UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FACULTY OF BUSINESS ADMINISTRATION AND ECONOMICS

Working Paper Series

Working Paper No. 2025-03

Evaluating competencies with spoken comments and machine learning

Jana Kim Gutt and Kirsten Thommes

February 2025

wiwi

Evaluating competencies with spoken comments and machine learning

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

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

Abstract

Employees are frequently evaluated on a numerical scale by their supervisors. These numerical assessments inform far-reaching managerial decisions, such as promotions, training opportunities, and dismissals. Yet, they often lack accuracy, are subject to supervisor bias, and do not provide justification for the ratings. In this paper, we address the limitations of numerical ratings by letting individuals provide spoken assessments of others and use a Random Forest algorithm to convert the spoken assessments into numbers (algorithmic ratings). Through this method, we combine the advantages of qualitative feedback and numerical ratings while potentially mitigating common biases. Our results suggest that the algorithmic ratings more accurately reflect the distribution of competencies (as measured by psychometric tests) than assigned numerical ratings (assigned ratings). The algorithmic ratings are considerably more nuanced and less skewed compared to the assigned ratings. Our findings highlight the potential of combining spoken comments with a machine learning model to enhance the accuracy of employee assessments in organizational settings.

Keywords: performance appraisals; rating prediction; machine learning; spoken comments

Funding Sources

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

1 Introduction

Assessing employees' competencies to identify gaps, foster personal development, and increase productivity is crucial for organizations in all industries (see, for instance, Alagaraja, 2012; Boxall, 1996; Datta et al., 2005; Thang et al., 2010). While many studies acknowledge the link between competence and performance (e.g., Catano et al., 2007; Fletcher, 2001), only a few studies address how organizations should assess and evaluate their employees' competencies. To this end, prior research has developed and studied psychometric scales (e.g., Morgeson et al., 2007) via self- or external assessment. The availability of psychometric scales notwithstanding, competence remains a vague construct that may include both; the underlying determinants of individual performance (such as cognitive abilities, personality, and skills) and the observable outcomes of performance, which are often the focus of organizational appraisals (Campbell & Wiernik, 2015). In practice, assessments frequently rely on single-question evaluations rather than validated scales and are typically integrated into supervisory performance appraisals, further blurring the distinction between competence and performance (Heinsman et al., 2006; Heinsman et al., 2008). Notably, the competence assessment then suffers from the same problems as performance appraisals, namely leniency and centrality biases, halo effects, favoritism, or inconsistent rating standards among evaluators (Demeré et al., 2019; Prendergast, 1999; Stauffer & Buckley, 2005). This approach adds little value to the assessment as developmental needs cannot be identified and potential rankings remain meaningless.

Even though most assessment forms consist of rating scales and comment sections, evaluations have become synonymous with numerical rating scales (Brutus, 2010). Yet, translating knowledge into a single number poses a challenge for individuals (Centeno et al., 2015). Assigning a value to multiple observations and impressions that were made over a certain period of time is demanding and possibly leads to incorrect ratings that do not reflect the individual competence (as demonstrated through performance). Since the ratings are a crucial basis for

decision-making (Merkel et al., 2021), incorrect assignments may have far-reaching consequences for the individual who was evaluated inaccurately and may prove a less solid framework for decision-making altogether. For instance, numbers cannot explain the assessment and offer little constructive guidance for the future (Rivera et al., 2023). Therefore, neither the evaluated individual nor a third party, such as the HR department, can get the intended use out of the evaluation.

In contrast, narrative comments that complement the rating scales in the appraisal can be rich in detail, provide context, and allow for more extensive descriptions (Speer, 2021). They are perceived as more helpful to employees who rather seek narrative feedback than mere numerical scores (Rivera et al., 2023). However, they fail to provide a ranking, and comments about two employees cannot be easily compared. Consequently, they serve as a more explanatory source of information than numerical ratings but may be less suited to get a comparable overview of the whole workforce.

Steering the middle course would be a combination of verbal explanations and numerical ratings. However, writing comments and manually transferring them into numbers is time-consuming and might be error-prone. Comments need to be written, read, and analyzed to utilize the information conveyed. Yet, with machine learning (ML), we might be able to effectively tap into the richness of qualitative data and automated data processing. Consequently, this paper shows a possible way to translate comments into numerical ratings via ML.

Considering the limited research on this subject and the benefits of narrative comments (Brutus, 2010; David, 2013; Speer, 2018), we aim to contribute to the literature by proposing ML methods as a useful means to bridge the gap between the advantages of numerical rating scales and comments. Specifically, we train and test a ML algorithm that predicts numerical ratings based on spoken comments. This way, we hope to (1) circumvent the problem of translating knowledge into numbers, (2) reduce the time and effort to produce narrative comments, and (3) test if spoken comments might be a better starting point for reflecting competence than

assigned numeral ratings. We focus mainly on interpersonal competencies that are particularly difficult to measure (Duckworth & Yeager, 2015). Accordingly, the study concentrates on communication skills, empathy, trust, adaptability, and cooperativeness. These competencies have been considered valuable teamwork skills (Mathieu et al., 2014; Salas et al., 2005; Stevens & Campion, 1994) and serve as exemplary competencies in this study.

In broad terms, we collect training data and train a machine learning (ML) algorithm using a Random Forest model. We then compare the predictions of our ML algorithm (hereafter referred to as algorithmic ratings) and the assigned ratings to a competence assessment using a psychometric test, which we take as the ground truth in this study. Our analysis provides preliminary evidence that algorithmic ratings more accurately reflect individual competencies (as measured by psychometric tests) than ratings assigned by team members. Overall, we show that ML can be used to translate verbal descriptions into numerical ratings. The method maintains verbal reasoning and eliminates parts of the numerical ratings' leniency and centrality bias.

The remainder of the paper is structured as follows: In Section 2, we present the related literature (2 Related literature). Section 3 presents the method and results (3 Method and results), and Sections 4 and 5 discuss (4 Discussion) and conclude (5 Conclusion) our work.

2 Related literature

2.1 Challenges in employee evaluation

To decide which employee is the best fit for a certain project or task, it is crucial to know and communicate their strengths and weaknesses (e.g., Maley, 2024; Sniezek et al., 2002). Yet, the assessment of employee competencies faces challenges on several levels concerning the (1) conceptualization, (2) subjectivity, and (3) format.

(1) A key issue in performance appraisals is the conflation of competence, which refers to an employee's underlying knowledge, skills, and abilities (Campion et al., 2011), with performance, which reflects their observable work behaviors (Campbell & Wiernik, 2015). While

these two concepts are related, they are distinct, and failing to separate them can lead to misleading assessments (Kurz & Bartram, 2002). Competence refers to what an employee is capable of, whereas performance is influenced not only by competence but also by external factors such as task difficulty, team support, and workplace conditions (e.g., Motowidlo et al., 1997). Assessing competence based solely on performance can result in inaccurate evaluations. Employees working in well-structured environments with strong managerial support may appear more competent, while equally skilled employees in challenging conditions may receive lower ratings due to circumstances beyond their control. Additionally, short-term fluctuations in performance**,** caused by workload variations or organizational changes, can further distort perceptions of an employee's actual abilities.[1]

(2) Since objective performance measures are not available for most complex tasks, performance appraisals consist of subjective judgments (Manthei & Sliwka, 2019). Correspondingly, human cognition and social processes become influencing factors of the appraisal (e.g., Lavigne, 2018). Spence and Keeping (2011) state that rating accuracy is not necessarily each evaluator's goal. They distinguish between conscious and subconscious rating distortion in the appraisal process. Evaluators may consciously assign inaccurate ratings because they, for example, want to avoid confrontation, wish to be perceived as a successful manager, or need to comply with organizational norms. Contrarily, subconscious rating distortion refers to the evaluator's cognitive process of forming a judgment (Spence & Keeping, 2011) and includes subconscious biases. The annual cycle of most performance appraisals (David, 2013) puts further demand on the evaluator's cognition as individuals tend to struggle with transferring their knowledge into a single number (Centeno et al., 2015) and possess a limited ability to process and communicate large amounts of information (Hitt & Brynjolfsson, 1997).

[1] While our study acknowledges the need to distinguish between competence and performance, it does not aim to theoretically delineate these constructs. Instead, we adopt a pragmatic approach by mirroring organizational practice, in which evaluators infer competence from observable behavior. Within this framework, we examine which appraisal format most accurately reflects participants' competencies by having evaluators provide performance appraisals focused on interpersonal competencies.

(3) Translating knowledge into a number also concerns the format of performance appraisals: Despite there being no prototypical appraisal process, most evaluations consist of both rating scales and comment sections but are widely associated with only numerical ratings (Brutus, 2010). As rating scales are closed-ended and only cover a limited range, the reflection of an employee's actual competence and respective performance is presumably reduced. Even though practice and research unanimously coincide that the current way of assessing competence and performance needs to be improved, suggestions from research to use standardized psychometric tests that are validated go unheard as they are typically lengthy scales that need some statistical knowledge to process the data. Moreover, they still miss the explanation for a certain assessment as the result is a number and does not provide reasoning. Although comment sections usually follow the rating scales, appraisal comments have scarcely been investigated (Silva & Ribeiro, 2021).

In this study, we build on the limited scholarly attention to appraisal comments by examining them from a new perspective, with a particular focus on their relationship to competence assessments through psychometric tests. As the lack of objective indicators persists, subjective appraisals remain the main measure to estimate employees' strengths and weaknesses. Inevitably, appraisals continue to be subject to human errors in judgment. However, by commenting on competence rather than assigning a number on a scale, some of these errors might be reduced. We investigate whether greater assessment accuracy might be achieved through the format of appraisal comments rather than through numerical ratings. While we acknowledge that social processes and personal goals undoubtedly play a role in organizational appraisal processes, we do not aim to solve the whole puzzle but study one piece, namely the appraisal format of comments, in detail.

2.2 Ratings, comments, and ML

2.2.1 Ratings

Numerical ratings are easy to process but lack the information depth of comments (Brutus, 2010). Previous studies state that employees pay more attention to comments than to numbers on a scale (Rivera et al., 2023; Smither & Walker, 2004), as comments offer reasoning and context (Speer, 2018). Brutus (2010), for instance, argues that receiving a message like "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) would be more informative than just receiving a below-par rating. However, the example also shows that, frequently, performance and competencies are mixed as Bob's behavior may be a result of his set of interpersonal competencies or his motivation, both eventually resulting in his evaluation.

The ease of processing numerical ratings particularly benefits HR departments as they need to compare a large number of appraisals. Moreover, ranking individuals in terms of performance is a popular means of gaining information and allows for arranging employees in a meaningful way (Gill et al., 2019). Consequently, employee ratings have been found to be an influential factor in hiring decisions on online platforms (Kokkodis & Ransbotham, 2023). This notwithstanding, performance ratings in organizations can be substantially biased. Rivera et al. (2021) investigated real-time performance feedback and found that supervisors use tit-for-tat strategies in numerical ratings for employees. In the same context, Petryk et al. (2022) find that the organizational relationship between rater and ratee biases the performance rating. Several studies have also been conducted on performance ratings in online labor markets. Category-specific performance ratings of workers in such markets can be used to predict future performance ratings (Kokkodis & Ipeirotis, 2016; Kokkodis & Ransbotham, 2023). These studies use performance ratings as inputs for Linear Dynamic Systems to predict future ratings (Kokkodis & Ipeirotis, 2016).

2.2.2 Comments and ML

Notwithstanding the easy use of numerical ratings, textual assessments might be more informative and have a higher predictive power (Archak et al., 2011). Still, appraisal comments have received relatively little attention in research (Silva & Ribeiro, 2021) and practice. Manual text processing poses challenges concerning the evaluators' time and the underlying coding scheme. All comments from different supervisors need to be coded homogeneously and cannot be subject to individual interpretation. Also, comments need to be transferred into numbers to allow for manageability and ranking.

To circumvent the shortcomings of manual coding, scholars have drawn upon ML approaches. In organizational contexts, Campbell and Shang (2022) used ML to predict corporate risk of misconduct based on employee reviews on Glassdoor. They developed misconduct word indices with a weighted regression approach. Subsequently, they used the word indices to show their power to predict misconduct above and beyond known observable firm characteristics. Hickman et al. (2021) trained ML algorithms to predict personality traits based on interview answers. They used the Linguistic Inquiry and Word Count (LIWC) to analyze the sentiments of the answers. With the sentiments and self- and other-rated personality traits, they trained the algorithms. The authors report that the algorithm is well-suited to predict other-reported personality traits.

A recurring theme in studies on performance appraisals is that comments are mostly positive (Smither & Walker, 2004; Wilson, 2010) and that highly positive comments are often inconsistent with the respective numerical ratings (Wilson, 2010). One source of inconsistency between comments and ratings stems from the ratee's ethnicity. In particular, Wilson (2010) has investigated whether similar comments lead to similar ratings across ethnic groups. She found that this is not the case for certain ethnic groups. Given the same number of positive comments received, specific ethnic groups receive lower numerical ratings, suggesting the existence of a discrimination bias.

In contrast to the existing studies on performance ratings and comments, to the best of our knowledge, no studies have applied a ML algorithm for text-based rating predictions in the appraisal process thus far. Silva and Ribeiro (2021) quantified performance comments manually and found low correlations between the comments and numerical ratings. Other existing studies have investigated the determinants and consequences of qualitative performance feedback. Prior research has found that the favorability of comments positively impacts performance (David, 2013; Smither & Walker, 2004). For example, one study discovered that the sentiment score of narrative performance appraisals explains a statistically significant share in the variation of the following year's numerical performance ratings of the ratee (Speer, 2018). This suggests that narrative elements can be a driver of ratee motivation that translates into performance increases in the next year. Other studies have characterized which comments are particularly substantial for ratee performance, observing that favorable comments and comments exhibiting interactional justice are most effective in stimulating future ratee performance (David, 2013). Perhaps most closely related to our study is the work of Kokkodis and Ransbotham (2023). This study uses performance comments to predict task-specific performance ratings for future jobs (e.g., building a website vs. creating a social media marketing campaign) using ML approaches. Yet, our study notably differs from their work in that we leverage spoken comments and differentiate between ratings and competencies measured through psychometric tests.

We propose that contemporary advancements in ML may be a way forward to combine the advantages of appraisal comments while still using numerical ratings in ranking individuals. To evaluate whether comments (1) may reflect true competence more closely than assigned ratings and (2) can be converted into numerical ratings for management purposes, we have collected data in terms of comments, numerical ratings, and competencies via psychometric tests. We focus on the assessment of interpersonal competencies as they are more likely to be regularly assessed, are needed in most work situations, and are usually directly related to

performance (Heckman & Kautz, 2012; Traylor et al., 2022). To enhance manageability, we opted to collect spoken comments instead of written ones. As writing is slower than speaking and puts higher demands on the working memory (Kellogg, 2007), we expect spoken comments to be less time-consuming for the evaluator and more valuable in content.

3 Method and results

In the first section of our study, we train two ML algorithms for rating prediction. The advantage of using an algorithm to predict numerical ratings for comments lies in the prediction not being confined by researchers' choice of certain words or sentiment operationalizations. Therefore, we use a holistic approach to analyze the relationship between ratings and comments. To this end, we employ the supervised machine-learning techniques Random Forest and XGBoost as they have proven to perform well in natural language prediction tasks (Hossain et al., 2021; Zhu et al., 2020). In the second section of our study, we test the algorithm on a separate dataset (see 3.2 Out-of-sample prediction) that is unrelated to the training data collected in the first section. By comparing the algorithmic ratings with the assigned ratings, we examine how the two differ from one another. However, the comparison does not allow for judgment on the rating quality. In performance appraisals, rating quality is conceptualized in terms of convergence with other performance measures (Speer, 2020). As interpersonal competencies are difficult to measure objectively (Duckworth & Yeager, 2015), we investigated how the algorithmic ratings and the assigned ratings relate to the results of psychometric tests (see 3.3 Results). The tests are directed at empathy, communication skills, trust, cooperativeness, and adaptability and match the competencies that participants were asked to assess verbally and numerically.

3.1 Training the algorithm

3.1.1 Data collection

We collected the data for this study via an online survey. Specifically, we collected spoken comments and numerical ratings on team members with exceptionally high or low interpersonal competencies. The comments and ratings were captured as voice recordings. The online survey was conducted from January to March 2020. Participants were recruited via convenience sampling. We chose an initial seed of participants to which we distributed the survey link. Respondents were then encouraged to forward the link to other potential participants. Each respondent was tasked with thinking about the best and the weakest team member they have worked with in the past. Specifically, they were asked to provide spoken assessments on eight key competencies of their team members: empathy, communication skills, trust, cooperativeness, social adaptability, task-related adaptability, learning orientation, and goal orientation. Additionally, two introductory questions about teamwork were included. Out of the initial 104 respondents, 77 completed the survey. After data cleaning, 1,194 answers were considered for further analysis. Although the data collection was conducted in German, for the subsequent analyses, the transcripts were translated into English. For the translation, we used the machine translator DeepL, which has previously been described as the "rising star in the field of […] machine translation providers" (Reber, 2019, p. 103). Even though a word-by-word translation cannot be achieved, inaccuracies have been stated to be of less concern with approaches in which words are analyzed on an individual level, like the TF-IDF approach that will be introduced in the following (Reber, 2019).

For our investigations that followed, it was important to obtain relatively equally distributed data. To accurately predict both high and low numerical ratings, we need training data that include ratings across the full range from 1 to 6. Yet, prior literature has demonstrated that people are hesitant to use the entire rating scale in rating situations and rather use positive ratings (Hu et al., 2017). To circumvent this and make participants use ratings all across the scale,

we deliberately asked for the most and the least competent team member the participants had experienced with regard to the eight competencies. They were free to talk about any person that came to mind and assign a numerical rating. The ratings followed the German school grading system (1 = very good, 6 = very poor), as it was expected to be familiar to participants. For each competence, one question regarding high and one regarding low performance was asked:

> "Describe the communication of the team member with the strongest communication skills. In which situations did the person communicate particularly strongly? Rate his/her communication skills with a grade from 1 to 6."
>
> "How did the team member with the weakest communication skills perform? Rate his/her communication skills with a grade from 1 to 6."

Table 1 displays the descriptive statistics on the response level. The data include the variables assigned rating, age, gender, response duration, and the number of characters for each answer. Out of the 77 respondents who completed the survey, the answers of 22 respondents did not meet the criteria for analysis (e.g., assigning a numerical rating). Thus, we considered the completed surveys of 55 participants, including 33 females, 20 males, and 2 respondents preferring not to specify their gender. In our analyses, we also studied the answers of respondents who did not complete the full survey in case their answers met our analysis criteria. As the sociodemographic data were queried at the end of the survey, the age and gender of respondents who dropped the survey were not collected.

Insert Table 1 about here

In Figure 1, we plotted the assigned ratings. As intended, the distribution shows little skewness as compared to typical distributions in a rating context. Most participants rated their team members' competence with a rating of 1 or 2. Slightly fewer participants gave a 5. Even though the questions were directed at the best and weakest performance, the participants assessed the perceived competence relatively rarely with a rating of 6. This distribution supports the assumption that individuals are more hesitant to give poor ratings than good ones.

Nevertheless, we gather from this distribution that our method of data collection was successful in receiving ratings and comments for good and poor performance. Therefore, we consider the dataset as suitable for training the ML algorithm.

Insert Figure 1 about here

3.1.2 Training process

As the dependent variable is a continuous variable from 1 to 6, we model the prediction as a regression problem. To process the text data and develop our model, we used the scikit-learn library in Python. Scikit-learn offers a wide variety of ML algorithms and provides extensive support for text processing (Varoquaux et al., 2015). First, we set all characters to lowercase. Then we removed standard English stop words from the texts, using the snowball stemmer of the scikit-learn library. We also explored using stemming to reduce the remaining words to their stems. However, our evaluation metric Root Mean Square Error (RMSE) was higher after stemming, which is why we refrained from stemming.

To predict the numerical rating of each comment, we operationalized a feature vector with 6,000 features. The feature vector can be regarded as independent variables used to explain the dependent variables. To construct the feature vector, we used Term Frequency Inverse Document Frequency (TF-IDF), one of the most common schemes for term weighting (Polpinij & Luaphol, 2021). TF-IDF tokenizes the documents, learns the vocabulary, and inverses the document frequency weightings (Hossain et al., 2021). Soucy and Mineau (2005) explain the weighting process in more detail and state that, in the first step, term frequency (TF) in a document is incorporated. Next, the IDF measures the infrequency of a word in the entire training collection. The authors further state that a highly frequent term in the text collection is not considered representative of the document, similar to stop words. Consequently, if a term appears infrequently in the text collection, it is considered relevant to the text collection (Soucy & Mineau, 2005).

For the rating prediction, we applied Random Forest and XGBoost models. Both models are ML methods widely used for prediction purposes (Oughali et al., 2019). Drawing on previous studies, we used 80 % of the data for training and 20 % for testing purposes (Hossain et al., 2021; Lei et al., 2016). The prediction performance of our training set was measured in RMSE. The lower the RMSE, the more accurate the prediction. Our RMSE is 0.94 for Random Forest and 0.69 for XGBoost. In summary, this indicates one can use comments to predict their attached numerical ratings reasonably well. Our RMSE falls in the same range as those obtained in the computer science literature that predict numerical ratings based on reviews. For example, Ganu et al. (2013) achieved an RMSE between 1.149 and 1.231 when predicting individual restaurant ratings (on a scale from 1 to 5) based on written text. Although the XGBoost model showed a lower RMSE on the training data than the Random Forest, we found that XGBoost predictions sometimes deviated from the information provided in the comment, as shown in the example below:

> "The team member showed himself to be very cooperative. Always tried to include the other solution proposals in order to then find a common solution path." (Transcript ID: K105.000213)

While the participant who made this comment gave a rating of 1, XGBoost predicted a much lower rating of 5. However, the transcript does not show any negative references that would explain this discrepancy. These results suggest that the XGBoost model overfits the training data. This means that the model works well on the training dataset but does not generalize to new data (Vabalas et al., 2019). Although no algorithm is flawless, this pattern was much less pronounced in the Random Forest (that predicted a 2.55 for the particular comment). Therefore, we decided to run the Random Forest on the out-of-sample data.

3.2 Out-of-sample prediction

To further test the model's performance, we conducted an out-of-sample prediction. The aim of the out-of-sample prediction is to demonstrate that our Random Forest does not merely

achieve a high training set performance but that the performance also generalizes to separate data in a similar context. Especially for predictions involving spoken language, this task is essential because spoken language can vary drastically between samples and human beings. Hence, the applicability of our algorithm across samples might be put into question.

For the out-of-sample prediction, we collected data on the participants' competencies, as well as spoken comments, and assigned ratings on their perceptions of each team member's competence. In this context, the teamwork consisted of solving an online escape game in randomized groups of four to five. Escape games are particularly well-suited as a team task, as they require close cooperation and interaction among participants. Before solving the escape game, participants filled in psychometric tests concerning empathy, communication skills, trust, cooperativeness, and adaptability. After the game, we asked the participants to speak about each team member's competence and assign a corresponding numerical rating from 1 to 6 (without the option to pick gradations, e.g., 1.5), with 1 being the best and 6 the weakest rating. The competencies evaluated by the team members were congruent with the assessed competencies before the game. For this data collection, the questions were not directed at competence extremes (best – worst) but asked for a skill assessment in general. We decided to give two examples for each competence to avoid misunderstandings.

> "How did the team member communicate? Explain your assessment. Think, for example, about aspects like comprehensibility, listening, etc. [On the next page] Rate the communication skills of the team member with a school grade."

We recruited participants in lectures at a medium-sized European university. For participation, students received course credits and financial compensation depending on the escape game performance. We started with the first round of data collection in September 2021 and collected the last data in April 2022. The responses were recorded and transcribed in German. In order to be able to run the algorithms on the audio comments, the German transcripts were translated into English as in the training data collection.

Of 240 participants who finished the evaluations, 231 recorded clean audio comments we could transcribe. We excluded comments with the following errors: (1) the comment was cut mid-sentence, (2) the participant did not refer to the specific team member the question was directed at but rather spoke about teamwork in general, (3) the participant described a different team member than the team member that the question was directed at, (4) the participant indicated that they were not able to answer the question, (5) the participant systematically gave the same answer for all their team members (e.g., "the team member was adaptable"), (6) the audio recording contained too much background noise to understand what was being said, and (7) it was incomprehensible what the participant tried to say (e.g., "was able to put himself into others' shoes well and accordingly showed himself likable as a team"). Additionally, all comments of participants were systematically excluded if (8) the scale was consistently applied incorrectly (which became obvious from comparing the comment with the assigned rating) or (9) the participant stated to struggle with identifying the team members.

In the end, 3,177 transcripts from 211 participants were considered for further analysis. The standard team size was four, but in some cases, the participants solved the task in teams of five. Accordingly, each subject received evaluations from three to four team members concerning five competencies. Overall, 227 participants received comments that matched the inclusion criteria. The discrepancy between the numbers (211 and 227) derives from the fact that some participants did not provide an evaluation for their team members (10), some systematically applied the scale incorrectly (4), and others were not able to remember their team members and could not tell who was who (2). Still, these participants filled in the psychometric tests and received evaluations from their team members which is why the number of evaluators is smaller than the number of evaluated individuals. The mean age was 23.93 (SD 3.19), and the participants were split into 126 female and 101 male participants.

3.3 Results

We compared the assigned ratings with the algorithmic ratings and came to the results displayed in Figure 2 and Figure 3.

Insert Figure 2 about here

In Figure 2, we plotted the assigned ratings from the out-of-sample collection as a histogram. The histogram shows the ratings across all competencies instead of one competence in particular. Participants were asked to assess their team members' competence with school grades from 1 to 6. The histogram demonstrates that participants rarely rated their team members' competence with a 6 and only very few with a 5. Predominantly, ratings of 1 or 2 were given. This might point to a positivity bias in the participants' rating behavior (Merkel et al., 2021).

Insert Figure 3 about here

Figure 3 shows a histogram of the algorithmic ratings for the out-of-sample data. The algorithmic ratings start at 1.498 and end at 4.725. We find the highest frequency between 2.5 and 3. Comparing the assigned ratings in Figure 2 with the algorithmic ratings in Figure 3 suggests that the tendency toward positivity persists as ratings of 5 and 6 were not predicted at all. Yet, the algorithmic ratings congregate around the area between 2.5 to 3 and are more differentiated. We are aware that participants did not have the option to differentiate between degrees and their assigned ratings, therefore, did not allow for gradations, e.g., 2.75. However, scales are naturally closed-ended and only allow respondents to choose from given options. Thus, the Random Forest predictions indicate that comments enhance precision and shift the focus away from overly positive assessments.

To estimate the quality of the algorithmic and the assigned ratings, we compared their relation to the individual competence we measured with psychometric tests. Since rating accuracy is challenging to determine (DeNisi & Murphy, 2017), we consult psychometric test results

as an indicator of true competence (ground truth). We expect the measured competencies to be related to the observable behavior during the escape game, i.e., we assume that individuals with high communication skills (as measured by the psychometric test) showed high communication skills during the escape game. For this comparison, we included data from 227 participants who completed the psychometric tests and played the escape game. 3,177 transcripts were considered for analysis.[2]

To allow for comparison between the psychometric test results, the assigned ratings, and the algorithmic ratings, we inverted the values of the ratings (which initially were coded as "1" being the best rating and "6" being the worst), calculated the average of all values, and mean-centered them. By mean-centering the values, we are able to compare ratings from different scales, i.e., we are able to compare six-point scales to a nine-point scale from a psychometric test.

We then compared how accurately assigned ratings and algorithmic ratings reflect the psychometric test results. Overall, we find a relatively low correlation between assigned ratings and psychometric tests, as well as algorithmic ratings and psychometric test results; no correlation coefficient is greater than 0.4. Table 2 shows the exemplary results for communication skills.

Insert Table 2 about here

However, we still observe differences in how the assigned and algorithmic ratings relate to the test results. In particular, we examine whether assigned and algorithmic ratings can correctly identify high and low psychometric test scores. To further explore the association, we calculate rating accuracy by subtracting the rating from the test score for every competence. Table 3 depicts the results for communication skills. Several observations stand out. The share

[2] Table A1 in Appendix A depicts the descriptive statistics of the competence assessment and the psychometric tests used.

of ratings that are in the range of -0.5 to 0.5 of deviation from the psychometric test score is much larger for the algorithmic ratings than for the assigned ratings. This means that algorithmic ratings – compared to assigned ratings – have a larger number of observations that deviate only slightly from the psychometric test. Moreover, for the classes of large deviations from the psychometric test score, such as -2.5 to -1.5 or 1.5 to 2.5, the assigned ratings have a larger share of observations than algorithmic ratings. This further substantiates that algorithmic ratings have higher accuracy in reflecting psychometric test scores than assigned ratings.

Insert Table 3 about here

To visualize the difference in accuracy, we also display the distribution of the psychometric test scores and differences in assigned and algorithmic ratings in Figure 4 and Figure 5, respectively.

Insert Figure 4 about here

Insert Figure 5 about here

Values of 0 indicate that the ratings correspond to the test scores whereas negative values imply that the rating is higher than the test score. Contrarily, positive values signal that the test score is higher than the rating. A comparison of the histograms reveals that the algorithmic ratings are more tightly centered around 0 and do not show as many outliers. By contrast, the assigned ratings are more dispersed away from 0, indicating lower precision. To summarize, on average, the algorithmic ratings show a higher accuracy in classification than assigned ratings. We come to similar results for the other competencies.[3]

[3] The results for the other competencies are shown in Appendix B - E.

4 Discussion

Our study aims to bridge the gap between the benefits of assessment ratings and comments through the use of a ML algorithm. By comparing the algorithmic and assigned ratings to psychometric test results, we find evidence that the algorithmic ratings reflect the test scores from scale-based competence assessments more closely and reduce leniency in terms of overly positive ratings.

In line with previous research (e.g., Prendergast, 1999; Spence & Keeping, 2011), our results support the view that individuals show a tendency towards leniency and centrality when assigning ratings. Consistent with earlier studies, our comments are not without positivity (Smither & Walker, 2004; Wilson, 2010). The algorithmic ratings do not reflect the full range of ratings as 5 and 6 were not predicted at all. Yet, the algorithmic ratings congregate around 2.5 and 3 (as compared to 1 and 2 in the assigned ratings) and show higher differentiation. When considering that comments are just as subjective as ratings and are made by the same person, it would be rather surprising to find no tendency towards positivity at all. Comments rely just as much on human perception as ratings. However, we show that the extent of certain biases can be reduced and that the level of differentiation can be increased by switching to a different appraisal format, i.e., comments.

The reasons why evaluation comments may be less prone to biases can be diverse. One potential explanation might be the context as comments help to clarify the reasoning behind individual judgments. The need for explicitly stating reasons in comments may lead to less bias and more substantiated evaluations. Also, numerical scales require the evaluator's subjective interpretation, i.e., what constitutes a rating of 4 vs. a rating of 5, whereas comments are a more direct medium to express thoughts and observations. Additionally, by articulating their thoughts, evaluators are encouraged to reflect more carefully as compared to giving a quick numerical rating. Lastly, numerical ratings are often influenced by anchoring effects

(Thorsteinson et al., 2008). While evaluators may use prior ratings as a reference point for subsequent ratings, their comments can stand independently. Arguably, comments might also be associated with higher accountability since evaluators need to reflect on the assessment in more detail and the evaluation seems more personal and traceable. As prior research has pointed out, accountability encourages evaluators to give more accurate ratings (Mero & Motowidlo, 1995).

By applying an algorithm for rating prediction, we address previous calls for future research to identify analysis methods for qualitative feedback in the appraisal process (e.g., Doldor et al., 2019). Our results tie well with previous studies wherein textual data provide more information than rating scales and have a higher predictive power (Archak et al., 2011; Rivera et al., 2023).

In our talk to managers, they have frequently revealed that the current methods of evaluation are time-consuming and not helpful as biases hinder effective responses. Running these rating systems leads organizations to collect data with limited practical value: The data do not reflect the actual competencies and fail to identify leverage points for improvements, as especially low performers are not identified (Golman & Bhatia, 2012). The idea of utilizing comments (maybe also from different sources as in 360-degree evaluations) seemed promising to them as this may reduce bias in judgments.

While our study highlights how the interplay of ML and comments enhances the appraisal process in organizational settings, our results might generalize to different contexts where accurate ratings are highly important. Beyond reviews on employers, products, or services, individuals also assign numerical ratings in medical (e.g., Karcioglu et al., 2018), psychological (e.g., Lesage et al., 2012), or risk assessments (e.g., Brody et al., 2004). Correctly rating the level of pain, stress, or air quality is crucial for deciding which steps to take next. In these contexts, taking the spoken comments as a starting point might also benefit the accuracy of the assessments. Of course, medical professionals do not exclusively rely on their patient's numerical indication of pain but listen to their verbal descriptions as well. However, the

descriptions might be a more reliable starting point, especially if an algorithm would be able to learn the patient's individual speaking style and tendencies to exaggerate or understate. Together with the professional's knowledge about the patient, algorithmic predictions might benefit the decision process.

Our first attempt to showcase how ML may improve organizational decision-making has some limitations: First and foremost, we assume that the psychometric test is the ground truth. This might not be correct as the "true" competencies may not have been observable for the evaluator during the task. Yet, this limitation is true for assigned ratings and comments and, therefore, should not systematically vary between assessment modes. Secondly, gaming the algorithm may occur if evaluators learn how to influence an algorithmic rating verbally. Gaming the system, however, does already occur in numerical rating systems where it is arguably also much easier.

Finally, ML inevitably triggers an ethical debate: Is it ethical to collect spoken comments and predict numerical ratings? What if the rating does not reflect the evaluator's intention? This is a crucial discussion on a general level, especially since judging the fairness of an algorithmic decision is a major challenge on its own (Kallus et al., 2022). How should human and algorithmic output be combined? Should humans always be allowed to overrule algorithmic advice? Should they be at least forced to verify? These are important questions that need to be addressed not only by scientists but also by society. However, they are not at the core of our research. For consideration, we add that existing rating and evaluation systems may also face the same ethical questions.

5 Conclusion

Our paper contributes to the scientific discourse in a meaningful way by exploring the largely neglected connections between comments and numerical ratings in an organizational context. Considering the limited time available to evaluators and the need for numerical ratings

in terms of data processing, we tested a way of reconciling the richness of qualitative information with the straightforward usability of quantitative data. By collecting voice recordings, we studied an appraisal format relatively uncommon today. While the benefits of spoken recollections have been recognized in other contexts, e.g., criminal investigations (Sauerland & Sporer, 2011), the potential of spoken evaluations in an organizational setting has not yet been fully explored. As spoken comments are spontaneous, less thought out, and possibly less censored than written text, they allow for an evaluation rich in information and low in production effort. Additionally, the open format of spoken comments allows the evaluator to freely decide which aspects to emphasize and gives important information on the behaviors that stood out the most. On an organizational level, the information collected in appraisal comments might yield essential knowledge on the qualities evaluators particularly focus on when reflecting on good and poor competence. Through our analyses, we made the first step in revealing the opportunities offered by spoken comments to assess employee competence. Our results bear valuable implications for improving the appraisal process in organizations.

References

Alagaraja, M. (2012). HRD and HRM Perspectives on Organizational Performance. *Human Resource Development Review, 12*(2), 117–143. https://doi.org/10.1177/1534484312450868.

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. https://doi.org/10.1287/mnsc.1110.1370

Beierlein, C., Kemper, C. J., Kovaleva, A., & Rammstedt, B. (2012). *Kurzskala zur Messung des zwischenmenschlichen Vertrauens: Die Kurzskala Interpersonales Vertrauen (KUSIV3)*. (GESIS-Working Papers, 2012/22). https://www.ssoar.info/ssoar/handle/document/31212

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

Brody, S. D., Peck, B. M., & Highfield, W. E. (2004). Examining Localized Patterns of Air Quality Perception in Texas: A Spatial and Statistical Analysis. *Risk Analysis, 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, D. W., & Shang, R. (2022). Tone at the Bottom: Measuring Corporate Misconduct Risk from the Text of Employee Reviews. *Management Science, 68*(9), 7034–7053. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3850554

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

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing Competencies Well: Best Practices in Competency Modeling. *Personnel Psychology, 64*(1), 225–262. https://doi.org/10.1111/j.1744-6570.2010.01207.x

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. https://doi.org/10.1111/j.1744-6570.2007.00070.x

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

Datta, D. K., Guthrie, J. F. & Wright, P. M. (2005). Human resource management and labor productivity: Does industry matter? *Academy of Management Journal, 48*(1), 135–145. https://doi.org/10.5465/amj.2005.15993158

David, E. M. (2013). Examining the Role of Narrative Performance Appraisal Comments on Performance. *Human Performance, 26*(5), 430–450. https://doi.org/10.1080/08959285.2013.836197

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

DeNisi, A. S., & Murphy, K. R. (2017). Performance Appraisal and Performance Management: 100 Years of Progress? *Journal of Applied Psychology, 102*(3), 421–433. https://doi.org/10.1037/apl0000085

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

Duckworth, A. L., & Yeager, D. S. (2015). Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. *Educational Researcher, 44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Fletcher, C. (2001) Performance appraisal and management: The developing research agenda. *Journal of Occupational and Organizational Psychology, 74,* 473 – 487. https://doi.org/10.1348/096317901167488

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. https://doi.org/10.1016/j.is.2012.03.001

Gill, D., Kissová, Z., Lee, J., & Prowse, V. (2019). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Management Science, 65*(2), 494–507. https://doi.org/10.1287/mnsc.2017.2907

Golman, R., & Bhatia, S. (2012). Performance Evaluation Inflation and Compression. Accounting, *Organizations and Society, 37*(8), 534–543. https://doi.org/10.1016/j.aos.2012.09.001

Heckman, J. J., & Kautz, T. D. (2012). Hard Evidence on Soft Skills. Working Paper. National Bureau of Economic Research, Cambridge. https://doi.org/10.1016/j.labeco.2012.05.014

Heinsman, H., de Hoogh, A. H. B., Koopman, P. L., & van Muijen, J. J. (2006). Competency Management: Balancing Between Commitment and Control. *Management Revue, 17*(3), 292–306. https://doi.org/10.5771/0935-9915-2006-3-292

Heinsman, H., de Hoogh, A. H. B., Koopman, P. L., & van Muijen, J. J. (2008). Commitment, Control, and the Use of Competency Management. *Personnel Review, 37*(6), 609–628. 10.1108/00483480810906865

Hickman, L., Saef, R., Ng, V., Woo, S. E., Tay, L., & Bosch, N. (2021). Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews. *Human Resource Management Journal, 34*(2). https://doi.org/10.1111/1748-8583.12356

Hitt, L. M., & Brynjolfsson, E. (1997). Information Technology and Internal Firm Organization: An Exploratory Analysis. *Journal of Management Information Systems, 14*(2), 81–101. https://www.jstor.org/stable/40398267

Hossain, M. I., Rahman, M., Ahmed, T., & Islam, A. Z. M. T. (2021). Forecast the Rating of Online Products from Customer Text Review Based on Machine Learning Algorithms. 2021 International Conference on Information and Communication Technology for Sustainable Development.

Hu, N., Pavlou, P. A., & Zhang, J. (2017). On Self-Selection Biases in Online Product Reviews. *MIS Quarterly, 41*(2), 449–471. https://www.jstor.org/stable/26629722

Jackson, C. L., Colquitt, J. A., Wesson, M. J., & Zapata-Phelan, C. P. (2006). Psychological Collectivism: A Measurement Validation and Linkage to Group Member Performance. *Journal of Applied Psychology, 91*(4), 884–899. https://doi.org/10.1037/0021-9010.91.4.884

Kallus, N., Mao, X., & Zhou, A. (2022). Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. *Management Science, 68*(3), 1959–1981. https://doi.org/10.1287/mnsc.2020.3850

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

Kellogg, R. T. (2007). Are Written and Spoken Recall of Text Equivalent? *The American Journal of Psychology, 120*(3), 415–428. https://doi.org/10.2307/20445412

Kokkodis, M., & Ipeirotis, P. G. (2016). Reputation transferability in online labor markets. *Management Science, 62*(6), 1687-1706. https://doi.org/10.1287/mnsc.2015.2217

Kokkodis, M., & Ransbotham, S. (2023). Learning to successfully hire in online labor markets. *Management Science, 69*(3), 1597-1614. https://doi.org/10.1287/mnsc.2022.4426

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.

Lavigne, E. (2018). How structural and procedural features of managers' performance appraisals facilitate their politicization: A study of Canadian university deans' reappointments. *European Management Journal, 36*(5), 638–648. https://doi.org/10.1016/j.emj.2018.06.001

Lei, X., Qian, X., & Zhao, G. (2016). Rating Prediction Based on Social Sentiment from Textual Reviews. *IEEE Transactions on Multimedia, 18*(9), 1910–1921. https://doi.org/10.1109/tmm.2016.2575738

Lesage, F., 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

Maley, J. F. (2024). Contextual complexity: Understanding unique appraisal reactions of top subsidiary managers. *European Management Journal.* https://doi.org/10.1016/j.emj.2024.09.006

Manthei, K., & Sliwka, D. (2019). Multitasking and Subjective Performance Evaluations: Theory and Evidence from a Field Experiment in a Bank. *Management Science, 65*(12), 5861–5883. https://doi.org/10.1287/mnsc.2018.3206

Mathieu, J. E., Tannenbaum, S. I., Donsbach, J. S., & Alliger, G. M. (2014). A Review and Integration of Team Composition Models. *Journal of Management, 40*(1), 130–160. https://doi.org/10.1177/0149206313503014

Merkel, S., Chan, H. F., Schmidt, S. L., & Torgler, B. (2021). Optimism and Positivity Biases in Performance Appraisal Ratings: Empirical Evidence from Professional Soccer. *Applied Psychology, 70*(3), 1100–1127. https://doi.org/10.1111/apps.12266

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://doi.org/10.1037/0021-9010.80.4.517

Monge, P. R., Bachman, S. G., Dillard, J. P., & Eisenberg, E. M. (1981). Communicator Competence in the Workplace: Model Testing and Scale Development. *Annals of the International Communication Association, 5*(1), 505–527. https://doi.org/10.1177/002194368902600202

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the Use of Personality Tests in Personnel Selection Contexts. *Personnel Psychology, 60*(3), 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x

Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (2014). A Theory of Individual Differences in Task and Contextual Performance. In W.C. Borman, & S.J. Motowidlo (Eds.), *Organizational Citizenship Behavior and Contextual Performance* (pp. 71–83). Psychology Press.

Oughali, M. S., Bahloul, M., & El Rahman, S. A. (2019). Analysis of NBA Players and Shot Prediction Using Random Forest and XGBoost Models. 2019 International Conference on Computer and Information Sciences (ICCIS). https://doi.org/10.1109/iccisci.2019.8716412

Paulus, C. M. (2009). Der Saarbrücker Persönlichkeitsfragebogen SPF(IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index. https://doi.org/10.23668/psycharchives.9249

Petryk, M., Rivera, M., Bhattacharya, S., Qiu, L., & Kumar, S. (2022). How network embeddedness affects real-time performance feedback: An empirical investigation. *Information Systems Research, 33*(4), 1467-1489. https://doi.org/10.1287/isre.2022.1110

Ployhart, R. E., & Bliese, P. D. (2006). Individual Adaptability (I-ADAPT) Theory: Conceptualizing the Antecedents, Consequences, and Measurement of Individual Differences in Adaptability. In Burke, C. S., Pierce, C., & Salas, E. (Eds.), *Understanding Adaptability: A Prerequisite for Effective Performance within Complex Environments* (pp. 3–39). Emerald Group Publishing Limited.

Polpinij, J., & Luaphol, B. (2021). Comparing of Multi-Class Text Classification Methods for Automatic Ratings of Consumer Reviews. In P. Chomphuwiset, J. Kim, & P. Pawara (Eds.), International Conference on Multidisciplinary Trends in Artificial Intelligence, 2021, 164–175.

Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature, 37*(1), 7–63. https://doi.org/10.1257/jel.37.1.7

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. https://doi.org/10.1080/19312458.2018.1555798

Rivera, M., Jiang, C., & Kumar, S. (2023). Seek and Ye Shall Find: An Empirical Examination of the Effects of Seeking Real-Time Feedback on Employee Performance Evaluations. *Information Systems Research, 35*(2), 783–806. https://ssrn.com/abstract=3553906

Rivera, M., Qiu, L., Kumar, S., & Petrucci, T. (2021). Are traditional performance reviews outdated? An empirical analysis on continuous, real-time feedback in the workplace. *Information Systems Research, 32*(2), 517-540. https://doi.org/10.1287/isre.2020.0979

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is There a "Big Five" in Teamwork? *Small Group Research, 36*(5), 555–599. https://doi.org/10.1177/1046496405277134

Sauerland, M., & Sporer, S. L. (2011). Written vs. Spoken Eyewitness Accounts: Does Modality of Testing Matter? *Behavioral Sciences and the Law, 29*(6), 846–857. https://doi.org/10.1002/bsl.1013

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. https://doi.org/10.1108/pr-08-2019-0444

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

Sniezek, J. A., Wilkins, D. C., Wadlington, P. L., & Baumann, M. R. (2002). Training for Crisis Decision-Making: Psychological Issues and Computer-Based Solutions. *Journal of Management Information Systems, 18*(4), 147–168. https://doi.org/10.1080/07421222.2002.11045704

Soucy, P., & Mineau, G. W. (2005). Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. 19th International Joint Conference on Artificial Intelligence, 2005, 1130–1135.

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

Speer, A. B. (2020). Judging Performance – General Mental Ability and the Convergence of Operational Performance Ratings. *Journal of Personnel Psychology, 19*(1), 44–48. https://doi.org/10.1027/1866-5888/a000244

Speer, A. B. (2021). Scoring Dimension-Level Job Performance from Narrative Comments: Validity and Generalizability When Using Natural Language Processing. *Organizational Research Methods, 24*(3), 572–594. https://doi.org/10.1177/1094428120930815

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

Stevens, M. J., & Campion, M. A. (1994). The Knowledge, Skill, and Ability Requirements for Teamwork: Implications for Human Resource Management. *Journal of Management, 20*(2), 503–530. https://doi.org/10.1111/j.1468-2389.2012.00578.x

Thang, N. N., Quang, T., & Buyens, D. (2010). The Relationship Between Training and Firm Performance: A Literature Review. *Research and Practice in Human Resource Management, 18*(1), 28–45.

Thorsteinson, T. J., Breier, J., Atwell, A., Hamilton, C., & Privette, M. (2008). Anchoring effects on performance judgments. *Organizational Behavior and Human Decision Processes, 107*(1): 29-40. https://doi.org/10.1016/j.obhdp.2008.01.003

Traylor, A. M., Reyes, D. L., & Holladay, C. L. (2022). Do We Practice What We Preach? The Association Between Judgements of Soft Skills and Performance Evaluations Over Time. *Current Psychology, 41*(10), 7208–7214. https://psycnet.apa.org/doi/10.1007/s12144-020-01276-0

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine Learning Algorithm Validation with a Limited Sample Size. *PloS One, 14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Varoquaux, G., Louppe, G., Pedregosa, F., Buitinck, L., Grisel, O., & Mueller, A. (2015). SCIKIT-LEARN: Machine Learning Without Learning the Machinery. *GetMobile: Mobile Computing and Communications, 19*(1), 29–33. https://doi.org/10.1145/2786984.2786995

Wilson, K. Y. (2010). An Analysis of Bias in Supervisor Narrative Comments in Performance Appraisal. *Human Relations, 63*(12), 1903–1933. https://psycnet.apa.org/doi/10.1177/0018726710369396

Zhu, A., Li, R., & Xie, Z. (2020). Machine Learning Prediction of New York Airbnb Prices. 2020 Third International Conference on Artificial Intelligence for Industries, 1–5. https://doi.org/10.1109/ai4i49448.2020.00007

Tables

TABLE 1 Descriptive statistics of the training data

Variable	N	Mean	Std. Dev.	Min.	Max.
Assigned rating	1,194	3.165829	1.799731	1	6
Age	1,058	27.52174	5.223482	19	56
Gender (female)	1,060	.640566	.480061	0	1
Response duration (seconds)	1,194	46.30402	35.74904	4	346
Number of characters	1,194	516.9464	454.0558	28	4776

TABLE 2 Correlation matrix for communication skills

	1.	2.	3.
1. Test results communication			
2. Assigned ratings communication	-.01		
3. Algorithmic ratings communication	-.03	.33***	

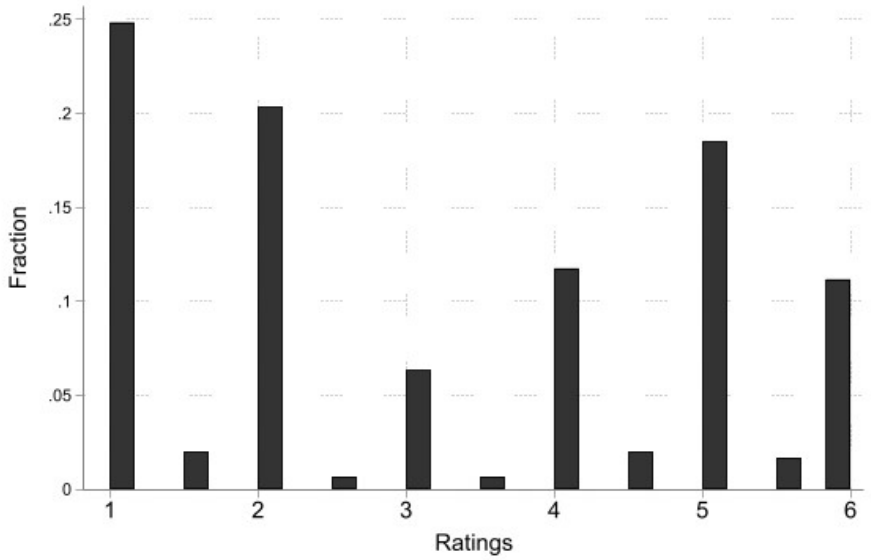
*p < 0.1, **p<0.05, ***p<0.01

TABLE 3 Share of observations deviating from the test results for communication skills

Deviation from Test Results	Share Assigned Rating	Share Algorithmic Rating
<-3.5	0	0
[-3.5; -2.5]	0.44	0
[-2.5; -1.5]	3.96	0.88
[-1.5; -0.5]	23.79	23.35
[-0.5; 0.5]	44.05	51.98
[0.5; 1.5]	18.94	22.91
[1.5; 2.5]	7.49	0.88
[2.5; 3.5]	0.88	0
>3.5	0.44	0

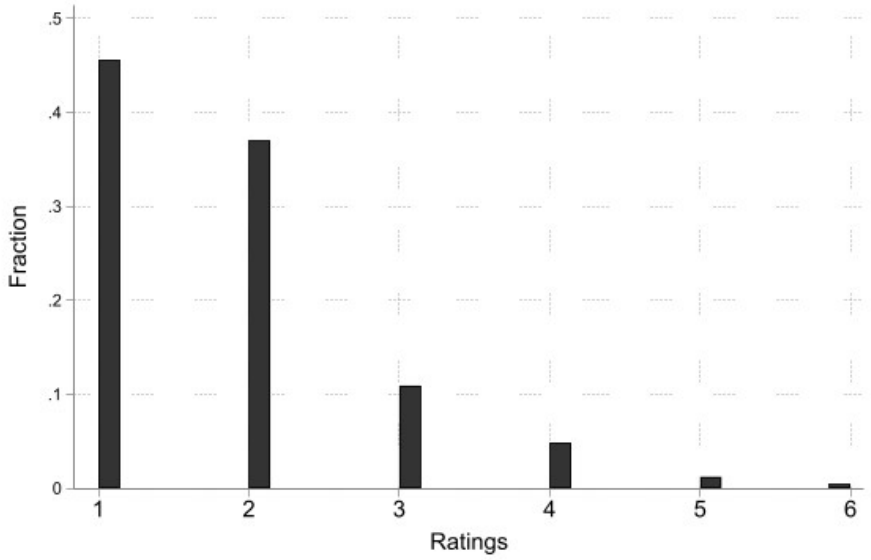
Figures

FIGURE 1 Distribution of participants' ratings in the training data



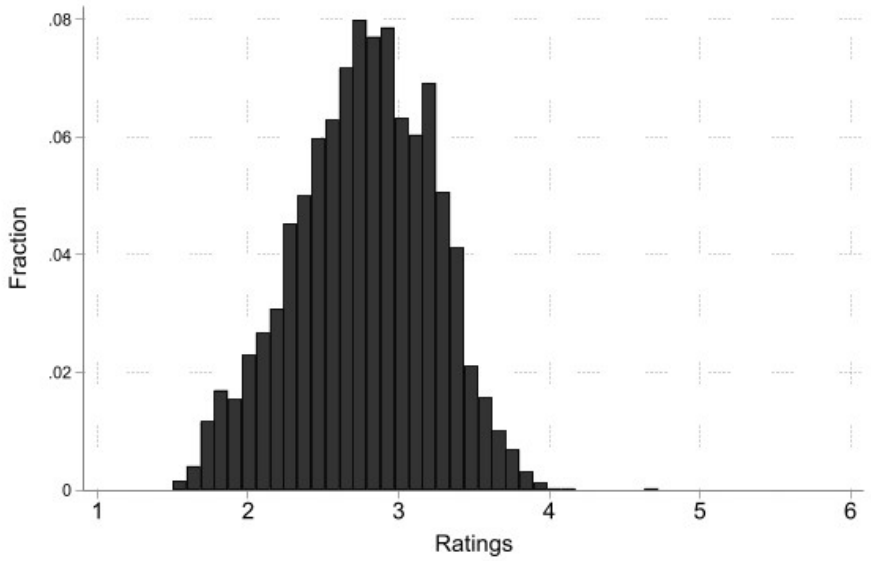
Note: This figure depicts the distribution of respondents' assigned ratings from the training data set. The ratings are based on the German school grading system (1 = very good, 6 = very poor). The distribution includes decimal numbers for cases in which participants assigned a 1-, for example.

FIGURE 2 Distribution of the assigned ratings in the out-of-sample data



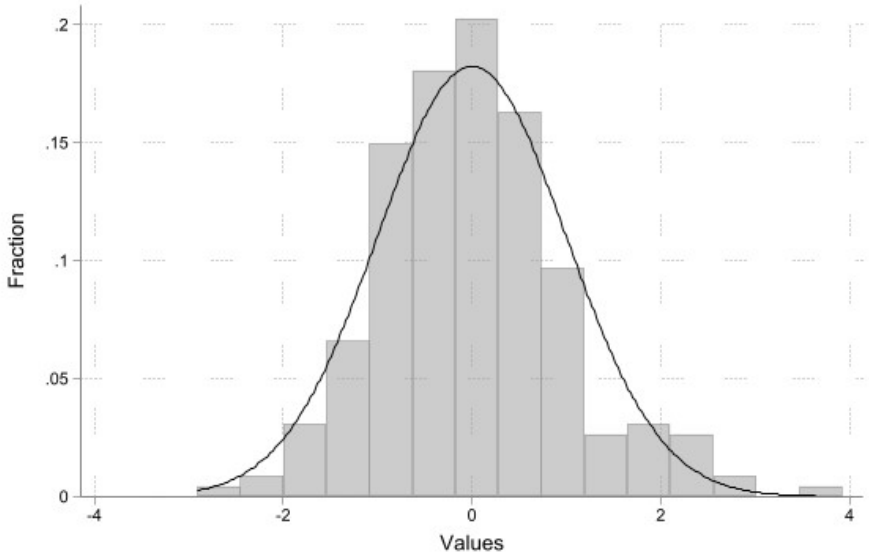
Note: The ratings are based on the German school grading system (1 = very good, 6 = very poor). Ratings did not include decimal numbers.

FIGURE 3 Distribution of algorithmic ratings for the out-of-sample data



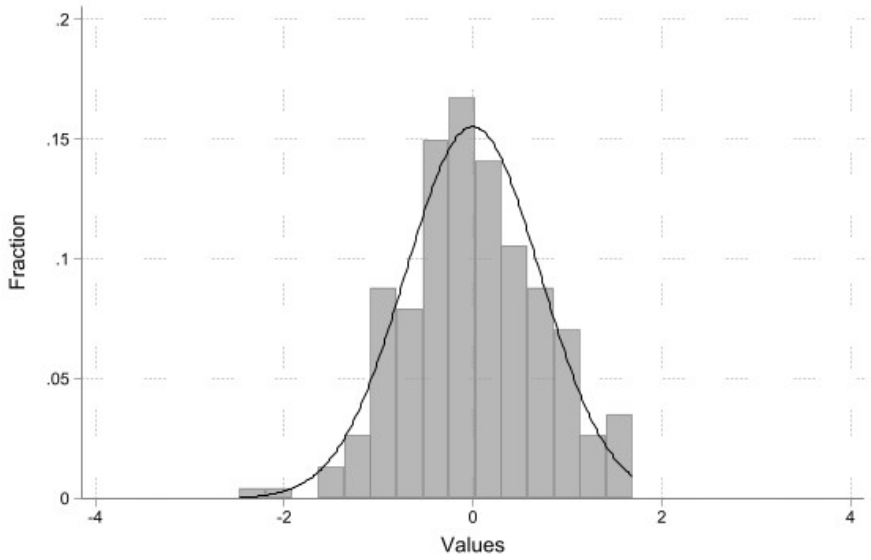
Note: The algorithmic ratings are based on the German school grading system (1 = very good, 6 = very poor).

FIGURE 4 Accuracy test results – assigned ratings for communication skills



Note: This figure depicts the deviation of the assigned from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.

FIGURE 5 Accuracy test results – algorithmic ratings for communication skills



Note: This figure depicts the deviation of the algorithmic ratings from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.

Appendix A

TABLE A1 Descriptive statistics of the psychometric tests

Competence	Psychometric Test	Mean	Std. Dev.	Min.	Max.	Obs.
Communication	Communicator competence questionnaire (Monge et al. 1981)	5.4	.62	3.33	6.83	227
Adaptability	*Work stress* and *interpersonal* subscales of the i-adapt measure (Ployhart and Bliese 2006)	3.75	.39	2.75	4.83	227
Cooperation	Psychological collectivism measure (Jackson et al. 2006)	3.33	.56	1.4	4.8	227
Trust	German short scale for interpersonal trust (Beierlein et al. 2012)	3.34	.75	1.33	5	227
Empathy	German short version of the interpersonal reactivity index (Paulus 2009)	3.56	.61	1.75	4.75	227

Appendix B: Adaptability

TABLE B1 Correlation matrix for adaptability

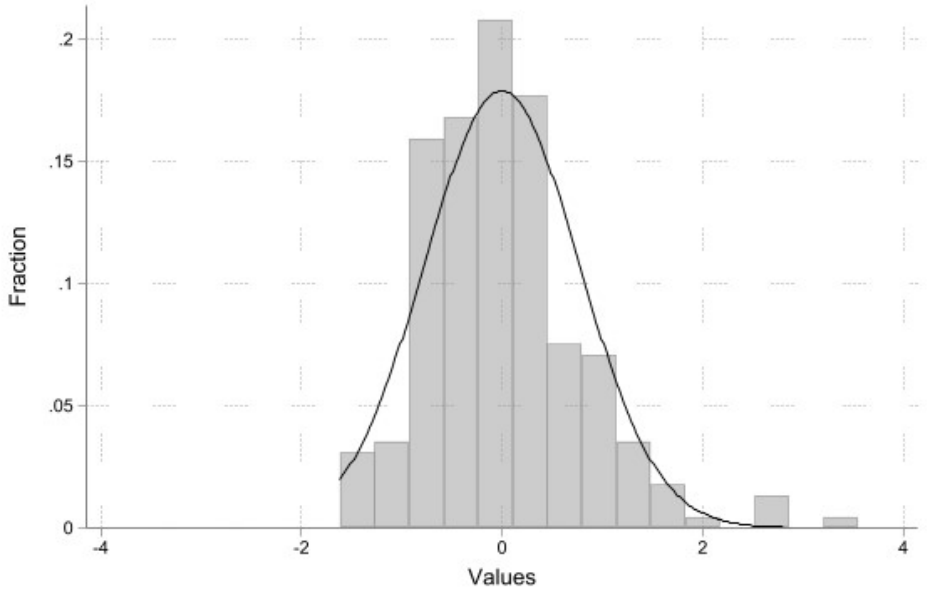
	1.	2.	3.
1. Test results adaptability			
2. Assigned adaptability	-.07		
3. Algorithmic ratings adaptability	-.03	.21***	

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

TABLE B2 Share of observations deviating from the test results for adaptability

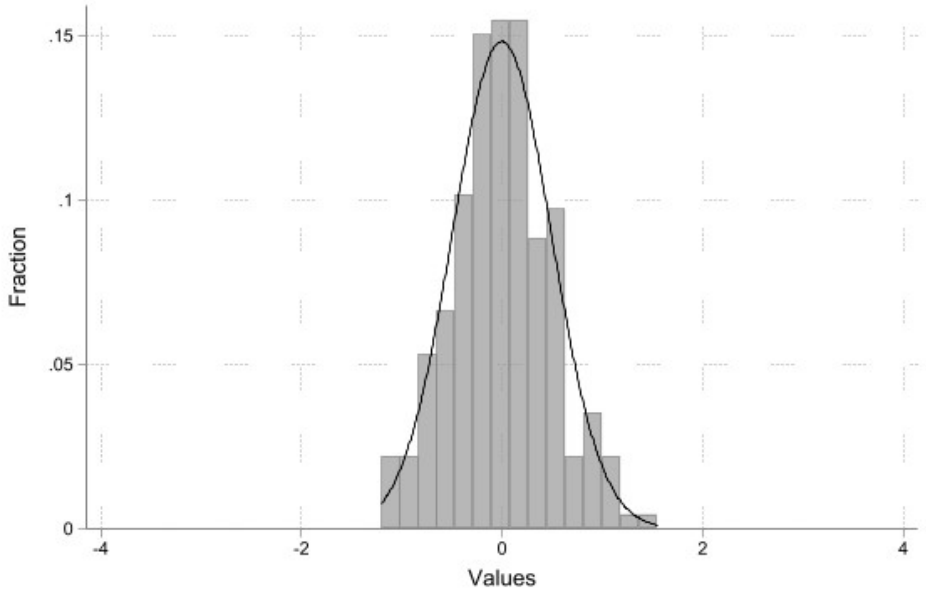
Deviation from Test Results	Share Assigned Rating	Share Algorithmic Rating
<-3.5	0	0
[-3.5; -2.5]	0	0
[-2.5; -1.5]	1.32	0
[-1.5; -0.5]	25.11	15.49
[-0.5; 0.5]	54.63	69.47
[0.5; 1.5]	14.54	14.6
[1.5; 2.5]	2.2	0.44
[2.5; 3.5]	1.32	0
>3.5	0.88	0

FIGURE B1 Accuracy test results – assigned ratings for adaptability



Note. This figure depicts the deviation of the assigned ratings from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.

FIGURE B2 Accuracy test results – algorithmic ratings for adaptability



Note. This figure depicts the deviation of the algorithmic ratings from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.

Appendix C: Cooperativeness

TABLE C1 Correlation matrix for cooperativeness

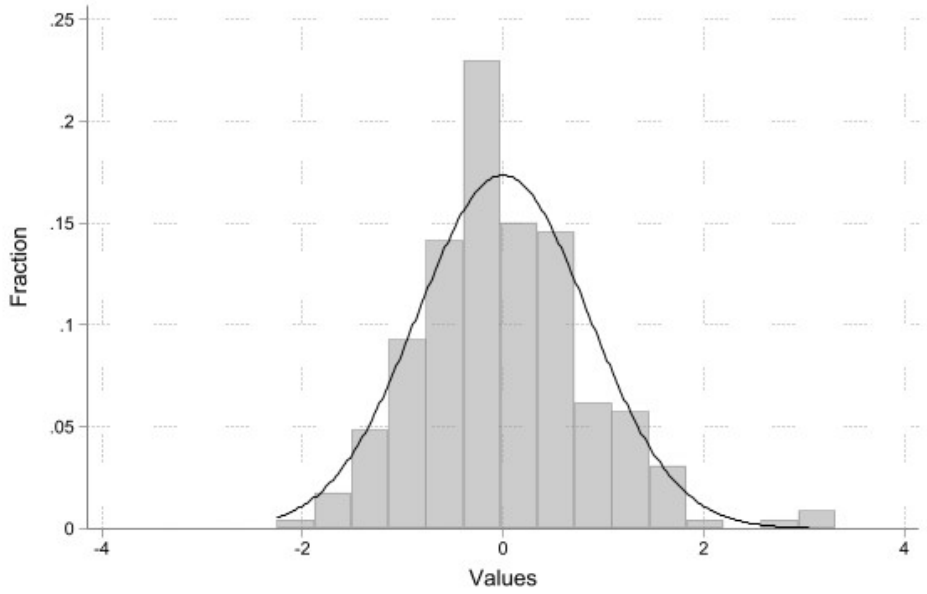
	1.	2.	3.
1. Test results cooperativeness			
2. Assigned ratings cooperativeness	.05		
3. Algorithmic ratings cooperativeness	-.06	.31***	

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

TABLE C2 Share of observations deviating from the test results for cooperation

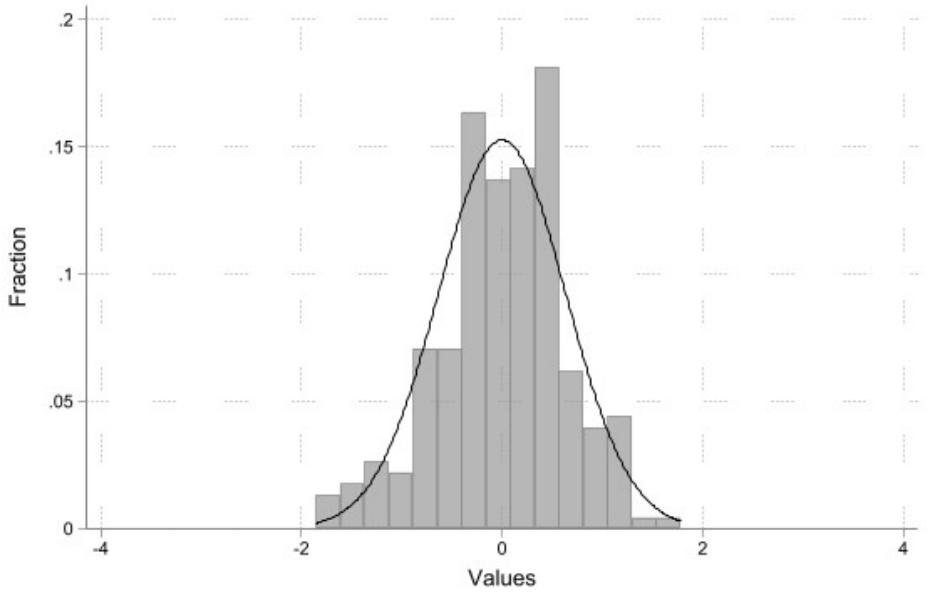
Deviation from Test Results	Share Assigned Rating	Share Algorithmic Rating
<-3.5	0	0
[-3.5; -2.5]	0	0
[-2.5; -1.5]	2.21	2.21
[-1.5; -0.5]	24.34	17.7
[-0.5; 0.5]	48.67	61.95
[0.5; 1.5]	20.8	17.7
[1.5; 2.5]	2.65	0.44
[2.5; 3.5]	1.33	0
>3.5	0	0

FIGURE C1 Accuracy test results – assigned ratings for cooperativeness



Note. This figure depicts the deviation of the assigned ratings from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.

FIGURE C2 Accuracy test results – algorithmic ratings for cooperativeness



Note. This figure depicts the deviation of the algorithmic ratings from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.

Appendix D: Trust

TABLE D1 Correlation matrix for trust

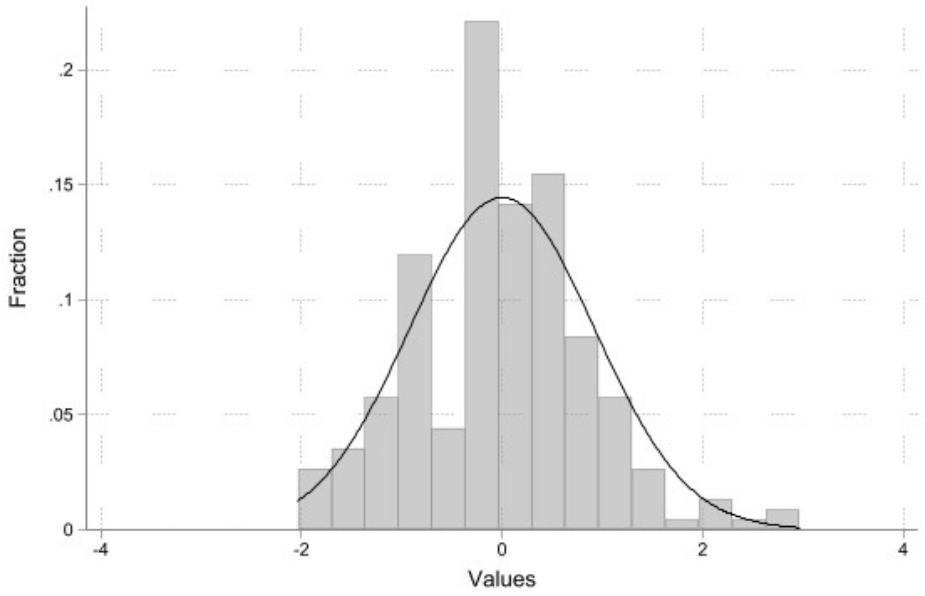
	1.	2.	3.
1. Test results trust			
2. Assigned ratings trust	.05		
3. Algorithmic ratings trust	-.07	-.09	

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

TABLE D2 Share of observations deviating from the test results for trust

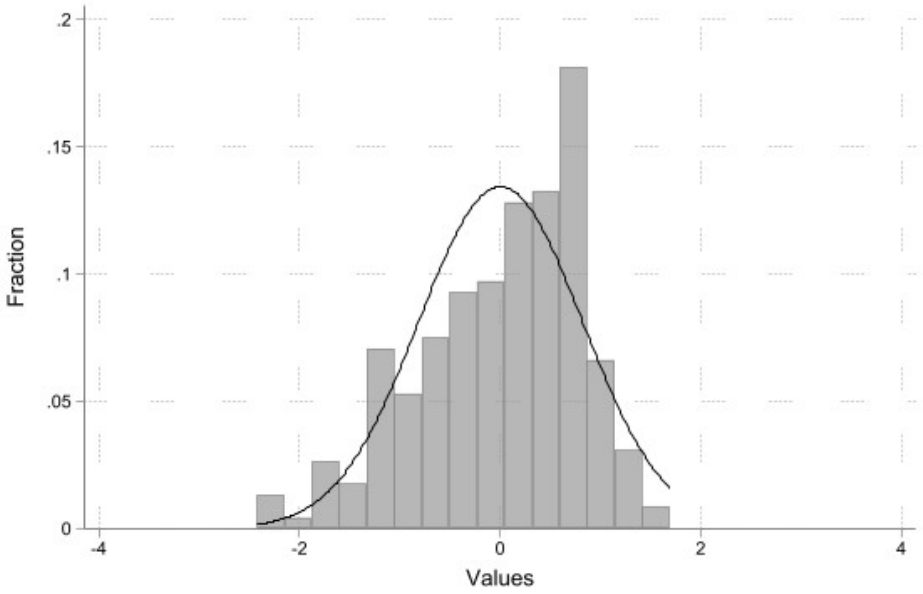
Deviation from Test Results	Share Assigned Rating	Share Algorithmic Rating
<-3.5	0	0
[-3.5; -2.5]	0	0
[-2.5; -1.5]	5.31	4.42
[-1.5; -0.5]	21.24	21.68
[-0.5; 0.5]	44.69	39.82
[0.5; 1.5]	23.89	33.19
[1.5; 2.5]	3.54	0.88
[2.5; 3.5]	1.33	0
>3.5	0	0

FIGURE D1 Accuracy test results – assigned ratings for trust



Note. This figure depicts the deviation of the assigned ratings from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.

FIGURE D2 Accuracy test results – algorithmic ratings for trust



Note. This figure depicts the deviation of the algorithmic ratings from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.

Appendix E: Empathy

TABLE E1 Correlation matrix for empathy

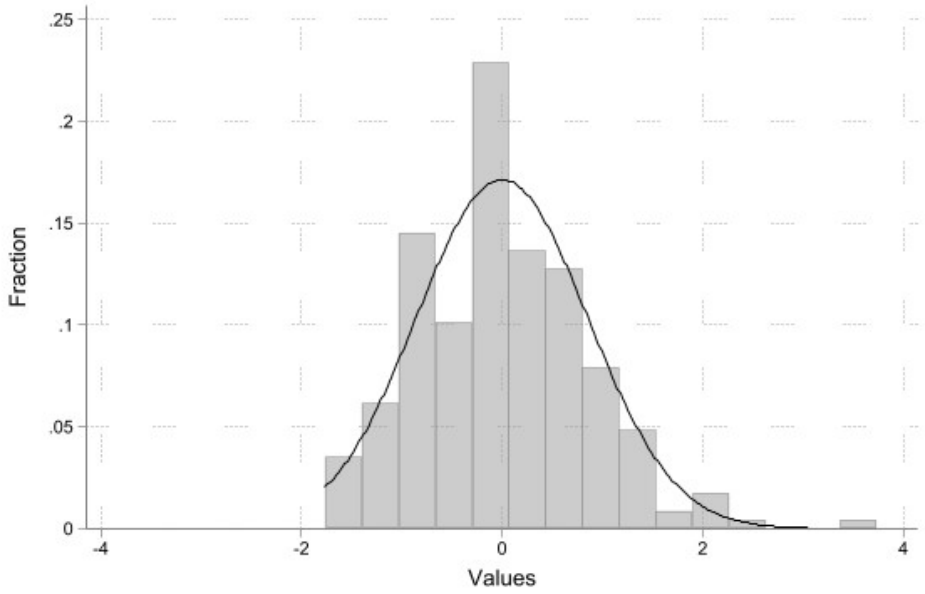
	1.	2.	3.
1. Test results empathy			
2. Assigned ratings empathy	.07		
3. Algorithmic ratings empathy	.003	.39***	

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

TABLE E2 Share of observations deviating from the test results for empathy

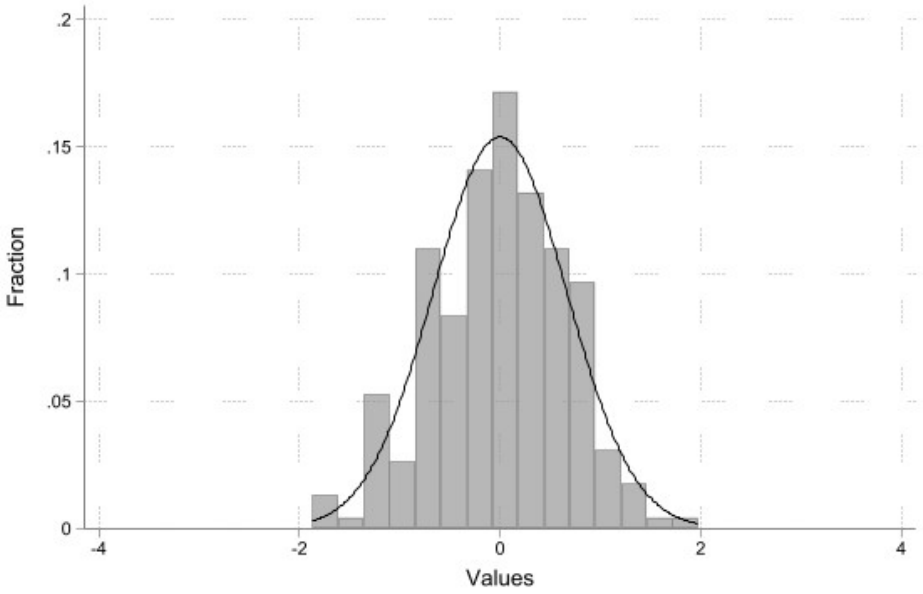
Deviation from Test Results	Share Assigned Rating	Share Algorithmic Rating
<-3.5	0	0
[-3.5; -2.5]	0	0
[-2.5; -1.5]	3.52	1.76
[-1.5; -0.5]	27.75	22.03
[-0.5; 0.5]	46.26	52.42
[0.5; 1.5]	18.94	22.91
[1.5; 2.5]	3.08	0.88
[2.5; 3.5]	0	0
>3.5	0.44	0

FIGURE E1 Accuracy test results – assigned ratings for empathy



Note. This figure depicts the deviation of the assigned ratings from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.

FIGURE E2 Accuracy test results – algorithmic ratings for empathy



Note. This figure depicts the deviation of the algorithmic ratings from the test results. The deviation has been calculated by subtracting the inverted mean-centered ratings from the mean-centered test results.