UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FACULTY OF BUSINESS ADMINISTRATION AND ECONOMICS

Working Paper Series

Working Paper No. 2025-05

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

Jana Kim Gutt, Kirsten Thommes and Miro Mehic

February 2025

wiwi

CAN VERBAL PERFORMANCE APPRAISALS AND MACHINE LEARNING IMPROVE THE ACCURACY OF PERFORMANCE EVALUATIONS?

Jana Kim Gutt
Paderborn University
jana.gutt@uni-paderborn.de

Kirsten Thommes
Paderborn University
kirsten.thommes@uni-paderborn.de

Miro Mehic
Paderborn University
miro.mehic@uni-paderborn.de

ABSTRACT

Performance appraisals are subject to recent debates with one common denominator: most discussions point to their lack of accuracy. In theory, performance appraisals aim to reflect an employee's performance over a certain period of time. However, recent research shows that appraisals fall short in reaching this goal. Although many studies acknowledge the benefits of performance comments over ratings on a scale, research has paid little attention to the potential of performance comments to achieve higher accuracy in performance evaluations. To approach this issue, we conducted a laboratory experiment and collected objective performance data as well as numerical and verbal performance appraisals. In particular, we compile numerical ratings, written comments, and spoken comments on performance from independent evaluators. To make the numbers (assigned ratings) and the comments comparable, we applied a Random Forest algorithm to transfer the comments into numerical ratings (algorithmic ratings). By analyzing each rating (assigned and algorithmic) in relation to the performance, we find evidence that spoken comments reflect performance differences most accurately within a team. Our results offer important insights into how performance appraisals may be approached to reflect objective performance more accurately.

Keywords: performance appraisal; rating accuracy; rating format; performance appraisal comment; rating scale

INTRODUCTION

In recent debates, performance appraisal systems and their shortcomings have been widely discussed (Adler, Campion, Colquitt, Grubb, Murphy, Ollander-Krane, & Pulakos, 2016; Brown, Inceoglu, & Lin, 2017; Murphy, 2020). In general, performance appraisals aim to reflect an employee's performance during a certain time period. However, low reliability of performance ratings and underlying biases complicate the process of reflecting true performance (Angelovski, Brandts, & Sola, 2016; Biernat, Tocci, & Williams, 2012; Bizzi, 2018; Catano, Darr, & Campbell, 2007; Kamphorst & Swank, 2018; Speer, 2018). In general, one can distinguish between issues concerning evaluations on the general level and fairness considerations. The former make it challenging to identify the strongest and the weakest performing employees, e.g., because of too much compression in ratings or too much leniency. Contrarily, the latter deal with the blurry lines on the group level as team outcomes cannot be attributed to one particular group member, e.g., the fair acknowledgement of achievements or the misattribution of mistakes. Taken together, these factors lead to a considerable decrease in the utility of performance appraisals for individuals and organizations (Murphy, 2020).

Although performance appraisals typically consist of both quantitative rating scales and qualitative comment sections, appraisals have become synonymous with numerical ratings (Brutus, 2010). In many practical applications, comment sections are intended to provide reasoning for a given rating. However, because comments are time-consuming to process, they often remain blank or go unanalyzed. In an effort to address the shortcomings of performance appraisals, improving numerical rating scales has been the most frequently proposed solution (Adler et al., 2016). Yet, research has shown that improving rating scales does not necessarily reduce bias or enhance accuracy, which has led to increased attention on rater training (Sánchez, Díaz-Cabrera, & Hernández-Fernaud, 2019). Despite this, studies also indicate that training alone is not sufficient to achieve

greater accuracy (e.g., Bernardin & Pence, 1980; Sánchez et al., 2019). Recognizing this limitation early on, Henemann, Moore, and Wexley (1987) emphasized the need to give greater consideration to the appraisal format when investigating rating accuracy.

However, few studies in the field of performance appraisals have sought to examine the role of performance appraisal comments when examining rating accuracy. One reason for the neglect of performance comments certainly lies in manageability. Numerical ratings are much easier to produce and process than comments (Brutus, 2010; Speer, 2018). Yet, with recent advances in natural language processing (NLP), texts can be collected, analyzed, and operationalized with low time effort. The benefits of comment over rating scales have been widely acknowledged. Previous studies point out that employees pay more attention to performance comments than to ratings (Smither & Walker, 2004) and that comments provide more reasoning and give context (Speer, 2018). Accordingly, employees have more information on why a certain assessment was made and get a clearer idea of the areas they need to improve.

Nonetheless, the advantages of one over the other do not allow for judgment on the accuracy of either appraisal format. If rating scales and narrative comments are both unrelated to the true performance, they neither hold value for the evaluated employee nor for individuals working with the appraisal data (Decotiis & Petit Andre, 1978). Thus, it is imperative to shed light on the relation between the appraisal format (i.e., ratings and comments) and the true performance.

In performance appraisal literature, the term "format" predominantly indicates variations in the design of graphic rating scales (Brutus, 2010). We use a broader definition of the term in this paper, including numerical but also verbal[1] elements. Numerical formats are limited in the topics

[1] In this paper, we adhere to the Merriam-Webster definition of "verbal" as "relating to or consisting of words" (Merriam-Webster, n.d.). Accordingly, "verbal" includes both spoken and written comments.

they address and in their response options (e.g., rating the communication skills of a co-worker on a five-point scale). Contrarily, verbal formats also address predefined topics but respondents are free to choose which aspects they want to mention and to what degree (e.g., evaluating the communication skills of a co-worker by writing a comment).

To judge the accuracy of numerical and verbal evaluation formats, it is crucial to know the extent to which each format is connected to the ground truth, i.e., the real performance that it aims to reflect. Yet, measures of individual performance are usually not available for employees, as they often work in teams, jobs require multi-tasking or the work output cannot directly be measured (e.g., in creative tasks).

With the aim of investigating the relationship between the appraisal format and the performance, we conducted a laboratory experiment. For this purpose, we collected spoken and written comments as well as numerical ratings to test different evaluation formats. Through the use of a trained Random Forest algorithm,[2] we predict numerical ratings for the written and spoken comments (algorithmic ratings) that allow for a comparison with the assigned ratings on the one hand and the performance measure on the other. We measure the participants' true performance by documenting their individual number of errors during the task. One important differentiation that we make in our study is analyzing between and within teams. This allows us to examine the evaluation behavior on a broader level but also on a more granular scale.

Our results fall into two broad categories, regarding (1) the level of analysis and (2) the appraisal format. The findings indicate that (1) neither assigned nor algorithmic ratings correlate with performance between teams. Yet, our within-teams-analysis reveals that evaluators differentiate

[2] Within the scope of a previous paper, we have already trained a Random Forest algorithm to translate verbal assessments into numbers (source blinded for double-blind review; the source will be added later).

the most (2) when speaking about performance, i.e., the differences in algorithmic ratings derived from spoken comments correlate with performance differences.

The results from the between- and within-team analysis are largely supported by the robustness check conducted using a generative AI approach (ChatGPT-4). The robustness check indicates that spoken comments most closely reflect the number of errors at both the between- and within-team levels. In sum, our analysis suggests 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).

Based on our findings, we suggest that performance comments, when combined with machine learning models, can significantly improve the accuracy of performance appraisals. By applying both a widely used supervised learning model (Random Forest) and a Large Language Model (ChatGPT), we account for the interpretability-accuracy trade-off in machine learning models (Feuerriegel et al., 2025). Our findings show that both models yield the same results, indicating that spoken comments relate most closely to performance. Our paper is a first step toward increasing the accuracy of performance appraisals through verbal comments and machine learning.

This paper is organized into the following sections: First, we review the existing literature on rating accuracy in performance appraisals. In the next step, we describe the data and our method. Specifically, we present our experimental design, empirical strategy, and the Random Forest algorithm which we applied to translate the appraisal comments into numerical ratings. Also, we depict our data collection and processing. Next, we describe our data analysis and show our results. In the last section of our paper, we discuss our results and conclude our work.

RATING ACCURACY

Rating accuracy is understood as a measure that reflects the degree to which behavior is accurately credited (Martell & Leavitt, 2002), or in other words, it describes the extent to which the rating reflects the true score (Funder, 1995; Speer, 2020). The accuracy of performance ratings has previously been described as "an issue of major concern" (Henemann et al., 1987: 431) in personnel decision-making, especially since accurate assessments of employee performance seem to pose an "impossible task" (Decotiis & Petit Andre, 1978: 635). One factor influencing rating accuracy is the approach to the true value, i.e. the measures that are applied. In the context of performance assessments, the measures are predominantly distinguished in terms of subjectivity or objectivity (Bayo-Moriones, Galdon-Sanchez, & Martinez-de-Morentin, 2020; Hoffman, Nathan, & Holden, 1991; Rojon, McDowall, & Saunders, 2015). In the following sections, we will review the literature with regard to the dimensions of both approaches.

Subjective measures

Over the last years, numerous studies have tried to estimate rating accuracy. For example, Borman (1979) investigated the relationship between rater training and rating biases and let participants watch videos that showed individuals performing a task. One group of participants underwent rater training in advance, while the other did not. Results showed that prior training did not improve accuracy in performance appraisals. To estimate which group rated the observed performance more accurately, expert ratings served as a true value (Borman, 1979).

Similarly, Mero, Motowidlo, and Anna (2003) explored the impact of accountability on rating accuracy and used expert ratings as the true score. In their study, participants also watched videotapes of employees performing a work task. If they were assigned to the treatment group, participants had to justify their performance rating. Being accountable improved the accuracy and

the effect was moderated by behavior during the observational period. In line with this approach, Yun, Donahue, Dudley, and McFarland (2005) drew on expert ratings when examining the effects of personality, social context, and the rating format on ratings. The performance that the participants were asked to assess was a video-taped teamwork task. They found that a rater's personality (i.e., agreeableness) affects rating accuracy.

In all of these examples, the accuracy of participants' ratings is measured through inter-rater reliability. Expert ratings serve as a reference point and both ratings (participant – expert) focus on one point in time, i.e., work performance shown in a video. In general, rating accuracy has predominantly been determined with different types of inter-rater agreement (Vanhove, Gibbons, & Kedharnath, 2016). In this respect, inter-rater reliability is not limited to experts' and participants' ratings. Self-ratings as well as supervisor-, subordinate-, or peer-ratings can also be considered when examining inter-rater agreement (Conway & Huffcutt, 1997).

Yet, self-ratings have been found to be more lenient than peer- and supervisor-ratings (e.g., Heidemeier & Moser, 2009; Hoffman et al., 1991; Ohland et al., 2012). This potentially leads to discrepancies between self-ratings and third-party-ratings and ultimately affects the reliability. Funder (1995) calls the inter-rater approach "constructivist" (p. 666) as it only requires inter-rater consensus. It ignores that raters may suffer from the same biases and therefore also overlooks that the inter-rater overlap is not equal to objective accuracy. Similarly, Funder and West (1993) argue that: "two judges can agree with each other perfectly yet both be perfectly wrong" (p. 458). Still, inter-rater agreement influences the confidence in accuracy (Funder, 2012). When different raters come to very different conclusions, the disagreement indicates that one rater must be wrong and vice versa (Funder, 2012). This means that inter-rater agreement serves as an approach for estimating accuracy but does not objectively reflect the rated criteria.

Beyond inter-rater comparisons, intra-rater reliability is another way to estimate rating accuracy (Viswesvaran, Ones, & Schmidt, 1996). For ratings over a period of time, rate-rerate reliability serves as a measure of accuracy, or more specifically for consistency (Dierdorff & Wilson, 2003). Here, the ratings of one rater are compared at different points in time (Viswesvaran et al., 1996). However, performance may change with time and variance in the (re-) rating process may be ascribed to differences in performance (Sturman, Cheramie, & Cashen, 2005). Additionally, Cronbach's alpha represents a way to compute the intra-rater reliability for multi-item scales (Viswesvaran et al., 1996).

Yet, all of the presented methods would fail if there were persistent and systematic measurement biases among raters or across time. In other words, when the true value is not really true, this complicates the process of measuring rating accuracy.

Objective measures

Apart from subjective approaches, performance can also be estimated by applying objective measures. Sundvik and Lindeman (1998) investigated rating accuracy and its predictors in a field setting. In particular, they collected supervisory performance ratings and used sales productivity as the true performance value. Since the company gave each salesperson points for every sold service, these points served as a measure of sales productivity. The authors found that rating accuracy varied significantly and that ratings were more accurate the longer the supervisor–subordinate relationship. More recently, Kusterer and Sliwka (2022) studied biases in performance evaluations by comparing subjective appraisals to objective data from a text-entering task. Their results indicate that when accuracy is rewarded, evaluations are less lenient. Also, their study shows that ratings are less accurate and more lenient when bonus payments depend on the evaluation.

Other studies have not explored rating accuracy but applied objective performance measures for different purposes. For example, Sturman and Trevor (2001) examined the relationship between turnover and performance. They used data from a financial services organization and measured performance in terms of fees that are generated from sold loans. Another example of applying objective measures of performance is the well-known study of Lazear (2000) who investigated the influence of monetary incentives on performance. In his work, performance was measured in terms of the number of pieces that employees installed, or more specifically the number of glass units. Mas and Moretti (2009) studied peer effects and assessed performance by counting the number of scanned products per second in a supermarket chain. In their study, Bandiera, Barankay, and Rasul (2005) investigated the role of relative incentives. By analyzing data from a fruit farm, they studied the effects of relative daily pay on performance. Here, the daily pay depended on the own performance relative to the average co-worker's performance. The authors determined individual performance by evaluating the kilograms of fruit that workers picked per hour.

Although these studies show that there are certainly ways to objectively measure individual performance, literature also demonstrates that these measures predominantly rely on quantity in a given time. Frederiksen, Lange, and Kriechel (2017) refer to these settings as "simple settings where simple measures can readily quantify the overall performance of employees reasonably well" (pp. 408). Yet, in most work settings, supervisors can only observe a total work result and cannot observe the individual contribution that an employee made to this result (Mas & Moretti, 2009). This explains why most studies in the performance appraisal literature focused mainly on subjective approaches to estimate the true value.

Relation between subjective and objective measures

Subjective appraisals and objective productivity may be very different measures and there is some evidence that supervisors inflate ratings for low-performers (Bol, 2011; Golman & Bhatia, 2012; Varma, Pichler, & Srinivas, 2005; Woods, 2012), especially when signals are noisy (Bolton, Kusterer, & Mans, 2019). While objective measures may only represent one single task, subjective measures are more holistic, yet prone to biases and therefore not necessarily more accurate.

We, therefore, test a different approach to capture performance, namely comments. Since objective key figures remain unavailable in most organizations (Kusterer & Sliwka, 2022), we explore the extent to which some subjective evaluations (i.e., comments) might be more accurate than others (i.e., ratings). Notably, the comments are made by individuals and underlie subjectivity. Yet, they might relate more strongly to performance than numerical ratings because they contain more information on the context and on factors influencing the evaluation. To the best of our knowledge, this method has not fully been explored in the past, typically because comments cannot be as easily compared to each other as numbers. In comments, evaluation criteria might vary and accordingly, employees' performance cannot be ranked intuitively. However, recent advancements in natural language processing may change the picture (Haller, Aldea, Seifert, & Strisciuglio, 2022; Hirschberg & Manning, 2015).

The benefits of comments have been widely acknowledged. Smither and Walker (2004) report that employees pay more attention to written texts than to numerical ratings (Smither & Walker, 2004), potentially because texts put the evaluation into context and provide reasons for certain decisions (Speer, 2018). Yet, text processing is demanding as the text needs to be written and coded. This concerns the evaluator writing texts but also managers in human resources and employees reading and processing these reports to take adequate action. For this reason, natural language processing has been on the rise (Speer, 2018). However, the content of narrative

comments and the identification of analysis methods have received little research attention (Doldor, Wyatt, & Silvester, 2019; Silva & Ribeiro, 2021).

Related research from online reviews suggests that textual assessments might be a richer source of information (e.g., Archak, Ghose, & Ipeirotis, 2011). Brutus (2010), for instance, shows exemplarily how a written comment provides more reasoning and reveals more improvement potential than just a low rating: "Bob is rarely willing to accept our input as to the best strategy to take. In the weekly meetings, he never seriously considers our suggestions about budget allocation. Most of us do not even prepare our numbers anymore" (p. 145).

Appraisal comments may also be advantageous in the sense that linguistic cues may convey the evaluator's level of uncertainty and reveal more information about potential biases and arguments taken into consideration. Accordingly, speaking or writing performance evaluations may influence accuracy. In this paper, we explore two ways of dealing with verbal formats, including written and spoken comments. Written comments may be more feasible in many situations; however, they are time-consuming and the wording may be more reflected. Spoken comments may be less time-consuming and bear more linguistic cues than thought-through written comments. However, they may be also more ad hoc and less prepared.

For both approaches, we use machine learning techniques to translate the comments into numerical ratings. To test numerical and verbal appraisal formats in terms of rating accuracy, we apply objective measures to approach the true value. For this purpose, we conducted a laboratory experiment in which we documented participants' individual number of errors in a picture-matching task. We use the experimental design introduced by Weber and Camerer (2003). Their experiment is divided into a pre- and post-merger phase. For our study, we use the experimental design of the pre-merger stage.

DATA AND METHOD

Experimental design

We conducted a laboratory experiment, in which participants were required to solve a picture-matching task. Originally, Weber and Camerer (2003)[3] used the task to investigate the influence of organizational cultures on the success of organizational mergers. In each session, we had four subjects: two serving as task-solvers and two as evaluators. The task-solvers were asked to perform the picture-matching task for 20 rounds in teams of two with a time limit of 150 seconds per round. As they were solving the task, they were observed by the two evaluators. The experimental design is depicted in Figure 1.

Insert Figure 1 about here

When arriving at the experiment, participants randomly drew a number, which determined their role (task-solver, evaluator in the writing condition, or evaluator in the speaking condition). Task-solving participants received a performance-based compensation $((1\ Euro - \frac{required\ time}{150}) * accuracy)$, while the evaluators were compensated with 12 Euros per hour. The performance-based compensation was calculated for each of the 20 rounds. Accuracy was coded as a dummy variable that took the value 0 if errors occurred and 1 if no errors occurred in a round.

The two task solvers had the same selection of 16 pictures laid out on the table in front of them. All pictures showed different office environments in black and white. The subjects were seated opposite each other but could neither see one another nor one another's pictures. Each subject was assigned one role during each round: manager or employee. The roles switched after every single round, i.e., the participant who acted as a manager in round 1, performed as the employee

[3] We use the original task and procedures from Weber and Camerer (2003). However, the authors recommended using updated pictures which we did.

in round 2 and vice versa. It was the manager's task to describe four predefined pictures to the employee. By listening actively and checking back, the employee tried to identify the four described pictures in the correct order within the time limit of 150 seconds. When the employee indicated that the four pictures were identified, the experimenter gave feedback on the required time and accuracy of the selected pictures. Since the roles changed after each round, the participants performed as the manager and the employee in equal parts. During the task, the interaction between both participants was audio-recorded. The goal was for the manager to describe and for the employee to identify the four correct pictures in a short amount of time.

Two evaluators were observing the task-solvers during the picture-matching task. Both evaluators were placed so that they could see both picture-matching partners, but could not have made eye contact with one another. The evaluators were not allowed to talk until we asked them for their assessment after the task was finished. During the working phase of the picture-matching partners, the evaluators could observe the whole working performance. After the working phase, both evaluators were brought to separate rooms. Here, they were asked for their assessment of the individual performance of both partners solving the picture-matching task.

To cover both numerical and verbal assessment formats, one evaluator was asked for a **written assessment**, while the other gave a spoken assessment of the observed performance. The two evaluators were asked about the overall performance of each individual task-solver and their communication skills during the task. We decided to ask for an assessment of the overall performance and the communication skills in order to animate the evaluators to give their appraisal some thought and reflect on both task-solvers in a detailed way. Following performance appraisals in organizations, the written assessments consisted of a numerical rating scale and a comment section. The

rating scale ranged from 1 to 6, with 1 indicating a very good and 6 implying a very poor performance. This rating scheme is based on the German school grading system which we expected the participants to be familiar with.

Evaluators who were assigned to the **spoken evaluation** format answered the same questions by speaking about the performance. The comments were recorded and transcribed. Each picture-matching team was assessed by a new pair of performance evaluators, meaning that each evaluator only evaluated one team instead of one after the other.

The experimenter documented the completion time and the errors during the 20 rounds. The timer was started as soon as the manager said "go" and stopped as soon as the employee indicated that they identified all four pictures in a round. After each round, the experimenter announced the completion time to the task-solvers and the evaluators simultaneously. Then the task-solver who acted in the role of the employee in that particular round was asked to read the pictures that they identified out loud, e.g., "16, 4, 9, and 3". The experimenter gave immediate feedback by either stating that the pictures were correct or by reporting the order that would have been correct, e.g., "That is incorrect. The correct order would have been 7, 4, 9, and 3".[4] This way, every participant in the room was aware of the performance in each round. Additionally, the experimenter documented the number of errors and who they thought was responsible (see chapter "Data processing"). The number of errors varied between 0 and 4 in each round as the task-solvers had to identify four pictures. We documented three types of errors: incorrect pictures, invalid wording, and time-outs. The experimenter assigned the responsibility for the respective errors on their protocol sheet and did not declare their decision audibly.

[4] By giving the performance feedback after each round, we made sure that participants did not give up on the round after one error.

Empirical Strategy

As a true value, we use the individual number of errors during the picture-matching task. Figure 2 illustrates our approach.

Insert Figure 2 about here

Through the picture-matching task, we collect two types of data: performance data from the participants who perform the task ("task-solvers") and appraisal data from the participants who evaluate the performance of the task-solvers ("evaluators"). The appraisal data consists of written and spoken evaluations. In the writing condition, participants rated the performance of the task-solvers on a rating scale and were asked to write a short comment on the individual performance. Contrarily, in the speaking condition, participants were only asked for a comment, without giving a numerical rating for the respective performance.

To make the three appraisal formats comparable, it is necessary to translate the written and spoken comments into a number. For this purpose, we use a Random Forest algorithm (see chapter "Algorithmic solution") that predicts a numerical rating for each comment (hereafter, algorithmic rating). An alternative strategy that has frequently been used in the literature would be to manually code all comments (e.g., David, 2013; Silva & Ribeiro, 2021; Wilson, 2010). Yet, this approach can be prone to subjectivity and does not scale well for large datasets. Our strategy circumvents these issues as the algorithm does not rely on subjective researcher choices.

With the numerical ratings that were directly assigned by the evaluators and the algorithmic ratings that were predicted by the Random Forest (based on the written and spoken comments), we compare how the ratings relate to true performance (see chapter "Data analysis"). In this study, we estimate true performance by documenting the individual number of errors caused by the task-solvers during the picture-matching task.

Algorithmic solution for ratings

To connect the richness of qualitative comments with the high usability of numbers, we trained and tested a Random Forest algorithm in a previous work (blind for review). The algorithm was trained on 1,193 audio responses and respective numerical ratings. Specifically, respondents were asked to describe their experiences with the strongest and weakest team members they have worked with in the past in terms of eight predefined subject areas. When speaking about the subject area, e.g., communication skills, participants were also asked to indicate which numerical rating they would give for the described experience. The rating system was based on German school grades. Accordingly, the scale covered the ratings from 1 to 6, with 1 indicating very good and 6 implying very poor performance. To ensure easier interpretation of the results, the ratings were inverted for the analyses (1 = very poor, 6 = very good).

Participants were free to decide who they wanted to describe. For example, they could choose one person for all eight subject areas or different people. Also, they could speak about someone they worked with in the past or somebody they currently work with. By directing the questions to whomever the respondents could think of, we received relatively equally distributed data (Figure 3). This data distribution offered a good basis for training an algorithm that can differentiate between good and bad performance. Had we trained it on a dataset that is skewed towards high ratings, the algorithm would arguably be worse at differentiating between good and bad performances.

Insert Figure 3 about here

The audio comments were recorded, transcribed, and translated into English. To process the transcripts and develop our model, we used the scikit-learn library in Python. For the prediction of

the ratings, we operationalized a feature vector by using Term Frequency Inverse Document Frequency (TF-IDF). Drawing on previous studies, we employed 80 percent of the data for training and 20 percent for testing purposes (e.g., Lei, Qian, & Zhao, 2016). Concerning the prediction performance of the Random Forest, we received a root mean square error (RMSE) of 0.94. Lower RMSEs stand for higher accuracy. When comparing our prediction performance to the reported performance in other scientific studies (e.g., Ganu, Kakodkar, & Marian, 2013), our RMSE falls in the same range. As the Random Forest showed promising performance on a separate test dataset, we are confident in applying the algorithm to the evaluation comments of the picture-matching task.

Data collection

We recruited participants in lectures at a medium-sized university in Europe. For participation, students received monetary compensation. The first round of data collection started in May 2022 and the last round finished in November 2022. From 184 participants who took part in the experiment, the data of 160 subjects were considered for further analysis. We matched the performance data of 80 task-solvers to the appraisals of 80 evaluators. 40 evaluators were in the writing and 40 in the speaking condition. Consequently, each of the 80 task-solvers was assessed through one written and one spoken evaluation because the evaluators observed both participants in the working team. The descriptive statistics are summarized in Table 1.

Insert Table 1 about here

On average, task-solvers made 3.05 mistakes (SD=3.89) in 20 rounds. This shows that most participants performed quite well. The average completion time for the task was 1,014 seconds (SD=251). With an overall time limit of 3,000 seconds (20 rounds*150 seconds), the fastest team completed the task in 586 seconds, while the slowest team required 1,798 seconds.

Data processing

To explore the relationship between the appraisal format and the true performance, we first analyzed the numerical ratings that were given by the evaluators in the writing condition, the written evaluation comments, and the spoken evaluations. In their evaluations, participants answered two main questions concerning the task-solvers general performance and their communication skills. For both questions, evaluators in the writing condition gave a numerical rating and wrote a comment. Evaluators in the speaking condition answered both questions in a comment without giving a numerical rating.

The spoken comments were recorded, transcribed, and translated into English. We used the machine translator DeepL since it has been found to achieve outstanding translation results (Reber, 2019). Although a word-by-word translation cannot be achieved without changing the meaning of the transcript, these inaccuracies have been reported to be of less concern when words are analyzed individually like in our approach (Reber, 2019). Just like the spoken evaluations, the written comments were also translated using DeepL. Student assistants checked the translations for errors and made sure that the original meaning of the transcripts was not changed. Additionally, the author team performed a final check. With the revised transcripts, we applied the Random Forest algorithm that predicted a numerical rating for each comment.

To estimate the true performance, we used the individual number of errors. For each team, the number of errors was documented by the experimenter who also recorded the error responsibility. In some cases, the responsibility was clear (e.g., “Sorry, I described the wrong pictures”), in others, it was more difficult to assign the error to one task-solver. In these cases, the experimenter assigned the error to the whole team.

DATA ANALYSIS

Between-team analysis

As the data from the descriptive statistics already suggested, the participants performed the task quite well with an average of 3.05 errors in 20 rounds. Figure 4 shows that 45 percent of the task-solvers completed the 20 rounds with one error or less. Overall, the majority of participants (56.25 percent) completed the task with 2 errors or less.

Insert Figure 4 about here

In comparison, the numerical ratings directly assigned by the evaluators in the writing condition (Figure 5), the algorithmic ratings for the written comments (Figure 6), and the algorithmic ratings for the spoken comments (Figure 7) are distributed as follows:

Insert Figure 5, Figure 6, and Figure 7 about here

It is important to note that the distribution of the assigned ratings (Figure 5) largely resembles the rating distribution commonly found in the literature: it is highly skewed towards positive ratings (Hu, Pavlou, & Zhang, 2017). Despite the fact that the errors are distributed in a similar way (many teams with very few errors), this does not permit us to conclude that numerical ratings assigned by evaluators are an appropriate measure to evaluate performance. For example, even though restaurant ratings are positively skewed, it would be misleading to conclude that there are only outstanding restaurants in a market. Schoenmueller, Netzer, and Stahl (2020) present evidence of such skewness in a wide variety of cases, e.g., Blablacar, Goodreads, and Netflix.

Although the distribution of the performance data and the assigned ratings look roughly similar, the ratings do not correlate with the number of errors (Table 2).

Insert Table 2 about here

The general picture emerging from the between-team analysis is that evaluators do not systematically differentiate between high and low performance between teams; regardless of using numerical ratings, written comments, or spoken comments.[5]

Within-team analysis

Since evaluators only observed a single team performing the task and did not get an overview of the other teams' performance, evaluators only had one reference point (the other task-solver). By watching two participants solve the task, evaluators were able to compare the performance of one task-solver relative to the other. This way, evaluators possibly determine high or low performance in relation to the other task-solver and differentiate on a much smaller scale, namely the team level instead of the whole sample. Consequently, we take a closer look inside the teams and investigate to what extent evaluators differentiate between the two task-solvers that they observed. For this purpose, we calculate the performance gap in each team (errors task-solver A – errors task-solver B) and compare this difference to the difference in the two ratings.

The analysis reveals that the difference in the number of errors between two task-solvers positively correlates with the difference in the algorithmic ratings in the speaking condition (Table 3).

Insert Table 3 about here

This correlation indicates that the evaluators in the speaking condition differentiate more strongly between two task-solvers the bigger their difference in the number of errors. Large differences in the number of errors imply that one task-solver performed much weaker than the

[5] The results, presented as scatter plots, are included in the Appendix (see Figure 1 – Figure 3).

other. Figure 8 illustrates the results for the algorithmic ratings in the speaking condition in a scatter plot.[6]

Insert Figure 8 about here

Robustness Check

In the next step, we conduct a robustness check to address concerns that our results may be merely an artifact of our prediction algorithm rather than a general pattern. To this end, we analyze the written and spoken comments using ChatGPT-4. Specifically, we prompt ChatGPT to assess the described performance on a scale from 1 (very good) to 6 (very poor). For the analyses, we inverted the ratings so that 1 corresponds to very poor and 6 to very good. The results of the between-team analysis are presented in Table 4.

Insert Table 4 about here

While most results remain consistent, we find a significant negative correlation between the number of errors and the algorithmic ratings for spoken comments, indicating that evaluators assess task-solvers more negatively when their number of errors is high. Figure 9 illustrates the results for the spoken comments using a scatter plot.[7]

Insert Figure 9 about here

For the within-team analysis, the results remain largely consistent as well. However, the positive correlation between the difference in the number of errors and the difference in ratings for spoken comments becomes even more pronounced (Table 5).

Insert Table 5 about here

[6] The scatter plots for the assigned ratings and algorithmic ratings (writing condition) are attached in the Appendix (see Figures 4 and 5).
[7] The scatter plot for the ChatGPT ratings in the writing condition is attached in the Appendix (see Figure 6).

The results indicate that evaluators in the speaking condition incorporated differences in the number of errors within a team into their assessments, whereas those in the writing condition did not. Figure 10 presents the results for spoken comments as a scatter plot.[8]

Insert Figure 10 about here

The robustness check provides strong support for the stability of our findings. Specifically, the results remain consistent when using an alternative algorithm for rating prediction, reinforcing the credibility of the approach. In our specific setting, spoken comments have shown to reflect performance most accurately.

SUMMARY OF RESULTS

In summary, the between- and within-team analyses provide evidence that spoken appraisal comments are a richer source of performance information than numerical ratings assigned by evaluators. Although the distributions of performance data and assigned ratings look roughly similar, we do not find support for the assumption that assigned numerical ratings are indeed an accurate measure for assessing performance. In particular, our data reveal no correlations between performance and the three appraisal formats on the between-team level (assigned ratings, algorithmic ratings for written comments, and algorithmic ratings for spoken comments).

To gain a deeper understanding of the underlying dynamics of the performance evaluations, we conducted a more detailed analysis within teams. Between teams, evaluators had no reference point, meaning it was challenging for them to determine good and poor performance as they only observed one team. However, they had a reference point inside their team: the other task-solver. Accordingly, we tested the extent to which evaluators differentiated between both task-solvers in

[8] The scatter plot for the ChatGPT ratings in the writing condition is attached in the Appendix (see Figure 7).

their evaluations with regard to the difference in their performance. First, we analyzed the appraisal formats in terms of the performance difference within teams. We find that the difference in the algorithmic ratings for spoken comments significantly relates to the difference in the number of errors within teams.

To assess the robustness of these results, we conducted an additional analysis using ChatGPT. This analysis reveals a significant negative correlation between the number of individual errors and ratings for spoken comments in the between-team analysis, indicating that evaluators take the number of errors into account when speaking of performance. Additionally, we observe a significant positive correlation between error differences and corresponding rating differences for spoken comments. Our findings provide evidence that evaluators differentiate more strongly when they have a reference point and that these differentiations predominantly become obvious in spoken comments. Our results are summarized in Table 6.

Insert Table 6 about here

DISCUSSION

Our study aimed to explore the accuracy of appraisal formats with regard to capturing objective performance indicators. While rating accuracy is an issue of high relevance (e.g., Rojon et al., 2015; Speer, 2020; Spence & Keeping, 2011), to the best of our knowledge, most research has investigated the accuracy of performance appraisals from a subjective standpoint. In this context, comparing participants' ratings to expert ratings has been the most common approach (e.g., Borman, 1979; Mero et al., 2003; Yun et al., 2005). As objective performance measures are more challenging to obtain, only a few studies have investigated rating accuracy in performance appraisals by applying objective performance indicators (e.g., Sundvik & Lindeman, 1998; Kusterer &

Sliwka, 2022). However, we are not aware of studies that explored the accuracy of numerical and verbal appraisal formats as compared to objective performance.

To explore the relationship between numerical and verbal appraisal formats and objective performance, we let participants solve a picture-matching task under the observation of two evaluators. One evaluator was asked to assess the performance of both task-solvers in writing and the other evaluated the performance by speaking. The individual number of errors served as an objective indicator of performance. We decided to let one pair of evaluators observe one pair of task-solvers only, as opposed to multiple task-solvers in a row, to circumvent sequential order bias (e.g., Chernev, 2011; Criscuolo, Dahlander, Grohsjean, & Salter, 2021; Highhouse & Gallo, 1997; Page & Page, 2010). By only watching one team solve the picture-matching task, evaluators solely have one reference point to classify high and low performance, namely the other task-solver in the team. We find evidence that participants tend to give positive ratings when they do not have an overview of the entire participant field but only one team, especially when participants make relatively few mistakes. Accordingly, the differences that we discover, are most pronounced on the within-team level as this is the only way for evaluators to make comparisons.

In general, this raises the widely discussed question if performance evaluations should rather be made in absolute (compared to a predefined standard) or relative terms (comparing to other employees) (e.g., Blume, Baldwin, & Rubin, 2009; Chattopadhayay & Ghosh, 2012; Roch, Sternburgh, & Caputo, 2007). Ideally, performance ratings would stand for themselves and give individuals the chance to make judgments on qualifications based on the rating at hand. When reading that an employee received a high rating for communication skills, this rating should be a reliable indicator that this particular employee is well suited to work in settings where high communication skills are required. However, individuals have a strong tendency towards positive ratings (e.g.,

Schoenmueller et al., 2020). Thus, the school grade B might express good communication skills in itself but when the rest of the team receives straight As, the grade B gets a new context and actually indicates that the ratee performed weakest in the team. Although performance appraisals should certainly stand for themselves, we find evidence that it is challenging for individuals to define high and low levels of performance when they do not have a reference point. Only when analyzing the evaluations on the team level, we see that evaluators differentiate between different levels of performance.

Additionally, our paper demonstrates that a Large Language Model (LLM) such as ChatGPT and a supervised learning model (i.e., our algorithm) arrive at the same results when predicting performance ratings. In conducting the robustness check, we follow the recommendation to assess model reliability by comparing complex models (ChatGPT) with simpler, more interpretable approaches (Random Forest) (Feuerriegel et al., 2025). Our finding strengthens confidence in the robustness of the results, as it suggests that both generalized language understanding (LLM) and task-specific learning (supervised model) detect the same underlying patterns. By showing this alignment, the study provides evidence that LLMs may potentially serve as viable alternatives to supervised models in performance assessment tasks, particularly when the availability of labeled training data is limited. However, it is important to consider that, unlike supervised models, LLMs do not rely on explicitly defined features but rather on large-scale data, making it difficult to understand how they arrive at their predictions (e.g., Budhwar et al., 2023). Additionally, data security is not guaranteed, and firms may ultimately prefer to train their own models based on their specific evaluation criteria. Nonetheless, our findings demonstrate that machine learning approaches may generally contribute to enhancing accuracy and that self-trained models can perform effectively even with small training datasets.

Our first approach to investigating the relationship between appraisal formats with regard to objective performance has some limitations. In our study, participants solved the task quite well and only made a few errors. Correspondingly, studies with a bigger heterogeneity in performance data might come to different results. Additionally, our results may depend on the operationalization of objective performance. In this study, we used the individual number of errors and assigned equal weight to the errors. However, organizational settings have a much higher complexity than laboratory experiments. Accordingly, some errors have more serious consequences than others and are weighted differently. Also, errors are more difficult to measure, and cannot easily be assigned to one person. As a consequence, our findings may have limited generalizability to settings where errors are harder to assess and need to be tested in a field experiment.

CONCLUSION

In this paper, we sought to investigate the accuracy of numerical and verbal appraisal formats in the reflection of objective performance. Our results show that between teams, neither evaluation format accurately predicts objective performance. Within teams, however, algorithmic ratings for spoken comments are superior in predicting objective performance compared to the other evaluation formats. This underscores the importance of harnessing qualitative information in the evaluation process. Although texts are more time-consuming regarding production and processing, they have a higher informative value when it comes to differentiating performance. Accordingly, this also underlines the importance of testing methods to quantify texts in organizational settings: performance appraisals aim to get accurate information on individual performance while still obtaining data that are easy to manage. Without any insight into who performed better than the other, organizations are missing an essential basis for decision-making. This not only concerns decisions

regarding who is the best fit for a certain task but also about who requires training and lacks important skills.

Especially when applying algorithms to transfer textual information into numerical ratings, it is crucial to build trust in machine learning and to highlight the connection between the human evaluation and the numerical prediction of the algorithm. The evaluation still stems from a human evaluator while the algorithm is used to assign a number to the described performance. Also, the prediction by the algorithm does not categorically exclude the option to conduct final manual checks on the ratings.

While there is wide research interest in rating accuracy, the literature on performance appraisals shows a gap in estimating true performance with objective measures. With this study, we take a first step in closing this gap and compare the accuracy of three different appraisal formats (assigned ratings, algorithmic ratings for written comments, and algorithmic ratings for spoken comments) with the individual number of errors in a picture-matching task. By using an algorithm to transfer text into numerical ratings, we point out the usefulness of machine learning and show an exemplary use case of machine learning in an organizational setting.

Acknowledgements

This research was funded by the Ministry of Economic Affairs, Innovation, Digitalization, and Energy of the State of North Rhine-Westphalia [005-2001-0017].

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.

Angelovski, A., Brandts, J., & Sola, C. 2016. Hiring and Escalation Bias in Subjective Performance Evaluations: A Laboratory Experiment. ***Journal of Economic Behavior & Organization*****,** 121(1): 114–129.

Archak, N., Ghose, A., & Ipeirotis, P. G. 2011. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. ***Management Science*****,** 57(8): 1485–1509.

Bandiera, O., Barankay, I., & Rasul, I. 2005. Social Preferences and the Response to Incentives: Evidence from Personnel Data. ***The Quarterly Journal of Economics*****,** 120(3): 917–962.

Bayo-Moriones, A., Galdon-Sanchez, J. E., & Martinez-de-Morentin, S. 2020. Performance Appraisal: Dimensions and Determinants. ***The International Journal of Human Resource Management*****,** 31(15): 1984–2015.

Bernardin, H. J., & Pence, E. C. 1980. Effects of rater training: Creating new response sets and decreasing accuracy. ***Journal of Applied Psychology*****,** 65(1): 60–66.

Biernat, M., Tocci, M. J., & Williams, J. C. 2012. The Language of Performance Evaluations: Gender-Based Shifts in Content and Consistency of Judgement. ***Social Psychological and Personality Science*****,** 3(2): 186–192.

Bizzi, L. 2018. The problem of employees' network centrality and supervisors' error in performance appraisal: A multilevel theory. ***Human Resource Management*****,** 57(2): 515–528.

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.

Bol, J. C. 2011. The Determinants and Performance Effects of Managers' Performance Evaluation Biases. ***The Accounting Review*****,** 86(5): 1549–1575.

Bolton, G. E., Kusterer, D. J., & Mans, J. 2019. Inflated Reputations: Uncertainty, Leniency, and Moral Wiggle Room in Trader Feedback Systems. ***Management Science*****,** 65(11): 5371–5391.

Borman, W. C. 1979. Format and Training Effects on Rating Accuracy and Rater Errors. ***Journal of Applied Psychology*****,** 64(4): 410–421.

Brown, A., Inceoglu, I., & Lin, Y. 2017. Preventing Rater Biases in 360-Degree Feedback by Forcing Choice. ***Organizational Research Methods*****,** 20(1): 121–148.

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.

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Lee Cooke, F., Decker, S., DeNisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., Rogelberg, S., Saunders, M. N. K., Tung, R. L., Varma, A. 2023. Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. ***Human Resource Management Journal*****,** 33(3): 606–659.

Catano, V. M., Darr, W., & Campbell, C. A. 2007. Performance Appraisal of Behavior-Based Competencies: A Reliable and Valid Procedure. ***Personnel Psychology*****,** 60(1): 201–230.

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.

Chernev, A. 2011. Semantic Anchoring in Sequential Evaluations of Vices and Virtues. ***Journal of Consumer Research*****,** 37(5): 761–774.

Conway, J. M., & Huffcutt, A. I. 1997. Psychometric Properties of Multisource Performance Ratings: A meta-Analysis of Subordinate, Supervisor, Peer, and Self-Ratings. ***Human Performance*****,** 10(4): 331–360.

Criscuolo, P., Dahlander, L., Grohsjean, T., & Salter, A. 2021. The Sequence Effect in Panel Decisions: Evidence from the Evaluation of Research and Development Projects. ***Organization Science*****,** 32(4): 987–1008.

David, E. M. 2013. Examining the Role of Narrative Performance Appraisal Comments on Performance. ***Human Performance*****,** 26(5): 430–450.

Decotiis, T., & Petit Andre 1978. The Performance Appraisal Process: A Model and Some Testable Propositions. ***The Academy of Management Review*****,** 3(3): 635–646.

Dierdorff, E. C., & Wilson, M. A. 2003. A Meta-Analysis of Job Analysis Reliability. ***Journal of Applied Psychology*****,** 88(4): 635–646.

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): 101308.

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B., Van Bavel, J. J. 2025. Using natural language processing to analyse text data in behavioural science. ***Nature Reviews Psychology*****:** 96–111.

Frederiksen, A., Lange, F., & Kriechel, B. 2017. Subjective Performance Evaluations and Employee Careers. ***Journal of Economic Behavior & Organization*****,** 134(2–3): 408–429.

Funder, D. C., & West, S. G. 1993. Consensus, Self-Other Agreement, and Accuracy in Personality Judgment: An Introduction. ***Journal of Personality*****,** 61(4): 457–476.

Funder, D. C. 1995. On the Accuracy of Personality Judgment: A Realistic Approach. ***Psychological Review*****,** 102(4): 652–670.

Funder, D. C. (2012). Accurate Personality Judgment. ***Current Directions in Psychological Science*****,** 21(3): 177–182.

Ganu, G., Kakodkar, Y., & Marian, A. 2013. Improving the Quality of Predictions Using Textual Information in Online User Reviews. ***Information Systems*****,** 38(1): 1–15.

Golman, R., & Bhatia, S. 2012. Performance Evaluation Inflation and Compression. ***Accounting, Organizations and Society*****,** 37(8): 534–543.

Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. 2022. Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers. *Working Paper*.

Heidemeier, H., & Moser, K. 2009. Self-Other Agreement in Job Performance Ratings: A Meta-Analytic Test of a Process Model. ***Journal of Applied Psychology*****,** 94(2): 353–370.

Henemann, R. L., Moore, M. L., & Wexley, K. N. 1987. Performance Rating Accuracy: A Critical Review. ***Journal of Business Research*****,** 15(5): 431–448.

Highhouse, S., & Gallo, A. 1997. Order Effects in Personnel Decision Making. ***Human Performance*****,** 10(1): 31–46.

Hirschberg, J., & Manning, C. D. 2015. Advances in Natural Language Processing. ***Science*****,** 349(6245): 261–266.

Hoffman, C. C., Nathan, B. R., & Holden, L. M. 1991. A Comparison of Validation Criteria: Objective versus Subjective Performance Measures and Self- versus Supervisor Ratings. ***Personnel Psychology*****,** 44(3): 601–618.

Hu, N., Pavlou, P. A., & Zhang, J. 2017. On Self-Selection Biases in Online Product Reviews. ***MIS Quarterly*****,** 41(2): 449–471.

Kamphorst, J. J.A., & Swank, O. H. 2018. The role of performance appraisals in motivating employees. ***Journal of Economics & Management Strategy*****,** 27(2): 251–269.

Kusterer, D., & Sliwka, D. 2022. Social Preferences and Rating Biases in Subjective Performance Evaluations. *IZA Discussion Paper Series*.

Lazear, E. P. 2000. Performance Pay and Productivity. ***American Economic Review*****,** 90(5).

Lei, X., Qian, X., & Zhao, G. 2016. Rating Prediction Based on Social Sentiment From Textual Reviews**.** ***IEEE Transactions on Multimedia*****,** 18(9): 1910–1921.

Martell, R. F., & Leavitt, K. N. 2002. Reducing the Performance-Cue Bias in Work Behavior Ratings: Can Groups Help? ***Journal of Applied Psychology*****,** 87(6): 1032–1041.

Mas, A., & Moretti, E. 2009. Peers at Work. ***American Economic Review*****,** 99(1): 112–145.

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.

Merriam-Webster. (n.d.). *Verbal*. In ***Merriam-Webster dictionary*****.** Retrieved February 9, 2025, from https://www.merriam-webster.com/dictionary/verbal

Murphy, K. R. 2020. Performance evaluation will not die, but it should. ***Human Resource Management Journal*****,** 30(1): 13–31.

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., Layton, R. A., Pomeranz, H. R., & Schmucker, D. G. 2012. The Comprehensive Assessment of Team Member Effectiveness: Development of a Behaviorally Anchored Rating Scale for Self- and Peer Evaluation. ***Academy of Management Learning & Education*****,** 11(4): 609–630.

Page, L., & Page K. 2010. Last shall Be First: A Field Study of Biases in Sequential Performance Evaluation on the Idol Series. ***Journal of Economic Behavior & Organization,*** 73(2): 186-198.

Reber, U. 2019. Overcoming Language Barriers: Assessing the Potential of Machine Translation and Topic Modeling for the Comparative Analysis of Multilingual Text Corpora. ***Communication Methods and Measures*****,** 13(2): 102–125.

Roch, S. G., Sternburgh, A. M., & Caputo, P. M. 2007. Absolute vs Relative Performance Rating Formats: Implications for Fairness and Organizational Justice. ***International Journal of Selection and Assessment*****,** 15(3): 302–316.

Rojon, C., McDowall, A., & Saunders, M. N. K. 2015. The Relationships Between Traditional Selection Assessments and Workplace Performance Criteria Specificity: A Comparative Meta-Analysis. ***Human Performance*****,** 28(1): 1–25.

Sánchez, C. R., Díaz-Cabrera, D., & Hernández-Fernaud, E. 2019. Does effectiveness in performance appraisal improve with rater training? ***PloS One*****,** 14(9).

Schoenmueller, V., Netzer, O., & Stahl, F. 2020. The Polarity of Online Reviews: Prevalence, Drivers and Implications. ***Journal of Marketing Research*****,** 57(5): 853–877.

Silva, V. V. M., & Ribeiro, J. L. D. 2021. A Discussion on Using Quantitative or Qualitative Data for Assessment of Individual Competencies. ***Personnel Review*****,** 50(6): 1460–1478.

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.

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

Speer, A. B. 2020. Judging Performance – General Mental Ability and the Convergence of Operational Performance Ratings. ***Journal of Personnel Psychology*****,** 19(1): 44–48.

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.

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. 2005. The Impact of Job Complexity and Performance Measurement on the Temporal Consistency, Stability, and Test–Retest Reliability of Employee Job Performance Ratings. ***Journal of Applied Psychology*****,** 90(2).

Sturman, M. C., & Trevor, C. O. 2001. The Implications of Linking the Dynamic Performance and Turnover Literatures. ***Journal of Applied Psychology*****,** 86(4).

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.

Vanhove, A. J., Gibbons, A. M., & Kedharnath, U. 2016. Rater Agreement, Accuracy, and Experienced Cognitive Load: Comparison of Distributional and Traditional Assessment Approaches to Rating Performance. ***Human Performance*****,** 29(5): 378–393.

Varma, A., Pichler, S., & Srinivas, E. S. 2005. The Role of Interpersonal Affect in Performance Appraisal: Evidence From Two Samples – the US and India. ***The International Journal of Human Resource Management*****,** 16(11): 2029–2044.

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.

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

Wilson, K. Y. 2010. An Analysis of Bias in Supervisor Narrative Comments in Performance Appraisal. ***Human Relations*****,** 63(12): 1903–1933.

Woods, A. 2012. Subjective Adjustments to Objective Performance Measures: The Influence of Prior Performance. ***Accounting, Organizations and Society*****,** 37(6): 403–425.

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.

TABLES

TABLE 1
Descriptive Statistics

Variable	N	Mean	Std. Dev.	Min.	Max.
Task-solver gender (female)	80	.50	.50	0	1
Evaluator speaking condition (female)	80	.53	.50	0	1
Evaluator writing condition (female)	80	.58	.50	0	1
No. errors	80	3.05	3.89	0	22
Completion time	80	1014.48	251.27	586	1798
Assigned ratings (scale)	80	5.2	.70	3	6
Algorithmic ratings (written comments)	80	4.12	.32	3.49	4.78
Algorithmic ratings (spoken comments)	80	4.28	.35	3.34	4.87

TABLE 2
Between-Team Analysis: Correlation Matrix for No. Errors and Ratings

Variable	1.	2.	3.	4.
1. No. errors	1			
2. Assigned ratings (writing condition)	-.14	1		
3. Algorithmic ratings (writing condition)	-.06	.17	1	
4. Algorithmic ratings (speaking condition)	.03	.03	-.00	1

*p < 0.1, **p<0.05, ***p<0.01

TABLE 3
Within-Team Analysis: Correlation Matrix for Differences in No. Errors and Ratings

Variable	1.	2.	3.	4.
1. Differences in no. errors	1			
2. Differences in assigned ratings (writing condition)	.14	1		
3. Differences in algorithmic ratings (writing condition)	-.01	.01	1	
4. Differences in algorithmic ratings (speaking condition)	.29*	.07	-.25	1

*p < 0.1, **p<0.05, ***p<0.01

TABLE 4
Between-Team Analysis: Correlation Matrix for No. Errors and Ratings (using ChatGPT)

Variable	1.	2.	3.	4.
1. No. errors	1			
2. Assigned ratings (writing condition)	-.14	1		
3. Algorithmic ratings (writing condition)	-.10	.05	1	
4. Algorithmic ratings (speaking condition)	-0.22**	.18	-.04	1

*p < 0.1, **p<0.05, ***p<0.01

TABLE 5
Within-Team Analysis: Correlation Matrix for Differences in No. Errors and Ratings (Using ChatGPT)

Variable	1.	2.	3.	4.
1. Differences in no. errors	1			
2. Differences in assigned ratings (writing condition)	.14	1		
3. Differences in algorithmic ratings (writing condition)	.15	.23	1	
4. Differences in algorithmic ratings (speaking condition)	.37**	.05	-.15	1

*p < 0.1, **p<0.05, ***p<0.01

TABLE 6
Overview of Results

	Between teams	**Within teams**
Ratings	no significant correlations between no. errors and ratings	significant positive correlation between the differences in no. errors and differences in algorithmic ratings for spoken comments
ChatGPT	significant negative correlation between no. errors and ratings for spoken comments	significant positive correlation between the differences in no. errors and differences in ratings for spoken comments

FIGURES

FIGURE 1
Schematic Illustration of the Experimental Design

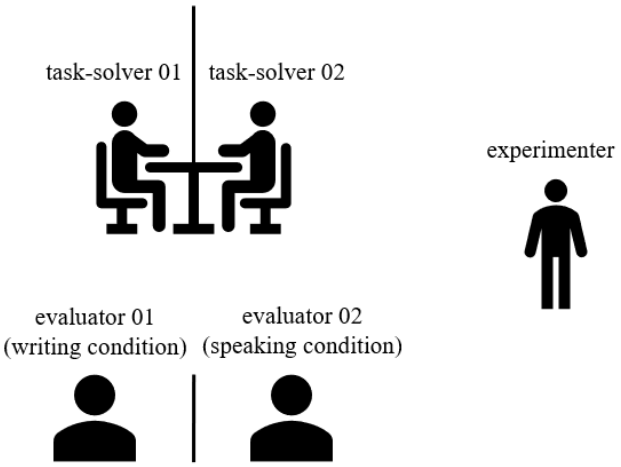


FIGURE 2
Schematic Illustration of the Empirical Approach

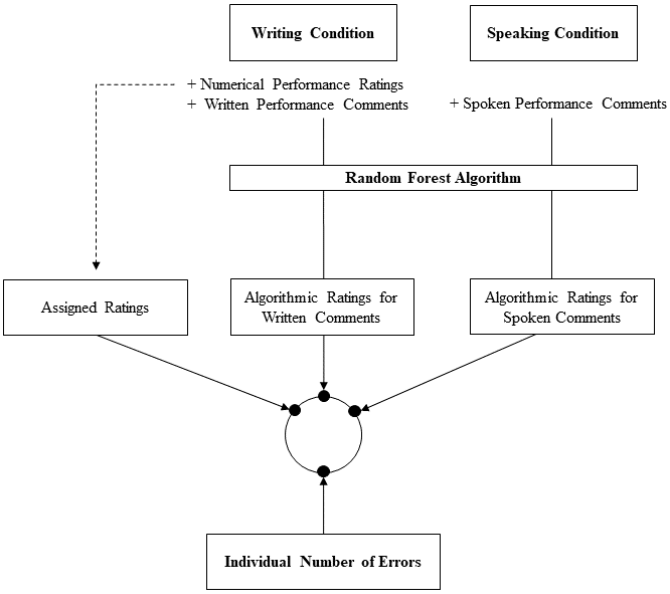


FIGURE 3
Distribution of Numerical Ratings in the Training Data Collection

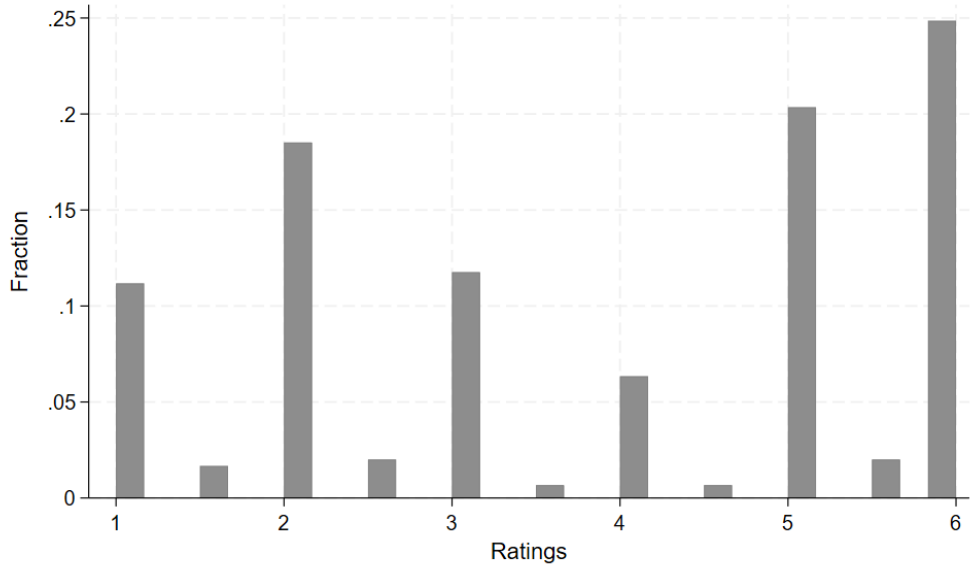


FIGURE 4
Distribution of Number of Errors in 20 Rounds

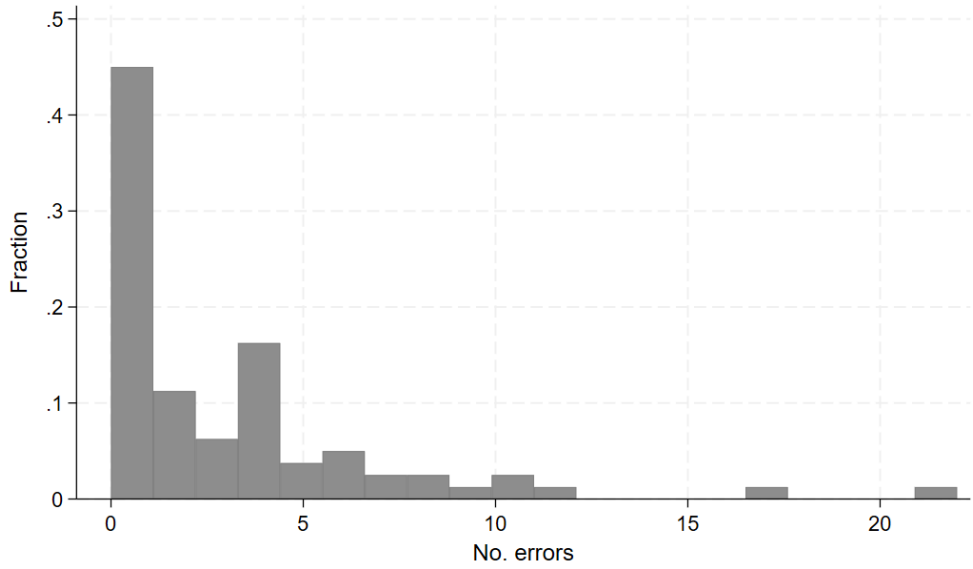


FIGURE 5
Distribution of the Ratings Assigned by the Evaluators in the Writing Condition

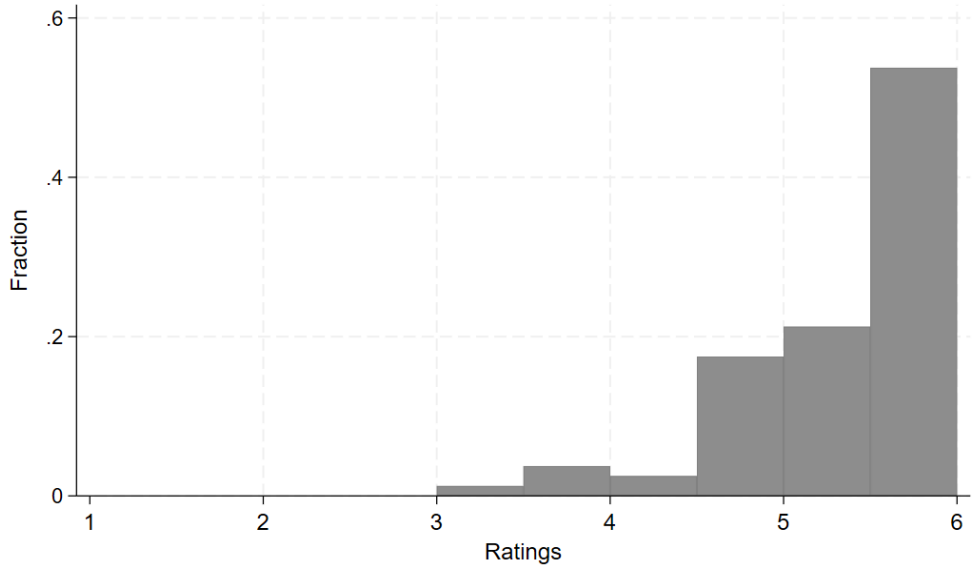


FIGURE 6
Distribution of the Text-Based Ratings for the Written Comments

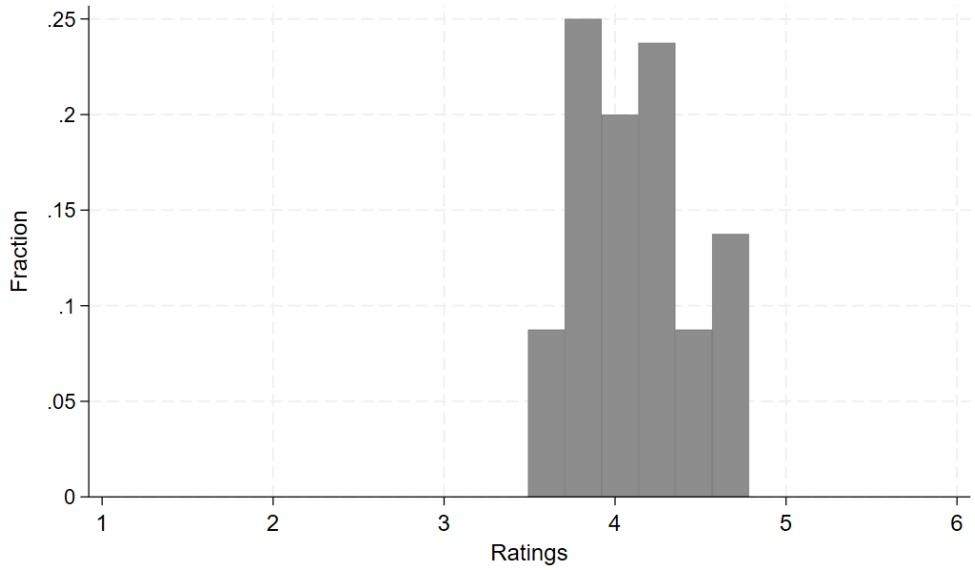


FIGURE 7
Distribution of the Text-Based Ratings for the Spoken Comments

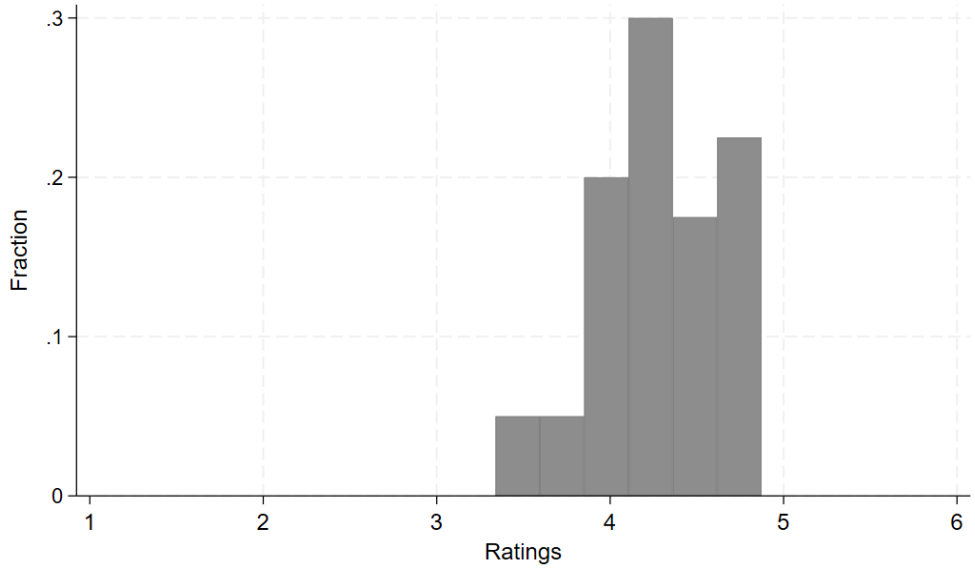


FIGURE 8
Within-Team Analysis: The Relationship Between the Difference in Errors and the Difference in Algorithmic Ratings (Spoken Comments)

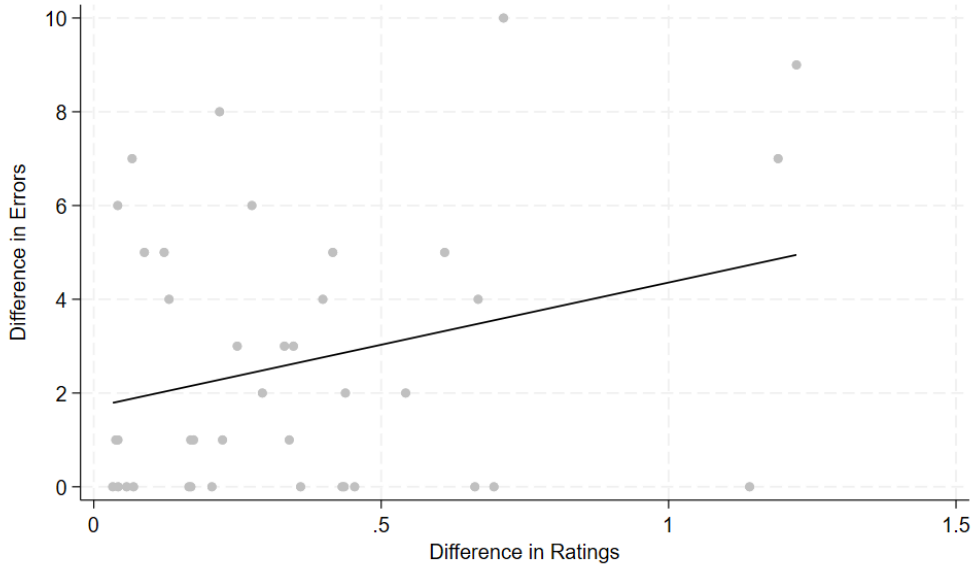


FIGURE 9
Between-Team Analysis: The Relationship Between Errors and ChatGPT Ratings (Spoken Comments)

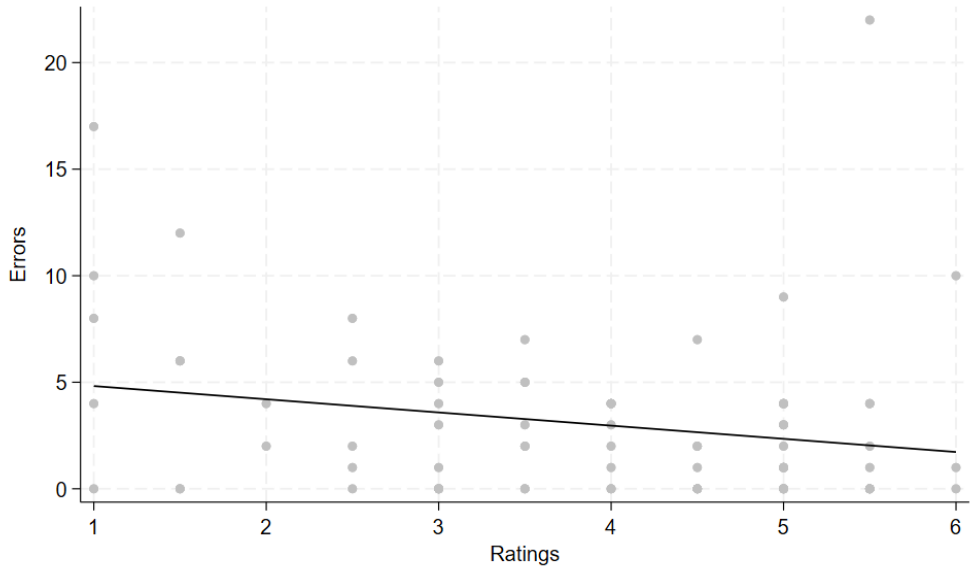
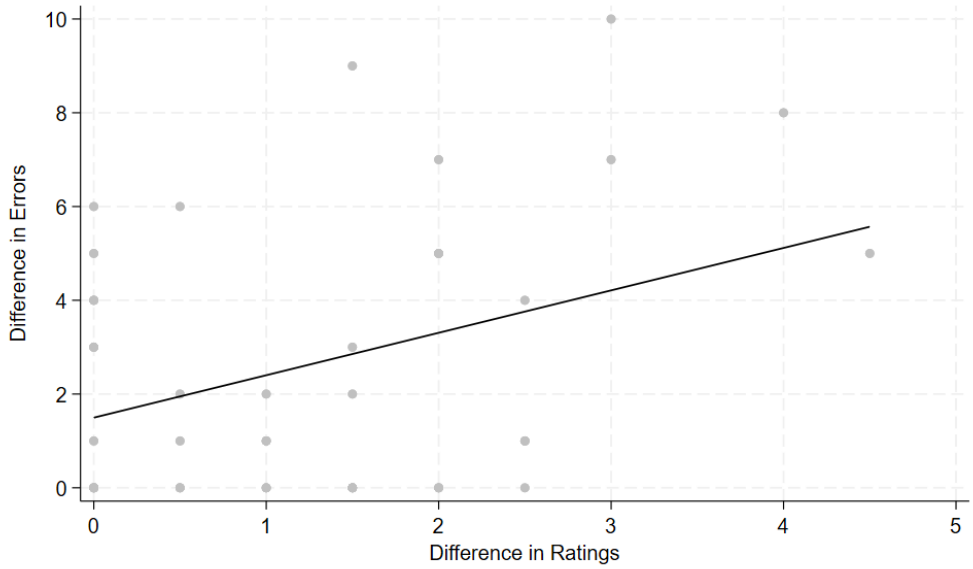


FIGURE 10
Within-Team Analysis: The Relationship Between the Difference in Errors and the Difference in ChatGPT Ratings (Spoken Comments)



APPENDIX

FIGURE 1
Between-Team Analysis: The Relationship Between Errors and Assigned Ratings

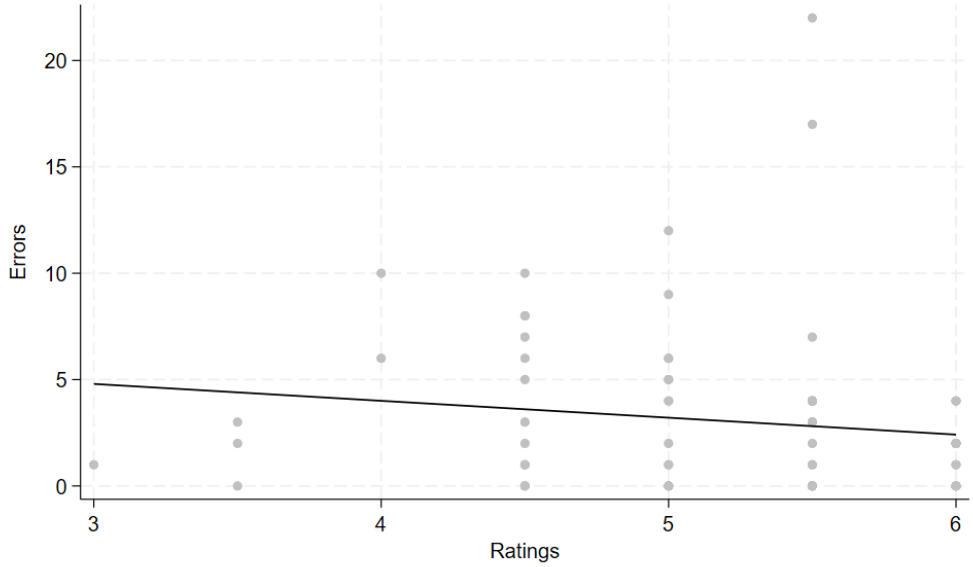


FIGURE 2
Between-Team Analysis: The Relationship Between Errors and Algorithmic Ratings (Written Comments)

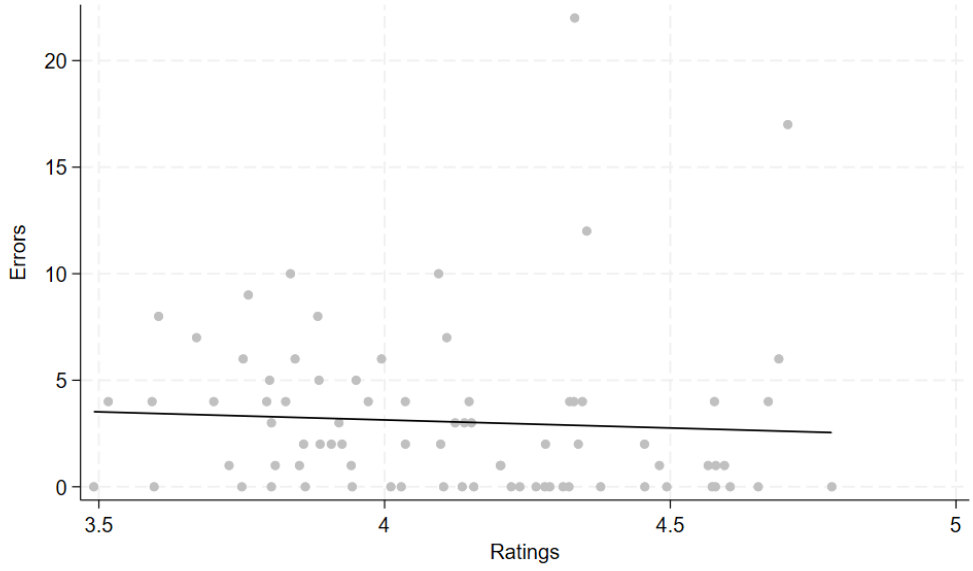


FIGURE 3
Between-Team Analysis: The Relationship Between Errors and Algorithmic Ratings (Spoken Comments)

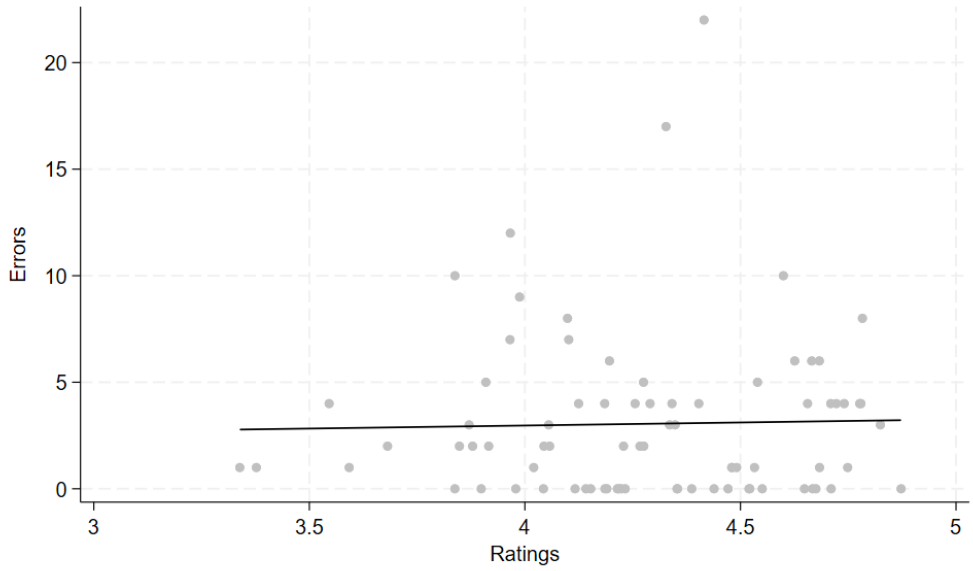


FIGURE 4
Within-Team Analysis: The Relationship Between the Difference in Errors and the Difference in Assigned Ratings

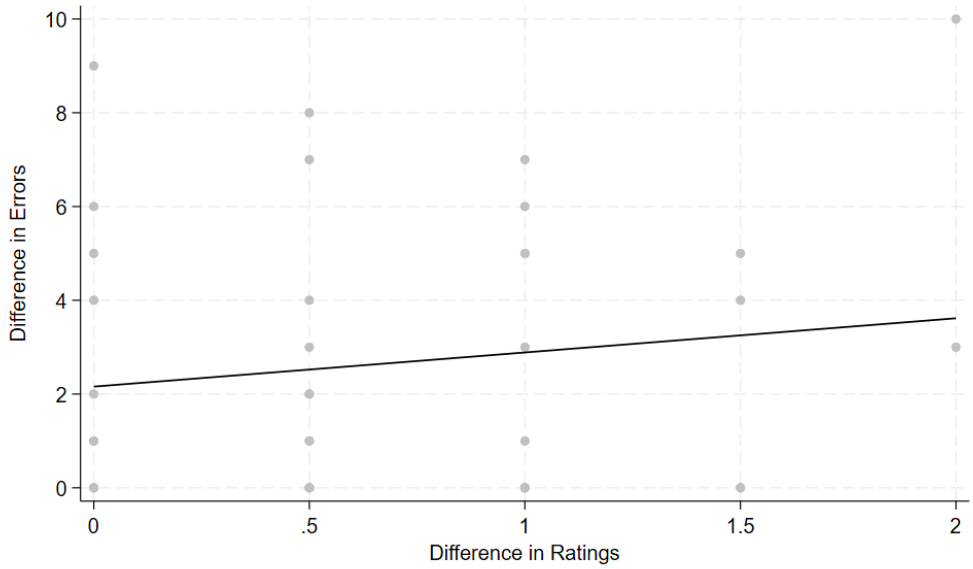


FIGURE 5
Within-Team Analysis: The Relationship Between the Difference in Errors and the Difference in Algorithmic Ratings (Written Comments)

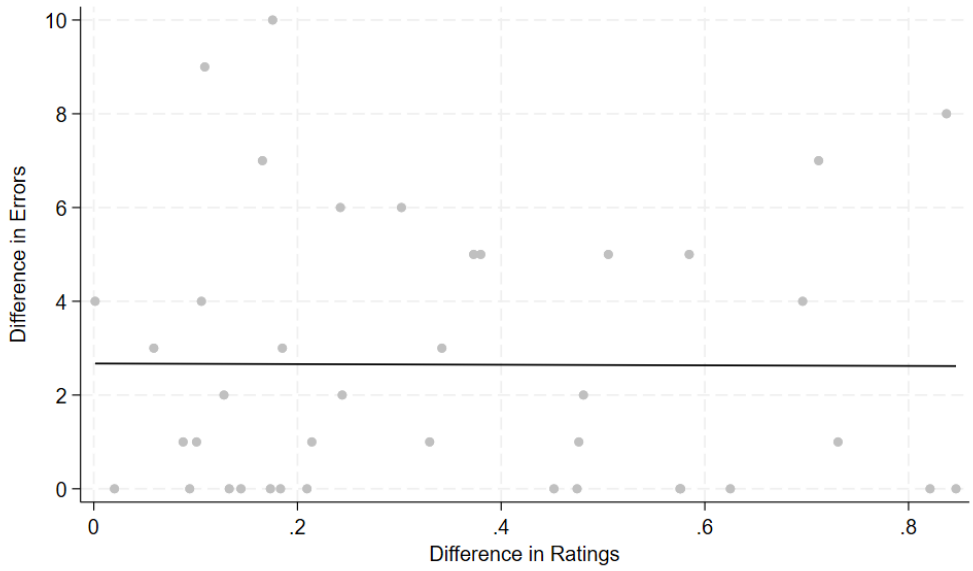


FIGURE 6
Between-Team Analysis: The Relationship Between Errors and ChatGPT Ratings (Written Comments)

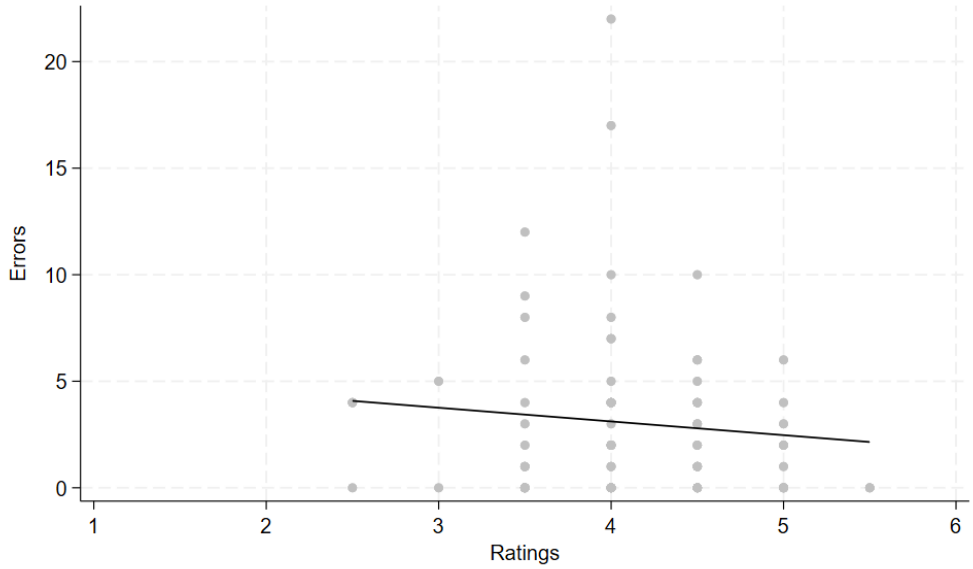


FIGURE 7
Within-Team Analysis: The Relationship Between the Difference in Errors and the Difference in ChatGPT Ratings (Written Comments)