darstellung der Mensch-Mensch-Übereinstimmungen der Kompetenzprofil-Zuordnung als Heatmap. Im Vergleich mit Abbildung 6.10 fällt auf, dass trotz der höheren Gesamtübereinstimmung die Mensch-Mensch-Übereinstimmung punktuell deutlich geringer ausfällt und ungleichmäßiger ist.

Zero-Shot Prompt mit Kodiermanual für Aufgabe 1a.

Leerzeilen und Formatierungen sind nur für die Darstellung hier eingefügt und haben keine Bedeutung für den tatsächlichen Prompt. Die Hashtag-Symbole („#“) dienen der Formatierung des Prompts auf eine Weise, die auch dem Sprachmodell zugänglich ist (sog. *Markdown-Syntax*).

```
Die Aufgabe, die bepunktet werden soll, lautet wie folgt:


<aufgabe>

Ein Lehrer hat das Wechselwirkungsprinzip "Actio=Reactio" (3. Newtonsches
Axiom) in einer 9. Klasse eingeführt. Nachfolgend spielt sich folgende Szene
ab.

Lehrer: Stellt euch jetzt einmal vor, ein Apfel hängt an einem Baum. Wo haben
   wir hier jetzt Actio und Reactio?

Schüler A: Na, ist doch klar, der Apfel zieht am Ast und der Ast hält den
   Apfel oben!

Die Klasse signalisiert Zustimmung

Lehrer: Ja richtig – schön, ihr habt es verstanden! Was ist denn dann, wenn
   der Apfel jetztherunterfällt? Also während des Fallens, wo ist da Actio
   und Reactio?

Ein Gemurmel stellt sich ein

Schüler B: Ja gilt das denn dann überhaupt noch? Ich meine, ist doch immer
   nur ideal, dass das gilt?!?

Schüler A: Klar hast du noch Actio und Reactio, nur Actio wird halt immer
   größer, der Apfel wird ja schließlich schneller beim Fallen!

Schüler B: Ich dachte, die müssen gleich sein? Wo willst du überhaupt Reactio
   haben, der fällt doch frei und wird nicht mehr gehalten!?!

Schüler A: Hm. Na Actio hast du auf jeden Fall schon mal, er bewegt sich ja.
   Und er wird ja auch nicht beliebig schnell, die Luftreibung bremst ihn
   ja. Das ist deine Reactio!

a) Offensichtlich haben die Schüler die Ausführungen des Lehrers nicht
richtig verstanden, die Übertragung auf die Situation mit dem frei fallenden
Apfel gelingt nicht. Analysieren Sie die Szene: Inwiefern ist das Vorgehen
des Lehrers nicht optimal?

</aufgabe>


Dann folgt ein Textfeld für die mögliche Antwort.

Der Erwartungshorizont für diese Aufgabe sieht wie folgt aus:


<erwartungshorizont>


Kodierung: Dichotom kodieren (0 oder 1 Punkt)

Ziel: Nennung von Aspekten, die die Lehrkraft in fachlicher Hinsicht nicht
optimal umgesetzt hat


## Erwartungshorizont korrekt
```

Fachlicher Bezug wichtig

Vermischungsaspekt:

Der Lehrer begeht einen Planungsfehler. Er verwirrt die Schüler, da er verschiedene physikalische Konzepte / Probleme verknüpft und vermischt: Zunächst handelt es sich um eine mechanische Wechselwirkung, im Fallen ist jedoch kein Angriffspunkt zu erkennen, da die Gewichtskraft berührungslos / gravitativ wirkt. Weiterhin handelt es sich zunächst um ein statisches Kräftegleichgewicht (1. Newtonsches Axiom), beim Fallen handelt es sich um einen dynamischen Fall (2. Newtonsches Axiom), d.h. es wird nicht zwischen Kräftegleichgewicht und Wechselwirkungsprinzip unterschieden.

Überforderungsaspekt:

Es wird zu schnell abstrahiert bei dieser komplexen Thematik, die Schüler sind überfordert. Mit einer mündlichen Erklärung allein ist ein solches Konzept nicht einzuführen, es wäre eine Begleitung – etwa durch ein Tafelbild mit eingezeichneten Kräften – nötig, in der die auftretenden Kräfte veranschaulicht werden.

Erwartungshorizont inkorrekt

Es werden Aussagen gemacht, die sich nicht auf das fachliche Problem beziehen, das der Lehrer durch seine Vorgehensweise provoziert

- z.B. "Der Lehrer überschätzt die Schüler."

Aussagen, die sich lediglich darauf beziehen, dass die Erklärung des Lehrers nicht ausreicht oder besser sein müsste, reichen nicht aus.

Aussagen, die sich lediglich darauf beziehen, dass die Erklärung des Lehrers nicht ausreicht oder besser sein müsste, reichen nicht aus.

Es werden Aussagen gemacht, die keinen direkten Bezug zur dargestellten Unterrichts-situation haben oder die physikdidaktisch nicht zutreffend sind.

- z.B. mangelnde Gesprächsführung)
- z.B. "Der Lehrer unterbricht die Schüler zu spät"
- "Der Lehrer gibt keine Hilfestellungen"

Es werden Aussagen gemacht, die nicht den Kern des Problems treffen.

- allg. Aussage zu mangelndem Verständnis
- z. B. "Übergeneralisierung nach Aussage von Schüler A. Vielleicht hat es die ganze Klasse eben nicht verstanden."
- "Das Prinzip actio=reactio wurde nicht richtig verstanden."

</erwartungshorizont>

Ordne die folgende Antwort zu deiner Aufgabe des Fragebogens gemäß diesem obigen Erwartungshorizont ein:

<scores>

[0, 1]

</scores>

Hier ist die Antwort des Probanden:

```
<antwort>
{response (hier wird die Antwort eingefügt)}
</antwort>

Antworte in folgendem Format:

<score>
Die Punktzahl folgt hier als Integer-Zahl.
</score>
```

Zero-Shot Prompt mit Kodiermanual für Aufgabe 3.

Leerzeilen und Formatierungen sind nur für die Darstellung hier eingefügt und haben keine Bedeutung für den tatsächlichen Prompt. Die Hashtag-Symbole („#“) dienen der Formatierung des Prompts auf eine Weise, die auch dem Sprachmodell zugänglich ist (sog. *Markdown-Syntax*).

```
Die Aufgabe, die bepunktet werden soll, lautet wie folgt:

<aufgabe>

Das Experiment spielt im Physikunterricht eine zentrale Rolle.
Nennen Sie bitte zwei verschiedene Ziele bzw. Funktionen des Experiments im
Physikunterricht.

</aufgabe>

Dann folgen zwei Textfelder für die möglichen Antworten.
Der Erwartungshorizont für diese Aufgabe sieht wie folgt aus:

<erwartungshorizont>

Kodierung: 1 Punkt pro richtige Funktion (max. 2 Punkte)
0 Punkte für gar keine Funktion

## Erwartungshorizont korrekt

### Pädagogische Funktion
Es trägt zur Bildung der Schüler bei, indem sie kausales und funktionales
Denken, Kreativität fördern

### Lernpsychologische Funktion:
  - Experimente motivieren, wecken Interesse, machen das Lernen erfahrbar.
    Grenzfall: "Gemeinschaftliches Event, hebt sich ab vom Lernalltag"
  - Physik in Technik und Alltag aufzeigen
  - Motivation durch kognitive Konflikte
  - „mehrkanaligen“ Zugang
  - Selbsttätigkeit
```

- mögliche Individualisierung
- Förderung des Selbstwertgefühls

Erkenntnistheoretische Funktion:

Das Experiment ist Methode der Erkenntnisgewinnung in der Physik

Überprüfung von physikalischen Gesetzen, Modellen

Fachliche Funktion:

Experimente visualisieren/veranschaulichen physikalische Sachverhalte, unterstützen die Bildung von Begriffen, Überführung von Theorie und Praxis ineinander, z. B.

- "Zum Erarbeiten eines physikalischen Konzepts.", oder
- "Praktische Anwendung von Modellen"

Praktische Funktion:

- Schüler üben den Umgang mit Messdaten, deren Auswertung, mit dem Umgang von Messgeräten
- Verantwortlicher Umgang
- Grenzfall: "Sorgfältiges Arbeiten lernen"
- experimentelle Kompetenzen erwerben

Leistungsbeurteilung:

Leistungen von Schülern im Rahmen einer experimentellen Aufgabe überprüfen

Soziale Kompetenzen

Kooperationsfähigkeit Kommunikationsfähigkeit

Methodologische Funktion:

Experiment als Lerninhalt, naturwissenschaftliche Arbeitsweisen (z.B. auch Beobachtung)

Sonstige Mögliche Funktionen:

- (Schüler-)Vorstellungen prüfen
- Handlungskompetenz erlernen
- Kritik- und Reflexionsfähigkeit
- Meilensteine unserer Kulturgeschichte aufzeigen

Erwartungshorizont inkorrekt

- "Abwechslung"
- "Experimente führen zu besserem Verständnis“ oder „Verständnis"
- Ziel oder Funktion im Unterricht ist die Abwechslung

Es werden keine oder Antworten gegeben, die sich keinem der fünf Bereiche des korrekten

Erwartungshorizontes zuordnen lassen,

- z.B. Experimente haben keine unterrichtliche Funktion
- Experimente müssen vorkommen aufgrund des Lehrplans
- Experimente machen den Unterricht zeitökonomischer
- "Durch Experimente behält man das Wissen eher im Kopf."

</erwartungshorizont>

Ordne die folgende Antwort zu deiner Aufgabe des Fragebogens gemäß diesem obigen Erwartungshorizont ein:

<scores>

[0, 1, 2]

</scores>

Hier ist die Antwort des Probanden:

<antwort>

{response (hier wird die Antwort eingefügt)}

</antwort>

Antworte in folgendem Format:

<score>

Die Punktzahl folgt hier als Integer-Zahl.

</score>

G. Auszüge aus der Dokumentation des Analysecodes

Teil des digitalen Begleitmaterials ist unter anderem auch die Dokumentation des verwendeten Analysecodes, der in Form eines Python-Pakets (ca. 13,000 Zeilen Code) strukturiert ist. Diese Dokumentation wurde mit dem Tool MkDocs (https://www.mkdocs.org/) erstellt und lässt sich im Webbrowser betrachten. Sie ist als Teil eines Open-Source-Projekts zur fortgeführten Nutzung des Codes gedacht und befindet sich somit unter aktiver Entwicklung. Sie hat zum Zeitpunkt der Fertigstellung dieser Arbeit noch nicht den Anspruch, jede Funktionalität bis ins Detail ausführlich zu erläutern. Die Darstellung hier dient der Illustration der Vision für die Fortführung der methodischen und technischen Ansätze des Projekts. Der Code ist auch in einer etwas entschlackten Version als Open-Source Projekt online verfügbar (https://github.com/JannisZeller/questionnaire-tools).

Die Willkommensseite der Dokumentation des Analysecodes sieht wie folgt aus:

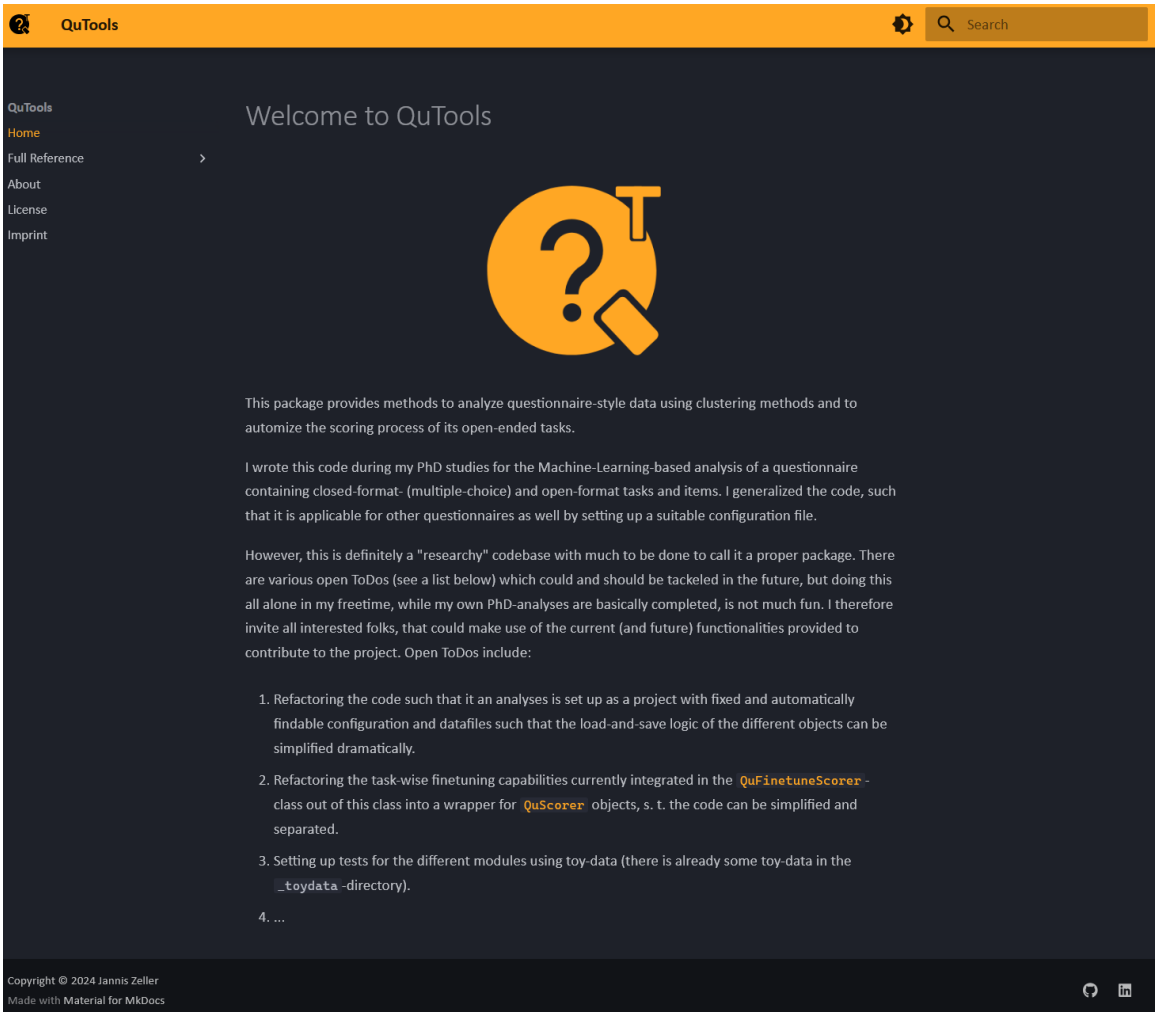


Abbildung A6 Willkommensseite der Dokumentation des Analysecodes.

Auf der linken Seite lassen sich dann die entsprechenden Seiten für einzelne Pakete unter „Full Reference“ öffnen, wobei sowohl Beispielcode…

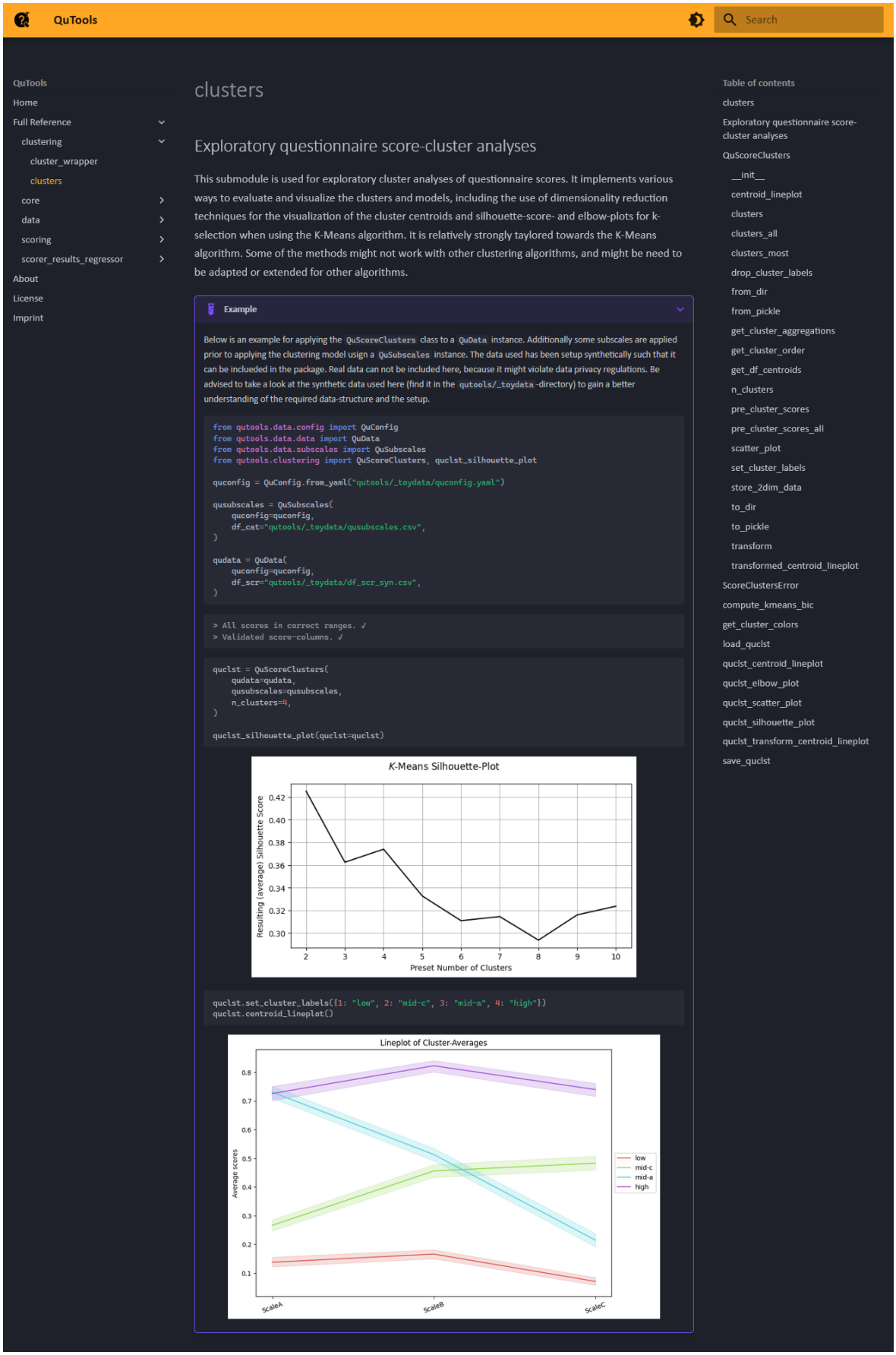


Abbildung A7 Beispielseite der Dokumentation des Analysecodes für Cluster-Analysen.

…als auch die Implementierung und Nutzung der einzelnen Methoden beschrieben werden.

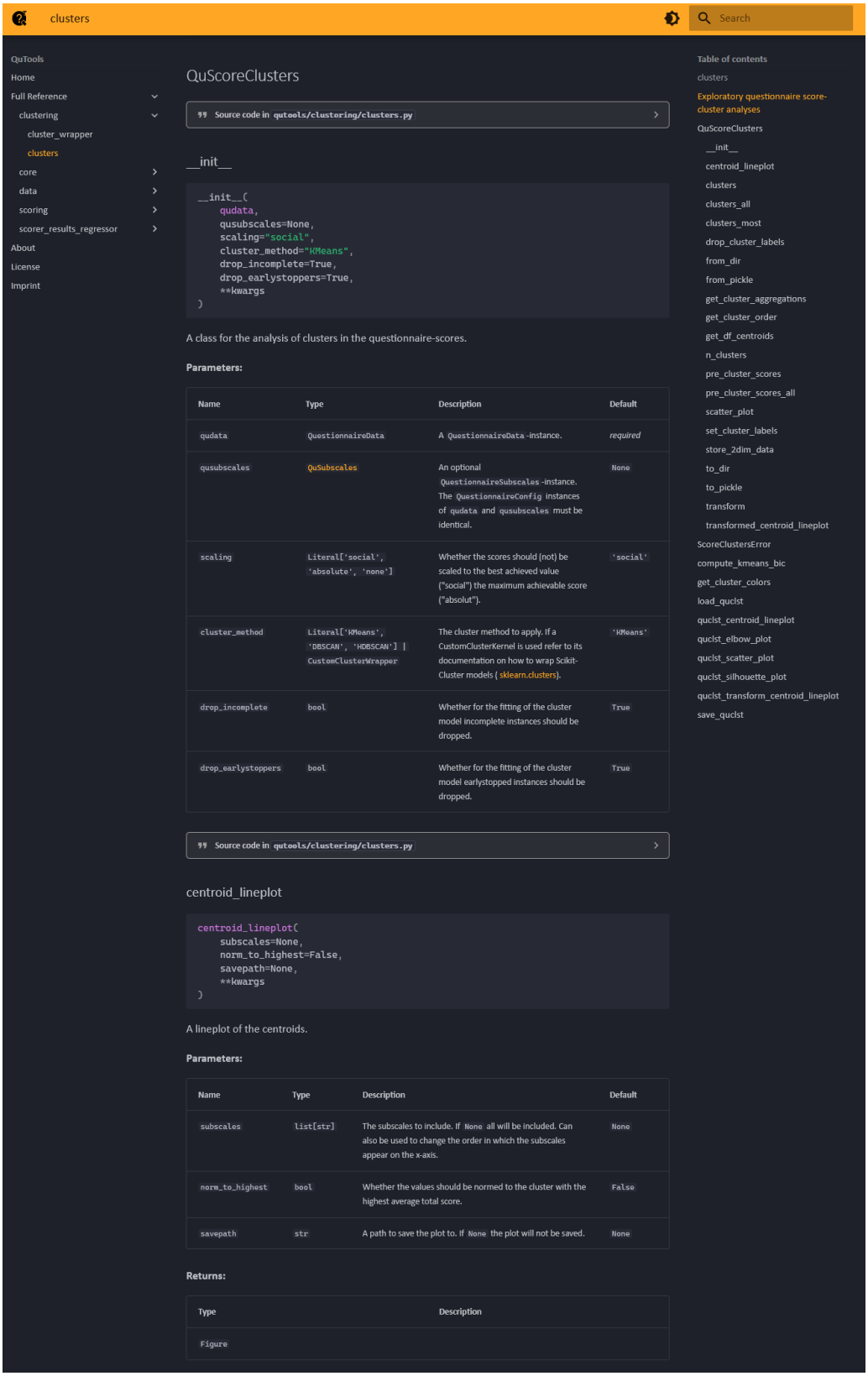


Abbildung A8 Beispielansicht der Einzelbeschreibungen der Methoden des Analysecodes.

H. Auszüge aus der Webumgebung für das Assessment

Im Folgenden werden einige Eindrücke des Webtools, welches als Proof of Concept zur Anwendung des Assessment Workflows in einem realen Assessment Setting entworfen wurde, dargestellt. Ähnlich wie auch das Paket für die Nachnutzung des Analysecodes (Anhang G) versteht sich auch das Webtool als Projekt unter aktiver Entwicklung und hat somit noch nicht den Anspruch einsatzbereit für ein reales Deployment zu sein. Es ist aktuell als Open-Source-Projekt (ca. 4,000 Zeilen Python-Code sowie einige hundert Zeilen JavaScript- und einige tausend Zeilen HTML-Code) unter https://github.com/JannisZeller/questionnaire-webtool hinterlegt.

Die Willkommensseite des Webtools sieht wie folgt aus:

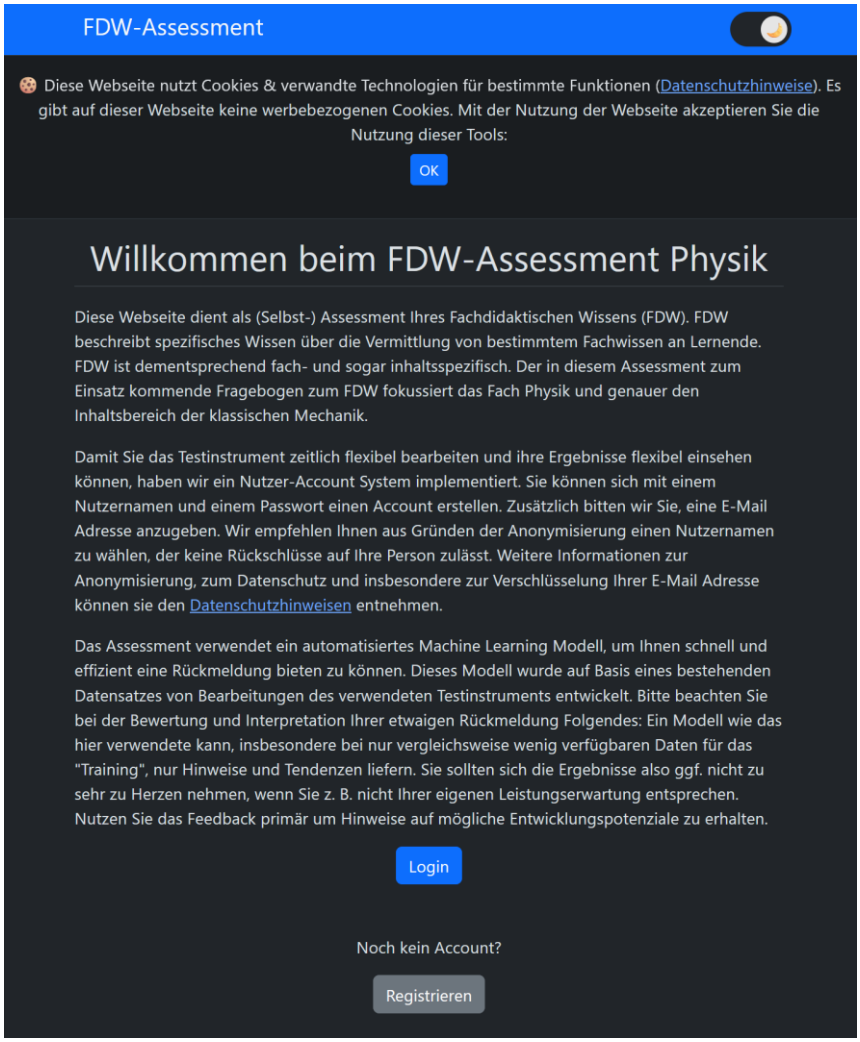


Abbildung A9 Willkommensseite des Assessment-Webtools. Dargestellt ist insbesondere das Cookie-Banner sowie die Login- und Registrierungsoptionen.

Nachdem die Cookies etc. akzeptiert sind kann man einen Account erstellen und sich einloggen. Die Bearbeitung des Testinstruments ist dann mithilfe von Eingabemasken wie der Folgenden möglich:

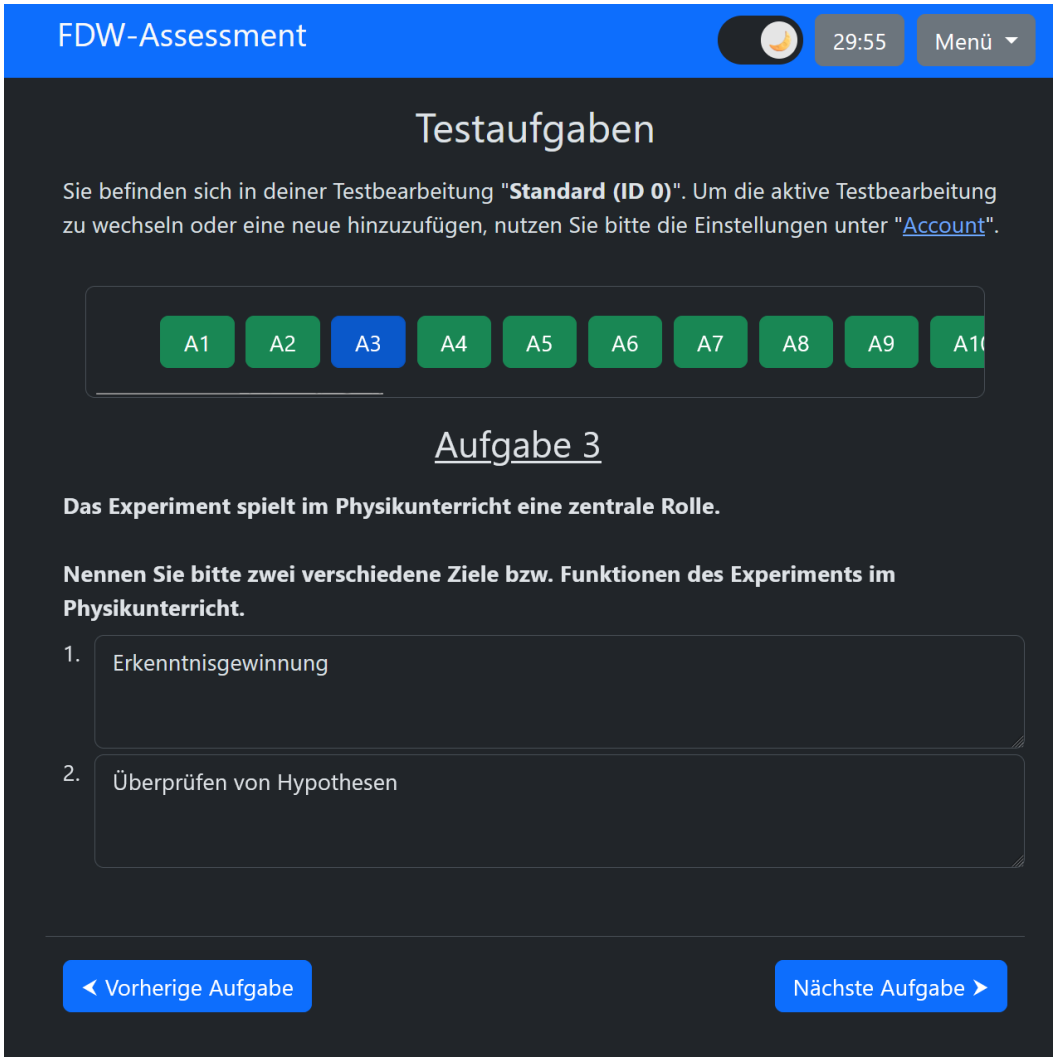


Abbildung A10 Beispielansicht einer Testaufgabe im Assessment-Webtool. Eingaben werden automatisch mit dem Server synchronisiert, um bei Verbindungsproblemen immer minimalen Datenverlust zu gewährleisten. Das Testinstrument kann zudem teilweise bearbeitet und zu einem späteren Zeitpunkt fortgesetzt werden.

Dabei werden alle Eingaben automatisch gespeichert. Im Menü unter „Report“ kann man nach der Bearbeitung des Tests dann eine Anfrage zur Erstellung eines Reports an den Server senden, wo dann basierend auf den Modellen aus Kapitel 6 eine automatisierte Bepunktung etc. stattfindet. Das Ergebnis hat dann die folgende Form:

Abbildung A11 Beispielhafte Darstellung des Ergebnisses eines automatischen Assessments. Diese Aufführungen sind eher als Platzhalter zu sehen. Zukünftig sind hier wahrscheinlich stärker inhaltliche Aussagen sinnvoller. Die Entwicklung und Evaluierung solcher finaler Gestaltungsfragen bezüglich des Feedbacks sind allerdings nicht mehr Teil dieses Dissertationsprojekts.

Bisher handelt es sich bei den dargestellten Informationen primär um Platzhalter, die Illustrieren, wozu das Modell grundsätzlich im Stande ist. Für reale Anwendungen wäre hier eine eher „prosaische Darstellung“ für eine Rückmeldung an die Proband:innen wahrscheinlich hilfreicher.