Evaluating Work: Verbal and Numerical Assessments Across Organizational Contexts

Der Fakultät für Wirtschaftswissenschaften

der Universität Paderborn

zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften

- Doctor rerum politicarum -

vorgelegte Dissertation

von

Jana Kim Gutt, M.A.

geboren am 02.09.1991

in Paderborn

Erscheinungsjahr 2025

Table of Contents

List of Papers

Paper 1

Gutt, Jana Kim*, & Thommes, Kirsten* (2025). Evaluating Competencies with Spoken Comments and Machine Learning. Working Papers Dissertations 130, Paderborn University, Faculty of Business Administration and Economics. https://ideas.repec.org/p/pdn/dispap/130.html

Paper 2

Gutt, Jana Kim* (2025). Evaluators' Consideration of Warmth and Competence in Verbal and Numerical Performance Assessments. Working Papers Dissertations 131, Paderborn University, Faculty of Business Administration and Economics. https://ideas.repec.org/p/pdn/dispap/131.html

Paper 3

Gutt, Jana Kim*, Thommes, Kirsten*, & Mehic, Miro* (2025). Can Verbal Performance Appraisals and Machine Learning Models Improve the Accuracy of Performance Evaluations? Working Papers Dissertations 132, Paderborn University, Faculty of Business Administration and Economics. https://ideas.repec.org/p/pdn/dispap/132.html

Paper 4

Gutt, Jana Kim*, & Knorr, Karin* (2025). Factors Influencing Organizations' Responses on Employer Review Platforms. Working Papers Dissertations 133, Paderborn University, Faculty of Business Administration and Economics. https://ideas.repec.org/p/pdn/dispap/133.html

Note:

* Paderborn University, Faculty of Business Administration and Economics, Chair of Organizational Behavior, Paderborn, Germany

Synopsis

1 Introduction

In order to survive, humans have been evolutionarily driven to make constant assessments of their surroundings (Kahneman, 2011). Beyond survival and safety, assessments extend to everyday situations, such as deciding whom to approach for directions (Ambady et al., 1995) or estimating how long a task will take to complete (Wiese et al., 2016). Evaluating information and forming beliefs are fundamental aspects of human thinking (Griffin & Tversky, 1992).

One area where assessments are routinely required is the workplace. Business decisions, ranging from entering a market to launching a product, demand a systematic assessment of the available information to determine a course of action (Moore & Cain, 2007). In many cases, however, the availability of information is limited, requiring organizations to rely on individuals' subjective judgment to make accurate estimations (Sniezek & Henry, 1990). For instance, due to the lack of objective measures, organizations often depend on supervisory assessments when evaluating employees' competence and respective job performance (Kusterer & Sliwka, 2024).[1]

Individuals make judgments based on relevant information they observe (Brashier & Marsh, 2020). Yet, a key challenge in subjective assessments is that humans are considered imperfect judges (Kahneman et al., 2021). As individuals are also not considered "optimal cognitive processors of observations" (Feiler et al., 2013, p. 575), they may overemphasize some observations and neglect others when forming their judgment (Morewedge & Kahneman,

[1] To clarify the terminology used in this dissertation, I adopt an input-based approach to competence, meaning that competence is understood as a set of underlying knowledge, skills, or abilities, in contrast to an output-based approach, which defines competencies as specific behaviors that contribute to achieving desired results or outcomes (e.g., Hoffmann, 1999; Kurz & Bartram, 2002). Organizations often adopt an output-oriented approach (e.g., Campbell & Wiernik, 2015). Accordingly, when conducting performance appraisals, the assessment may focus on competence (e.g., communication skills) but is based on observable behavior (e.g., the employee's actual communication).

2010). This challenge is particularly evident in supervisory performance assessments (e.g., Demeré et al., 2019; Prendergast & Topel, 1993; Stauffer & Buckley, 2005).

When assessing performance, social dynamics and human cognition inevitably become influencing factors in the evaluation process (Spence & Keeping, 2011). Since performance assessments serve as a crucial basis for organizational decisions, errors in judgment become particularly problematic (Adler et al., 2016). These errors have far-reaching implications at multiple levels, affecting the labor market, organization, and individual employee. In the labor market, errors in judgment can lead to inefficient talent allocation; in organizations, they may result in poor decision-making and reduced productivity; and for individual employees, they can cause unfair career outcomes, lower motivation, and job dissatisfaction. As a result, considerable research has focused on assessment decisions and their accuracy, particularly the extent to which numerical ratings capture employee performance (e.g., Bernardin et al., 2016; Decotiis & Petit, 1978; Mero & Motowidlo, 1995).

According to Funder (1987), studying rating accuracy and human error is a remarkable task as it relies on the assumption that the true state of affairs can be determined with certainty. Consequently, evaluating the extent to which ratings deviate from the truth becomes inherently challenging. In some instances, identifying the true state of affairs is relatively straightforward. The weather, for example, confirms or disproves the weather forecaster's predictions, serving as the apparent truth (Griffin & Tversky, 1992; Murphy & Winkler, 1977). In other contexts, determining the truth proves more challenging. For instance, while sentencing guidelines for judges exist, there is no apparent true value against which criminal sentences can be measured (Kahneman et al., 2021). Similarly, there is no apparent truth to validate performance ratings.

Over the years, substantial research has focused on the quality of performance ratings by estimating various types of true values. For example, inter- and intra-rater reliability (e.g., Mero et al., 2003; Viswesvaran et al., 1996; Yun et al., 2005), sales productivity (Sundvik &

Lindeman, 1998), and the number of correct entries in a text-entering-task (Kusterer & Sliwka, 2024) have been used to approximate a true value against which performance ratings can be compared. In this context, it is important to note that if a true, observable value capturing all dimensions of work performance existed, subjective ratings and research on rating accuracy would no longer be necessary (Demeré et al., 2019). This further reflects the inherent complexity of work performance, as tasks in modern work environments encompass multiple dimensions that make it difficult to capture all relevant aspects within a single measure. For example, production workers may be expected to maintain high production output while also ensuring the proper functioning of their machines (Holmstrom & Milgrom, 1991). Beyond their core tasks, they may also engage in valuable activities such as substituting for colleagues or suggesting improvements (Neckermann et al., 2014). Accordingly, research can only seek to approximate true performance from different angles and estimate how subjective ratings relate to the operationalized truth.

Given the complexity of performance, scholars have emphasized the need to account for the context of performance assessments (e.g., Johns, 2006; Levy & Williams, 2004) and have called for further research on rating formats (Heneman et al., 1987). In the performance assessment literature, however, the term “format” primarily refers to differences in the design of graphic rating scales (Brutus, 2010). In light of research acknowledging the context of appraisals as a significant influencing factor (Spence & Keeping, 2011), the predominant focus on scales seems counterintuitive as rating scales limit the ability to explain and elaborate (Brutus, 2010). While performance appraisals typically include both rating scales and comment sections (Brutus, 2010), the relation between assessment comments and true performance remains underexplored. Building on these considerations of human assessments in organizational contexts – specifically assessment quality, the significance of contextual factors, and the untapped potential of verbal assessment formats – the following research gaps emerge.

First, due to the inherent complexity of performance, few studies have examined how performance ratings relate to measurable performance indicators, such as sales productivity (Sundvik & Lindeman, 1998) or the number of errors during a task (Kusterer & Sliwka, 2024). Although these outcomes quantify performance and do not rely on subjective interpretation, they may also be influenced by factors beyond the employee's control (Campbell & Wiernik, 2015). For example, declining sales figures may not necessarily indicate poor sales performance by the employee but could instead be attributed to external factors, such as the customer's lack of decision-making power during the interaction or decreasing demand due to inflation (e.g., Motowidlo et al., 1997). Accordingly, defining performance outcomes as true performance reflects only one specific perspective, as it overlooks determinants – the foundational qualities and resources that enable performance. These include, among other factors, employees' knowledge, skills, and abilities, all of which influence how effectively they carry out their tasks (e.g., Carpini et al., 2017). Similarly, focusing on performance outcomes neglects the processes through which performance unfolds, such as how employees approach problem-solving, adapt to customer needs, or manage their time (e.g., Campbell & Wiernik, 2015; Motowidlo et al., 1997). Determinants and processes shape performance but may not be immediately reflected in outcomes like sales figures or task completion rates (Campbell & Wiernik, 2015).

This distinction is also evident in how performance is assessed in practice, where supervisors are unlikely to rely solely on outcome-based metrics such as sales figures. Instead, they may also consider how employees interact with customers and the effort they invest in preparing for meetings. Although evaluating performance through measurable outcomes offers an objective approach that minimizes subjective interpretation, a comprehensive understanding of performance requires considering not only outcomes but also the determinants and processes that drive them. This leads to the following research gap:

Research gap (1): Limited research has investigated the relationship between performance assessments and performance along the dimensions of determinants, processes, and outcomes.

Second, previous research on achieving higher rating quality has largely overlooked the role of the assessment format. Although scholars have called for further investigation into assessment formats, these calls have primarily focused on different types of rating scales, such as behaviorally anchored rating scales or behavioral observation scales (e.g., Brutus, 2010; Heneman et al., 1987). Plausibly, the main reason for the neglect of evaluation comments lies in their manageability. Writing comments requires time that supervisors often lack, and they are less straightforward to analyze and interpret (Brutus, 2010). Considering this argument while keeping recent technological advances in mind, it is notable that, to the best of my knowledge, no research has examined the potential of spoken assessment comments as an alternative form of verbal evaluation. This is particularly relevant given the advantages associated with verbal evaluations. For instance, scholars have shown that supervisors exhibit gender biases when using numerical rating scales but not when writing about performance (Biernat et al., 2012). Similarly, research indicates that supervisors show racial biases in performance ratings, yet these biases do not appear in their written comments (Wilson, 2010). Research attributes these findings to the cognitive effort involved (e.g., Fehrenbacher et al., 2018). It is reasoned that verbal accounts require more cognitive effort, i.e., individuals are forced to reflect on their assessment and to give it thought whereas numerical ratings may be assigned more automatically with less reflection (Fehrenbacher et al., 2018; Wilson, 2010).

Given these findings, extending research on performance appraisal quality to spoken comments could be a valuable consideration. The most apparent reason is that spoken comments require less production effort than written comments. Also, spoken language has distinctive characteristics that differentiate it from written language. For example, it is characterized as spontaneous and may contain repair mechanisms, particles, or fillers (Tannen, 1982). These linguistic details may seem minor at first, but they can carry significant informative value. Fox Tree and Schrock (1999) examined the discourse particle "oh", for instance, and reported its various functions, such as signaling an upcoming linguistic repair, adding emphasis, or marking an upcoming nonserious thought, among many other functions. These linguistic markers may convey significant implications in performance appraisals by reflecting the supervisor's hesitation or capturing initial thoughts that may have been erased from written texts. As research suggests that recent advances in natural language processing (NLP) enhance the manageability of written texts (e.g., Brutus, 2010), the same holds true for spoken comments, which additionally require a speech-to-text conversion compared to written evaluations. Moreover, NLP can detect linguistic subtleties and nuances that may further contextualize the evaluation. This reasoning highlights the following research gaps:

Research gap (2): The potential of spoken performance assessment comments is underexplored.

Research gap (3): A systematic comparison of how verbal and numerical performance assessments relate to performance has received limited research attention.[2]

The fourth research gap extends beyond employee performance to explore another form of verbal assessment within the organizational context, namely employee reviews. Review platforms allow employees to publicly assess the organization they work for or have worked for in the past (Dellarocas, 2003). In this scenario, the organization, now being the subject of the evaluation, has limited control over how it is assessed and about the arguments that are being made (van Hoye & Lievens, 2007). As organizations aim to be an attractive choice for both current and future employees, they may strategize on how to react to the increased transparency brought about by reviews (Dube & Zhu, 2021). To maintain a positive online reputation despite negative reviews, publicly responding to reviews has proven to be a

[2] In this dissertation, "numerical rating" refers to the ratings that are assigned on numerical rating scales. Although the written and spoken comments are converted into numbers, they will be referred to as "comments" or "algorithmic ratings".

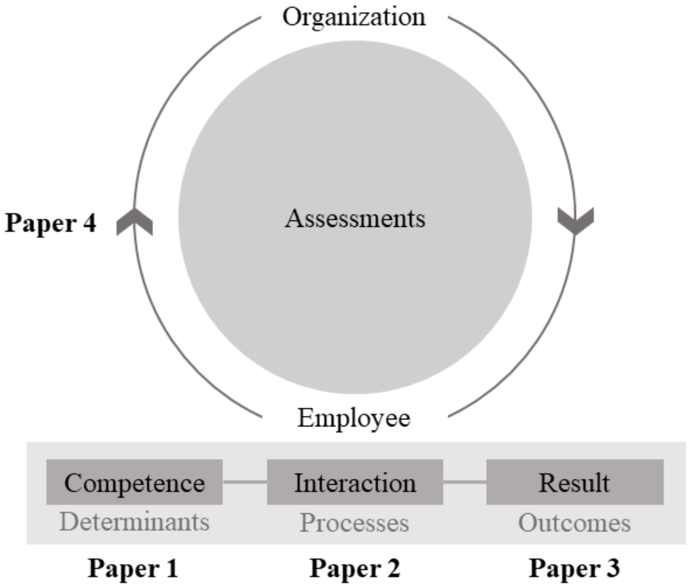
promising strategy, as it allows to address various aspects of the review and demonstrate that the evaluation is taken seriously (e.g., Chevalier et al., 2018; Kollitz et al., 2022; Proserpio & Zervas, 2017). While response strategies and consumer behavior have received considerable research attention (e.g., Ravichandran & Deng, 2023; Sparks et al., 2016), there is limited understanding of responses on employer review platforms (Yu et al., 2023), especially regarding the factors that prompt organizations to engage with employee reviews and to respond.

Understanding what drives organizations to respond is crucial for several reasons (e.g., Chevalier et al., 2018; Proserpio & Zervas, 2017). Organizations may already have an implicit awareness of their response behavior, yet actively reflecting on these patterns ensures a more deliberate and consistent approach to employer branding and reputation management. For employees and job applicants, insights into response patterns can provide valuable cues about an organization's culture, transparency, and willingness to engage with employee concerns. Moreover, from a research perspective, examining these response mechanisms can shed light on how organizations navigate external feedback, balance reputational concerns, and adapt to the growing influence of online platforms. This argumentation leads to the following research gap:

Research gap (4): The factors that prompt organizations to respond to employee reviews have yet to be investigated, leaving an important gap in understanding how organizations manage their employer reputation in the face of public scrutiny.

This dissertation aims to address these gaps and contribute to the existing discourse through four papers. Figure 1 illustrates how each paper is thematically positioned within the broader context of assessments in organizations.

Figure 1 Dissertation Framework



Paper 1 focuses on the relationship between numerical ratings and spoken assessment comments, specifically how they relate to interpersonal competencies. To make the comments comparable to the numerical ratings, we train and test a machine learning algorithm that converts the comments into numbers.[3] Our results suggest that the algorithmic ratings more accurately reflect the distribution of competencies within the sample. In this paper, we investigate the link between verbal and numerical performance assessments and competencies as performance determinants (Research gap 1, Research gap 3) and explore the potential of spoken performance appraisals (Research gap 2).

Paper 2 explores the process stage of performance. In particular, I investigate how warmth- and competence-related behaviors displayed by participants during a picture-matching task relate to numerical ratings, as well as to written and spoken appraisal comments. The findings indicate that the consideration of warmth and competence is influenced not only by the appraisal format but also by the gender of both the evaluator and the evaluated participant. This paper focuses on the performance process (Research gap 1) by investigating how warmth- and competence-

[3] The algorithm is also applied in Paper 2 and 3 to convert the comments into numerical ratings.

related behaviors during the team interaction translate into verbal and numerical performance assessments (Research gap 3). Additionally, spoken comments are tested as an alternative appraisal format (Research gap 2).

Paper 3 shifts the focus to the performance outcome, namely to how numerical ratings, along with written and spoken appraisal comments, correspond to the number of errors in a picture-matching task. Our findings suggest that spoken comments most accurately represent performance, in terms of error count, within teams. Paper 3 addresses the relationship between verbal and numerical performance appraisal formats and a performance outcome (Research gap 1 and Research gap 3), while further exploring the potential of spoken appraisal comments (Research gap 2).

Paper 4 takes a broader perspective on assessments in the organizational context by examining the factors that prompt organizations to respond to employee reviews (Research gap 4). Among other aspects, we examine whether the way organizations verbally present themselves to job applicants is linked to the response behavior they demonstrate on employer review platforms. For this purpose, we conduct a sentiment analysis of organizational identity claims in job advertisements and investigate how their verbal self-description relates to their response behavior. Our results show that the sense of community communicated through organizational identity claims, as well as the level of consensus a review receives, correlate with organizational response behavior.

2 Presentation of Papers

This dissertation includes four research articles, each differing in research objectives, scope, and formatting, as they have been prepared for submission to peer-reviewed scholarly journals independently.

(1) J. K. Gutt, K. Thommes

“Evaluating Competencies with Spoken Comments and Machine Learning”

Measuring employees’ competencies is essential for identifying gaps, promoting personal development, and increasing productivity (Boxall, 1996; Datta et al., 2003; Alagaraja, 2013). Competence is a vague construct that, in practice, is often assessed through simple questions rather than validated scales and is commonly evaluated alongside performance in supervisory appraisals, blurring the line between the two constructs (Campbell & Wiernik, 2015; Heinsman et al., 2006; Heinsman et al., 2008). This approach introduces biases and errors that impact the accuracy of evaluations (Demeré et al., 2019; Prendergast, 1999; Stauffer & Buckley, 2005). While rating scales dominate appraisal forms, assigning a single number to various observations can be challenging for supervisors, often complicating the assessment and failing to reflect true competence (Brutus, 2010; Centeno et al., 2015). Although evaluation comments provide richer context and feedback, they are challenging to obtain, do not allow for straightforward ranking, and are difficult to compare across employees (Brutus, 2010).

In our study, we trained and tested a Random Forest algorithm to predict numerical ratings for spoken assessment comments. To evaluate the rating quality, we collected data through an online escape game. Participants solved the escape game in randomized teams of four to five, completed psychometric tests on interpersonal competencies beforehand, and then evaluated each team member’s interpersonal competencies by assigning ratings on a scale and providing spoken comments. We applied the algorithm to predict numerical ratings from the spoken comments. In our analysis, we compared how the assigned ratings and the algorithmic ratings relate to the psychometric test results.

The results show that the algorithmic ratings are more nuanced and align more closely with the psychometric test results than the assigned numerical ratings. Although both types of ratings exhibit leniency, the algorithmic ratings provide a more detailed reflection of the competencies measured by the psychometric tests. This suggests that spoken comments, when processed by a machine learning algorithm, can provide a more precise and meaningful assessment of employee competencies.

Our study aims to bridge the gap between the advantages of assessment ratings and comments by employing a Random Forest algorithm. In applying the algorithm, we respond to previous calls for research on methods to analyze verbal feedback in the appraisal process (e.g., Doldor et al., 2019) and contribute to the ongoing discourse on the role of narrative comments in organizational performance appraisals.

Table 1 Presentations at Scientific Conferences and Publications in Scientific Journals

Workshops and Conferences	• Faculty Workshop, 2021, Paderborn, Poster Presentation: Jana Kim Gutt • European Academy of Management (EURAM) Conference, 2022, Winterthur / Zurich, Speaker: Jana Kim Gutt • 38th European Group for Organizational Studies (EGOS) Colloquium 2022, Vienna, Speaker: Jana Kim Gutt • 82nd Annual Meeting of the Academy of Management (AOM), 2022, Seattle, Washington, Speaker: Jana Kim Gutt • Faculty Workshop, 2022, Melle, Speaker: Jana Kim Gutt • Interdisciplinary Anniversary Conference “Data Society. Opportunity - Innovation - Responsibility”, 2022, Paderborn, Poster Presentation: Jana Kim Gutt
Status	Reject and Resubmit at European Management Journal (received on 02/20/2025)

(2) J. K. Gutt

“Evaluators’ Consideration of Warmth and Competence in Verbal and Numerical Performance Assessments”

Prior research has established that two fundamental dimensions – warmth and competence – are central to how individuals perceive themselves and others (e.g., Cuddy et al., 2008; Wojciszke & Abele, 2008; Martin & Slepian, 2021). Warmth, linked to social connection, is stereotypically associated with femininity, while competence, tied to goal achievement and task completion, stereotypically aligns with masculinity (e.g., Rudman & Phelan, 2008; Cuddy et al., 2011). Since performance appraisals inherently involve subjective judgment, they are vulnerable to stereotypical perceptions (Cuddy et al., 2011). In this context, studies have shown that the appraisal format influences the evaluation, as supervisors demonstrate racial or gender biases in numerical ratings but not in written comments (Wilson, 2010; Biernat et al., 2012). These inconsistencies between numerical ratings and comments suggest that biases may be less pronounced in verbal assessments than in numerical ratings. However, existing research has yet to address the consideration of warmth and competence in performance appraisals, especially across verbal and numerical formats, taking the rater and ratee gender into account.

To bridge this gap, I conducted a laboratory experiment[4] and analyzed the communication in a two-person task, coding it for warmth and competence. I investigated how both dimensions are reflected in numerical ratings, written comments, and spoken comments. The evaluations are provided by evaluator pairs who observed the task-solving participants (task-solvers) during the picture-matching task. I applied a

[4] The data analyzed in Paper 2 were collected as part of the same experiment described in Paper 3.

Random Forest algorithm to quantify the comments, making them comparable to the assigned numerical ratings.

Findings reveal that the consideration of warmth and competence depends not only on the appraisal format but also on the evaluator and task-solver gender. In particular, the analysis of spoken comments shows no significant gender effects, while the results for written comments and numerical ratings are mixed. Male evaluators tend to disapprove of male task-solvers' warmth statements when rating their communication on a numerical scale, whereas female evaluators approve of female task-solvers' warmth statements in their written comments on communication. Additionally, a structural topic model indicates that evaluators use more comparative language when assessing task-solvers of the same gender. This study highlights how the evaluation format, along with the gender of both the evaluator and the task-solver shape the consideration of warmth- and competence-related behaviors. The results underscore the need for organizations to carefully select the evaluative frameworks they adopt, as they can influence the judgment and contribute to disparities in performance assessments.

Table 2 Presentations at Scientific Conferences and Publications in Scientific Journals

Conferences	• European Academy of Management (EURAM), 2024, Bath, accepted for presentation • 84th Annual Meeting of the Academy of Management (AOM), 2024, Chicago, Illinois, accepted for presentation • 119th American Sociological Association (ASA) Annual Meeting, 2024, Montréal, Québec, accepted for presentation • Americas Conference on Information Systems (AMCIS), 2024, Salt Lake City, Utah, conditionally accepted for presentation
Status	Submitted to Review of Managerial Science (03/16/2025)

(3) J. K. Gutt, K. Thommes, M. Mehic

“Can Verbal Performance Appraisals and Machine Learning Models Improve the Accuracy of Performance Evaluations?”

The accuracy of performance appraisals has long been a significant concern in personnel decision-making, particularly because achieving accurate assessments of employee performance is often considered a challenging, if not impossible, task (Decotiis & Petit, 1978; Heneman et al., 1987). While the benefits of narrative comments over rating scales are well-recognized, such as providing context and reasoning (Smither & Walker, 2004; Speer, 2018), it remains unclear whether one format is more accurate than the other. To assess the accuracy of evaluation formats, it is essential to understand how each format correlates with the true performance it seeks to reflect. However, objective measures of individual performance are often unavailable (Kusterer & Sliwka, 2024), as employees typically work in teams, perform various tasks at the same time, or produce outcomes that cannot be directly measured (e.g., in creative roles).

To investigate the relationship between appraisal formats and performance, we conducted a laboratory experiment in which participants completed a picture-matching task. This task was originally used by Weber and Camerer (2003) to study the impact of organizational cultures on the success of mergers. We collected spoken and written comments, along with numerical ratings, and applied a Random Forest algorithm to predict numerical ratings for the comments. By measuring participants’ true performance through their error count and analyzing data both between and within teams, we were able to compare evaluation behaviors on both a broader and more detailed level.

Our results suggest that spoken appraisal comments are the most accurate in reflecting performance based on error count within a team. The robustness check using a generative AI approach (ChatGPT-4) largely supports the results from our analyses, indicating that spoken comments most closely reflect the number of errors both between and within teams. Overall, our findings suggest that the degree of differentiation in evaluations strongly depends on the chosen reference points (between or within teams) and the evaluation format (numerical ratings, written comments, or spoken comments). They contribute to the ongoing debate on whether performance appraisals should rely on absolute (compared to a predefined standard) or relative methods (compared to other employees) (e.g., Blume et al., 2009; Roch et al., 2011). Our results indicate that, without a reference point, individuals struggle to distinguish between high and low levels of performance. At the same time, our findings emphasize the importance of incorporating narrative comments in performance appraisals, as spoken comments were the only format that accurately reflected differences in individual performance.

Table 3 Presentations at Scientific Conferences and Publications in Scientific Journals

Conferences	• 25th Colloquium on Personnel Economics (COPE), 2023, Amsterdam, Speaker: Miro Mehic • European Academy of Management (EURAM), 2023, Dublin, Speaker: Jana Kim Gutt • 39th European Group for Organizational Studies (EGOS) Colloquium, 2023, Cagliari, Speaker: Jana Kim Gutt • 83rd Annual Meeting of the Academy of Management (AOM), 2023, Boston, Massachusetts, Speaker: Jana Kim Gutt
Status	• Published in the Academy of Management Best Paper Proceedings (2023) • Under Review at Academy of Management Discoveries (02/11/2025)

(4) J. K. Gutt, K. Knorr

"Factors Influencing Organizations' Responses on Employer Review Platforms"

In the digital age, online reviews hold significant power, making it crucial for organizations to manage their reputations effectively (Etter et al., 2019; Proserpio & Zervas, 2017). While individuals can easily share their opinions with the global community (Dellarocas, 2003), organizations have limited control over this shared information (Dube & Zhu, 2021; van Hoye & Lievens, 2007). Negative reviews are particularly challenging, as they remain online, cannot be removed, and significantly impact an organization's reputation (Proserpio & Zervas, 2017; Sparks et al., 2016). Although research has extensively investigated response strategies to consumer reviews, less attention has been given to how organizations respond on employer review platforms, particularly regarding the factors prompting a response.

Our study contributes to the limited research on organizational response behavior by proposing that an organization's decision to respond to an online review correlates with (1) the sense of community that organizations express linguistically in their organizational identity, (2) the consensus a review receives, and (3) the reviewer's job position. Additionally, we examine whether these relationships are moderated by the organization's overall rating on the review platform. Our analysis is based on a dataset of 872 job advertisements from 270 Germany-based organizations, matched with 74,786 ratings and reviews from Kununu.

Our findings reveal that organizations that linguistically emphasize community within their organizational identity are more likely to respond to reviews from employees. Moreover, we find that the consensus that an online review receives is related to an organization's responsiveness. However, contrary to our expectations, a reviewer's job position – particularly that of a manager – does not influence organizational responsiveness. Also, we find no statistically significant moderating effect of overall ratings. Our paper advances the understanding of organizational responses on employer review platforms by examining when and to what extent organizations choose to engage. While organizations may believe they are responding effectively to employee feedback, a lack of insight into underlying response patterns can lead to inconsistent reactions. Identifying which reviews receive responses – and why – enables organizations to refine their approach, ensuring more strategic and meaningful engagement.

Table 4 Presentations at Scientific Conferences and Publications in Scientific Journals

Conferences	• 40th European Group for Organizational Studies (EGOS) Colloquium, 2024, Milan, accepted for presentation • 119th American Sociological Association (ASA) Annual Meeting, 2024, Montréal, Québec, accepted for presentation • Americas Conference on Information Systems (AMCIS), 2024, Salt Lake City, Utah, conditionally accepted for presentation
Status	Under Review at The International Journal of Human Resource Management (03/12/2025)

3 Conclusion

In a broad sense, this dissertation examines assessments within the organizational sphere from two perspectives: how the organization, represented by the supervisor, evaluates employees, and how employees evaluate the organization. By addressing both perspectives, this dissertation aims to enhance the understanding of the underlying dynamics in organizational assessments and potentially provide guidance on key considerations related to the assessment format, framework, and variability (Paper 1, Paper 2, Paper 3), as well as on responding to employees' assessments when the roles of ratee and rater are reversed (Paper 4).

3.1 Assessment Format

The findings of this dissertation suggest that spoken assessments are more nuanced than numerical ratings and more accurately reflect the distribution of psychometric test results (Paper 1) and performance differences within teams (Paper 3). These results hold true when evaluators assess performance as part of a team they worked in (Paper 1) and when they evaluate performance as external observers (Paper 3). The reasons why spoken comments align more closely with operationalized true performance may vary. One reason might be that evaluators reflect more carefully on the observed performance when speaking about it, as they are encouraged to elaborate on their assessment. Additionally, spoken assessments, by being voice-recorded, inherently carry a more personal tone than ratings on a scale. The specific choice of words, use of active or passive voice, and the overall tone (e.g., overly excited or rather monotone) can provide insights into the evaluator's attitude and personality and may, in some cases, foster a greater sense of accountability when speaking about performance. As previous research has indicated, accountability encourages evaluators to provide more accurate ratings (e.g., Kusterer & Sliwka, 2024), which may help explain our findings. Also, spoken comments are more direct than numerical ratings as they do not require the evaluator to interpret what distinguishes, for example, a rating of 3 from a 4.

Spoken appraisals provide a valuable alternative to numerical ratings, as they allow for more detailed and nuanced evaluations. In certain circumstances, they may reduce some biases and enhance evaluator reflection, leading to a more accurate representation of performance. However, their accuracy depends on the specific criteria being analyzed, i.e., one might come to different results when comparing assessments to completion times instead of error counts. Rather than viewing spoken appraisals as universally superior, they should be considered as one of several tools that can be beneficial in specific situations where capturing the nuances of performance, reducing ambiguity, and encouraging evaluator reflection is particularly important.

3.2 Assessment Framework

Additionally, the dissertation highlights that evaluators tend to give relative performance appraisals, even when they are not explicitly instructed to assess performance in relation to others (Paper 3). Across the entire participant field, the assessments provide little meaningful insight, as they are uniformly positive and do not systematically reflect performance differences. However, the fact that evaluators still accounted for performance differences is only evident at the within-team level – and even then, only in spoken assessments. As evaluators in this study only observed two participants, they seemingly used the performance of one participant as a reference point to determine the assessment of the other. Additionally, I find that evaluators use comparative language more frequently when describing participants of the same gender (Paper 2), underlining that individuals may rely on within-group standards as reference points for their judgments.

These findings contribute to the ongoing debate on whether performance assessments should be conducted in relative or absolute terms (e.g., Blume et al., 2009; Chattopadhayay & Ghosh, 2012; Wagner & Goffin, 1997). Although performance appraisals should ideally function independently (i.e., a high rating in communication skills should represent the same level of ability, regardless of how others are rated), our results indicate that individuals struggle to reflect high and low performance levels in their assessments without a reference point, suggesting that assessments are inherently comparative.

If assessments are inherently relative, they must be interpreted with caution, as the perceived distinction between positive and negative assessments may be influenced by the reference group rather than actual performance. Our results highlight the need for a performance assessment framework that accounts for the evaluators’ tendency to compare (e.g., by leveraging the comparison intentionally through structured comparisons or strategically combining relative and absolute assessments).

3.3 Assessment Variability

By analyzing how specific participant behaviors in the performance process are reflected across different assessment formats (Paper 2), the dissertation also offers insights into factors contributing to assessment variations. The analyses show that the (dis-) approval of warmth-related behaviors is influenced by both rater and ratee gender and the assessment format. For example, the results indicate that female evaluators approve of female participants' warmth-related behaviors in written comments on communication or that male evaluators disapprove of male participants' warmth-related behaviors in numerical ratings on communication. Neither effect holds true for assessments of overall performance or for different formats. This is particularly noteworthy, as both numerical ratings and written comments in our study were provided by the same evaluator. While one might expect evaluators to assess performance consistently, the format appears to influence the weight they assign to their observations. Additionally, the evaluators' consideration of participant behavior is influenced by both their own and the participants' gender, suggesting that participants are not held to the same standard.

Although it is widely recognized that performance appraisals contain "noise" (Kahneman et al., 2021) or unwanted variability (e.g., Bernardin et al., 2016; Prendergast & Topel, 1993; Wilson, 2010), it remains essential to understand the factors that contribute to this variability. Performance is assessed through the evaluator's personal lens and every judgment is inevitably shaped by individual perception. As individuals will remain both imperfect judges (Kahneman et al., 2021) and imperfect processors of observations (Feiler et al., 2013), it is impossible to completely eliminate noise from performance appraisals. However, it is crucial to remain aware of the inherent subjectivity, understand the conditions under which it is more pronounced, and explore mechanisms that can help reduce noise, even if only marginally.

3.4 Responding to Employee Reviews

Although this dissertation primarily focuses on supervisory assessments, employees today also have the opportunity to take on the role of the evaluator and assess their organization. While research recognizes that an organization's responsiveness to such evaluations can have a positive impact (e.g., Proserpio & Zervas, 2017; Chevalier et al., 2018), our findings shed light on the factors that correlate with organizations' tendency to respond – namely, when they linguistically emphasize community within their organizational identity and when the consensus that the review receives is high.

While organizations may perceive their responses to employee feedback as effective, a lack of awareness regarding underlying response patterns can result in inconsistencies. Without a clear understanding of what drives their engagement, organizations may unintentionally overlook critical feedback while responding to less impactful reviews. Analyzing why reviews receive responses allows organizations to develop a more structured and intentional approach. By identifying patterns in their engagement, organizations can ensure that their responses are not only consistent but also strategically aligned with their broader communication and employee relations goals.

4 Limitations

Despite its contribution, this dissertation has certain limitations, which I will outline briefly in the following. Conducting most of the studies in a laboratory setting (Paper 1, Paper 2, Paper 3) provides a controlled but limited perspective on the complexities involved in evaluating performance in real organizational contexts, including factors like social dynamics and political influences (e.g., Levy & Williams, 2004; Spence & Keeping, 2011). Although laboratory experiments cannot fully capture the complexity of organizational reality, they help reduce noise in the observation and evaluation processes, making it easier to study the underlying mechanisms more clearly.

Recognizing that there is no single objective truth in measuring work performance (e.g., Demeré et al., 2019), this dissertation examines performance from multiple perspectives. As a result, the findings may be influenced by the specific operationalization of performance in each study and may not fully generalize to other contexts. However, by incorporating different operationalizations of performance, this dissertation aims to provide a more comprehensive and well-rounded understanding of performance assessments. Additionally, this limitation applies to both assigned ratings and comments and should therefore not systematically favor one assessment format over another.

A similar concern regarding operationalization arises in Paper 4, which takes a broader perspective by investigating the factors that prompt organizations to respond to employee reviews. In this study, we derive linguistic cues from job advertisements and operationalize community through sentiment analysis. While this approach captures certain aspects of how organizations present themselves, it is limited to a single communication channel (job advertisements), and the findings may vary depending on how community is defined. Nonetheless, the study provides valuable insights by demonstrating that organizations that use specific language in their job advertisements, among other factors, are more likely to engage with employee reviews, highlighting the role of organizational identity claims in shaping responsiveness.

5 Future Research

The findings of this dissertation suggest several directions for future research. A natural next step would be to test the results in organizational settings to determine whether the relationship between appraisal formats and performance holds or shifts due to the complex social dynamics of the workplace. In this context, an important area of investigation is how human assessments and algorithmic outputs can be effectively combined in practice. For instance, this could include examining whether evaluators should have the discretion to override algorithmic ratings, how much weight should be given to human judgment on algorithmic outputs, or whether safeguards are needed to ensure transparency and accountability in decision-making.

Another important avenue for future research is examining how employees perceive and respond to performance appraisals that are complemented by machine learning. A key question is whether employees view such evaluations as fairer and more objective compared to traditional human assessments even though the actual evaluation is still made by a human. Building on the finding that organizations describing themselves linguistically as a community are more likely to respond to employee reviews, future studies could explore whether such organizations are also more inclined to acknowledge employee concerns about performance appraisals that incorporate machine learning. For instance, organizations that emphasize a community-oriented identity may be more responsive to employee feedback on these systems, adapting their appraisal processes accordingly. Investigating these dynamics could provide valuable insights into how organizational identity claims influence responsiveness to concerns on emerging workplace technologies.

While this dissertation demonstrates how the interplay of machine learning and verbal comments can enhance appraisals in organizational settings, these findings may also be relevant in other contexts where accurate ratings are important. Beyond performance assessments, numerical ratings play a role in medical (e.g., Karcioglu et al., 2018), psychological (e.g., Lesage et al., 2012), and risk assessments (e.g., Brody et al., 2004), where decisions often rely on both quantitative and qualitative input. In these areas, incorporating spoken comments could contribute to more precise evaluations, as professionals already consider verbal descriptions alongside numerical ratings. An algorithm trained to recognize an individual's speaking style and tendencies to exaggerate or understate could offer additional insights. When combined with a professional's expertise, spoken comments combined with machine learning models may help support well-founded judgments.

References

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting Rid of Performance Ratings: Genius or Folly? A Debate. *Industrial and Organizational Psychology*, *9*(2), 219–252. https://doi.org/10.1017/iop.2015.106

Alagaraja, M. (2013). HRD and HRM Perspectives on Organizational Performance. *Human Resource Development Review*, *12*(2), 117–143. https://doi.org/10.1177/1534484312450868

Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On Judging and Being Judged Accurately in Zero-Acquaintance Situations. *Journal of Personality and Social Psychology*, *69*(3), 518–529. https://doi.org/10.1037/0022-3514.69.3.518

Bernardin, J. H., Thomason, S., Buckley, R. M., & Kane, J. S. (2016). Rater Rating-Level Bias and Accuracy in Performance Appraisals: The Impact OF Rater Personality, Performance Management Competence, and Rater Accountability. *Human Resource Management*, *55*(2), 321–340. https://doi.org/10.1002/hrm.21678

Biernat, M., Tocci, M. J., & Williams, J. C. (2012). The Language of Performance Evaluations. *Social Psychological and Personality Science*, *3*(2), 186–192. https://doi.org/10.1177/1948550611415693

Blume, B. D., Baldwin, T. T., & Rubin, R. S. (2009). Reactions to Different Types of Forced Distribution Performance Evaluation Systems. *Journal of Business and Psychology*, *24*(1), 77–91. https://doi.org/10.1007/s10869-009-9093-5

Boxall, P. (1996). The Strategic HRM Debate and the Resource-Based View of the Firm. *Human Resource Management Journal*, *6*(3), 59–75. https://doi.org/10.1111/j.1748-8583.1996.tb00412.x

Brashier, N. M., & Marsh, E. J. (2020). Judging Truth. *Annual Review of Psychology*, *71*, 499–515. https://doi.org/10.1146/annurev-psych-010419-050807

Brody, S. D., Peck, M., & Highfield, W. E. (2004). Examining localized patterns of air quality perception in Texas: a spatial and statistical analysis. *Risk Analysis: An International Journal*, *24*(6), 1561–1574. https://doi.org/10.1111/j.0272-4332.2004.00550.x

Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review*, *20*(2), 144–157. https://doi.org/10.1016/j.hrmr.2009.06.003

Campbell, J. P., & Wiernik, B. M. (2015). The Modeling and Assessment of Work Performance. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 47–74. https://doi.org/10.1146/annurev-orgpsych-032414-111427

Carpini, J. A., Parker, S. K., & Griffin, M. A. (2017). A Look Back and a Leap Forward: A Review and Synthesis of the Individual Work Performance Literature. *Academy of Management Annals*, *11*(2), 825–885. https://doi.org/10.5465/annals.2015.0151

Centeno, R., Hermoso, R., & Fasli, M. (2015). On the inaccuracy of numerical ratings: dealing with biased opinions in social networks. *Information Systems Frontiers*, *17*(4), 809–825. https://doi.org/10.1007/s10796-014-9526-1

Chattopadhayay, R., & Ghosh, A. K. (2012). Performance appraisal based on a forced distribution system: its drawbacks and remedies. *International Journal of Productivity and Performance Management*, *61*(8), 881–896. https://doi.org/10.1108/17410401211277138

Chevalier, J. A., Dover, Y., & Mayzlin, D. (2018). Channels of Impact: User Reviews When Quality Is Dynamic and Managers Respond. *Marketing Science*, *37*(5), 688–709. https://doi.org/10.1287/mksc.2018.1090

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. *Advances in Experimental Social Psychology*, *40*, 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

Cuddy, A. J.C., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, *31*, 73–98. https://doi.org/10.1016/j.riob.2011.10.004

Datta, D. K., Guthrie, J. P., & Wright, P. M. (2003). *HRM and Firm Productivity: Does Industry Matter?* Department of Human Resource Studies School of Industrial and Labor Relations, Cornell University, New York. https://ecommons.cornell.edu/bitstream/handle/1813/77113/WP03_02.pdf?sequence=1

Decotiis, T., & Petit, A. (1978). The Performance Appraisal Process: A Model and Some Testable Propositions. *Academy of Management Review*, *3*(3), 635–646. https://doi.org/10.5465/amr.1978.4305904

Dellarocas, C. (2003). The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Management Science*, *49*(10), 1407–1424. https://doi.org/10.1287/mnsc.49.10.1407.17308

Demeré, B. W., Sedatole, K. L., & Woods, A. (2019). The Role of Calibration Committees in Subjective Performance Evaluation Systems. *Management Science*, *65*(4), 1562–1585. https://doi.org/10.1287/mnsc.2017.3025

Doldor, E., Wyatt, M., & Silvester, J. (2019). Statesmen or cheerleaders? Using topic modeling to examine gendered messages in narrative developmental feedback for leaders. *The Leadership Quarterly*, *30*(5), 1–21. https://doi.org/10.1016/j.leaqua.2019.101308

Dube, S., & Zhu, C. (2021). The Disciplinary Effect of Social Media: Evidence from Firms' Responses to Glassdoor Reviews. *Journal of Accounting Research*, *59*(5), 1783–1825. https://doi.org/10.1111/1475-679X.12393

Etter, M., Ravasi, D., & Colleoni, E. (2019). Social Media and the Formation of Organizational Reputation. *Academy of Management Review*, *44*(1), 28–52. https://doi.org/10.5465/amr.2014.0280

Fehrenbacher, D. D., Schulz, A. K.-D., & Rotaru, K. (2018). The moderating role of decision mode in subjective performance evaluation. *Management Accounting Research*, *41*(1), 1–10. https://doi.org/10.1016/j.mar.2018.03.001

Feiler, D. C., Tong, J. D., & Larrick, R. P. (2013). Biased Judgment in Censored Environments. *Management Science*, *59*(3), 573–591. https://doi.org/10.1287/mnsc.1120.1612

Fox Tree, J. E., & Schrock, J. C. (1999). Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language*, *40*, 280–295. https://doi.org/10.1006/jmla.1998.2613

Funder, D. C. (1987). Errors and Mistakes: Evaluating the Accuracy of Social Judgment. *Psychological Bulletin*, *101*(1), 75–90.

Griffin, D., & Tversky, A. (1992). The Weighing of Evidence and the Determinants of Confidence. *Cognitive Psychology*, *24*(3), 411–435. https://doi.org/10.1016/0010-0285(92)90013-R

Heinsman, H., de Hoogh, Annebel H. B., Koopman, P. L., & van Muijen, J. J. (2006). Competency management: Balancing between commitment and control. *Management Revue*, *17*(3), 292–306. http://www.jstor.org/stable/41783523

Heinsman, H., Hoogh, A. H.B. de, Koopman, P. L., & van Muijen, J. J. (2008). Commitment, control, and the use of competency management. *Personnel Review*, *37*(6), 609–628. https://doi.org/10.1108/00483480810906865

Heneman, R. L., Moore, Michael, L., & Wexley, K. N. (1987). Performance-Rating Accuracy: A Critical Review. *Journal of Business Research*, *15*(5), 431–448. https://doi.org/10.1016/0148-2963(87)90011-7

Hoffmann, T. (1999). The meanings of competency. *Journal of European Industrial Training, 23*(6), 275 – 286. https://doi.org/10.1108/03090599910284650

Holmstrom, B., & Milgrom, P. (1991). Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *The Journal of Law, Economics, and Organization*, *7*, 24–52. https://doi.org/10.1093/jleo/7.special_issue.24

Johns, G. (2006). The Essential Impact of Context on Organizational Behavior. *Academy of Management Review*, *31*(2), 386–408. https://doi.org/10.5465/amr.2006.20208687

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise. A Flaw in Human Judgment*. Little, Brown Spark.

Karcioglu, O., Topacoglu, H., Dikme, O., & Dikme, O. (2018). A systematic review of the pain scales in adults: Which to use? *The American Journal of Emergency Medicine*, *36*(4), 707–714. https://doi.org/10.1016/j.ajem.2018.01.008

Kollitz, R., Ruhle, S., & Wilhelmy, A. (2022). How to deal with negative online employer reviews: An application of image repair theory. *International Journal of Selection and Assessment*, *30*(4), 526–544. https://doi.org/10.1111/ijsa.12392

Kurz, R., & Bartram, D. (2002). Competency and Individual Performance: Modelling the World of Work. In I. T. Robertson, M. Callinan, & D. Bartram (Eds.), *Organizational Effectiveness. The Role of Psychology* (pp. 227–255). John Wiley & Sons, Ltd.

Kusterer, D. J., & Sliwka, D. (2024). Social Preferences and the Informativeness of Subjective Performance Evaluations. *Management Science*, *132*(646), 2101. https://doi.org/10.1287/mnsc.2022.02267

Lesage, F.-X., Berjot, S., & Deschamps, F. (2012). Clinical stress assessment using a visual analogue scale. *Occupational Medicine (62*(8), 600–605. https://doi.org/10.1093/occmed/kqs140

Levy, P. E., & Williams, J. R. (2004). The Social Context of Performance Appraisal: A Review and Framework for the Future. *Journal of Management*, *30*(6), 881–905. https://doi.org/10.1016/j.jm.2004.06.005

Martin, A. E., & Slepian, M. L. (2021). The Primacy of Gender: Gendered Cognition Underlies the Big Two Dimensions of Social Cognition. *Perspectives on Psychological Science*, *16*(6), 1143–1158. https://doi.org/10.1177/1745691620904961

Mero, N. P., & Motowidlo, S. J. (1995). Effects of Rater Accountability on the Accuracy and the Favorability of Performance Ratings. *Journal of Applied Psychology*, *80*(4), 517–524. https://psycnet.apa.org/doi/10.1037/0021-9010.80.4.517

Mero, N. P., Motowidlo, S. J., & Anna, A. L. (2003). Effects of Accountability on Rating Behavior and Rater Accuracy. *Journal of Applied Social Psychology*, *33*(12), 2493–2514. https://doi.org/10.1111/j.1559-1816.2003.tb02777.x

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes*, *103*(2), 197–213. https://doi.org/10.1016/j.obhdp.2006.09.002

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, *14*(10), 435–440. https://doi.org/10.1016/j.tics.2010.07.004

Motowildo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance*, *10*(2), 71–83. https://doi.org/10.1207/s15327043hup1002_1

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature. *National Weather Digest*, *2*(2), 2–9.

Neckermann, S., Cueni, R., & Frey, B. S. (2014). Awards at work. *Labour Economics*, *31*(1), 205–217. https://doi.org/10.1016/j.labeco.2014.04.002

Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature*, *37*(1), 7–63.

Prendergast, C., & Topel, R. (1993). Discretion and bias in performance evaluation. *European Economic Review*, *37*, 355–365.

Proserpio, D., & Zervas, G. (2017). Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews. *Marketing Science*, *36*(5), 645–665. https://doi.org/10.1287/mksc.2017.1043

Ravichandran, T., & Deng, C. (2023). Effects of Managerial Response to Negative Reviews on Future Review Valence and Complaints. *Information Systems Research*, *34*(1), 319–341. https://doi.org/10.1287/isre.2022.1122

Roch, S. G., McNall, L. A., & Caputo, P. M. (2011). Self-Judgments of Accuracy as Indicators of Performance Evaluation Quality: Should We Believe Them? *Journal of Business and Psychology*, *26*(1), 41–55. https://doi.org/10.1007/s10869-010-9173-6

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior*, *28*(3), 61–79. https://doi.org/10.1016/j.riob.2008.04.003

Smither, J. W., & Walker, A. G. (2004). Are the Characteristics of Narrative Comments Related to Improvement in Multirater Feedback Ratings Over Time? *The Journal of Applied Psychology*, *89*(3), 575–581. https://doi.org/10.1037/0021-9010.89.3.575

Sniezek, J. A., & Henry, R. A. (1990). Revision, Weighting, and Commitment in Consensus Group Judgment. *Organizational Behavior and Human Decision Processes*, *45*(1), 66–84. https://doi.org/10.1016/0749-5978(90)90005-T

Sparks, B. A., So, K. K. F., & Bradley, G. L. (2016). Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. *Tourism Management*, *53*(3), 74–85. https://doi.org/10.1016/j.tourman.2015.09.011

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299–333. https://doi.org/10.1111/peps.12263

Spence, J. R., & Keeping, L. (2011). Conscious rating distortion in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review*, *21*(2), 85–95. https://doi.org/10.1016/j.hrmr.2010.09.013

Stauffer, J. M., & Buckley, M. R. (2005). The Existence and Nature of Racial Bias in Supervisory Ratings. *Journal of Applied Psychology*, *90*(3), 586–591. https://doi.org/10.1037/0021-9010.90.3.586

Sundvik, L., & Lindeman, M. (1998). Performance Rating Accuracy: Convergence Between Supervisor Assessment and Sales Productivity. *International Journal of Selection and Assessment*, *6*(1), 9–15. https://doi.org/10.1111/1468-2389.00067

Tannen, D. (1982). Oral and Literate Strategies in Spoken and Written Narratives. *Language*, *58*(1), 1–21. https://doi.org/10.2307/413530

van Hoye, G., & Lievens, F. (2007). Investigating Web-Based Recruitment Sources: Employee testimonials vs word-of-mouse. *International Journal of Selection and Assessment*, *15*(4), 372–382. https://doi.org/10.1111/j.1468-2389.2007.00396.x

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative Analysis of the Reliability of Job Performance Ratings. *Journal of Applied Psychology*, *81*(5), 557–574. https://psycnet.apa.org/doi/10.1037/0021-9010.81.5.557

Wagner, S. H., & Goffin, R. D. (1997). Differences in Accuracy of Absolute and Comparative Performance Appraisal Methods. *Organizational Behavior and Human Decision Processes*, *70*(2), 95–103. https://doi.org/10.1006/obhd.1997.2698

Weber, R. A., & Camerer, C. F. (2003). Cultural Conflict and Merger Failure: An Experimental Approach. *Management Science*, *49*(4), 400–415. https://doi.org/10.1287/mnsc.49.4.400.14430

Wiese, J., Buehler, R., & Griffin, D. (2016). Backward planning: Effects of planning direction on predictions of task completion time. *Judgment and Decision Making*, *11*(2), 147–167. https://doi.org/10.1017/S1930297500007269

Wilson, K. Y. (2010). An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations*, *63*(12), 1903–1933. https://doi.org/10.1177/0018726710369396

Wojciszke, B., & Abele, A. E. (2008). The primacy of communion over agency and its reversals in evaluations. *European Journal of Social Psychology*, *38*(7), 1139–1147. https://doi.org/10.1002/ejsp.549

Yu, K. Y. T., Goh, K. H., YU, S., & Soo, C. W. L. (2023). An Impression Management Approach to Managing Online Employer Reviews. *Academy of Management Best Paper Proceedings*. https://doi.org/10.5465/AMPROC.2023.244bp

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater Personality, Rating Format, and Social Context: Implications for Performance Appraisal Ratings. *International Journal of Selection and Assessment*, *13*(2), 97–107. https://doi.org/10.1111/j.0965-075X.2005.00304.x