Journal of Business Economics (2021) 91:35–59
https://doi.org/10.1007/s11573-020-00976-0

ORIGINAL PAPER

Likeability in subjective performance evaluations: does it bias managers' weighting of performance measures?

Kai A. Bauch[1] · **Peter Kotzian**[2] · **Barbara E. Weißenberger**[2]

Published online: 19 March 2020

Abstract
In this paper, we investigate how subordinate likeability induces bias in managers' subjective performance evaluations. Based on the affect-consistency heuristic, we expect managers who use multiple performance measures to subjectively evaluate their subordinates' performance to place greater weight on likeability-consistent performance measures than on likeability-inconsistent measures. Hence, we predict that likeability and performance information interact in affecting managers' performance evaluations. The results of our experiment support this prediction. In line with prior research, we find evidence of likeability bias in subjective performance evaluations: likeable subordinates receive more favorable evaluations than dislikeable ones. We further find that participants adjust their performance evaluations in the presence of likeability-consistent performance information to a greater extent than in the presence of likeability-inconsistent performance information. Thus, in accordance with the affect-consistency heuristic, our results indicate that likeability bias occurs due to a differential, biased weighting of performance measures. Additionally, we find that perceived likeability is also affected by subordinates' performance, which in turn partially mediates the effect of subordinate performance on evaluations: good performers are more likeable than poor performers. Hence, this can exacerbate likeability bias. We discuss the implications of our findings for the design of performance evaluation systems in practice.

Keywords Likeability bias · Affect · Subjective performance evaluation

JEL Classification M12 · M41 · M52

✉ Kai A. Bauch
kai.bauch@iuc.unibe.ch

1 University of Bern, Engehaldenstr. 4, 3012 Bern, Switzerland

2 Heinrich Heine University Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany

1 Introduction

As an integral part of many organizations' control systems, managers are frequently tasked to subjectively evaluate their subordinates' performance. However, a growing number of studies reports on *bias* in such performance evaluations (e.g., Fehrenbacher et al. 2018; Kaplan et al. 2017; Bol 2011; Maas and Torres-González 2011; Moers 2005; Lipe and Salterio 2000). Biased evaluation behavior is detrimental for organizations, especially when managers evaluate similar performances differently, as it can impair the perceived fairness of evaluation systems (Voußem et al. 2016). Furthermore, biased evaluations can affect personnel decisions (e.g., Moers 2005) and lead to inappropriate feedback, thus impairing the overall effectiveness of organizations' control systems. Hence, it is important to obtain a sound understanding of the various determinants and mechanisms of bias in subjective performance evaluations.

Since both the sender (i.e., the manager) and the receiver (i.e., the subordinate) of subjective performance evaluations are humans, research on the antecedents of this behavior indicates that the manager–subordinate relationship can cause such biased performance evaluations (Bol 2011). While different facets of this relationship may cause bias, prior research suggests that especially subordinate *likeability* is crucial (e.g., Robbins and DeNisi 1998). Likeability is an interpersonal affective reaction between individuals (e.g., Ding and Beaulieu 2011; Antonioni and Park 2001) and can be positive as well as negative. For example, similarities (or dissimilarities) in demographics or work style can determine likeability (Carmona et al. 2014; Xu and Tuttle 2005). Positive or negative likeability is likely to develop quickly in every manager–subordinate relationship but should not be relevant to performance evaluations (Kaplan et al. 2007). However, a solid body of evidence indicates that this claim does not hold (e.g., Carmona et al. 2014; Xu and Tuttle 2005; Antonioni and Park 2001). While this stream of literature has mainly focused on likeability as a determinant of bias in performance evaluations (i.e., the observation of a positive association between likeability and bias in performance evaluations), the relevant underlying mechanism (i.e., how likeability causes bias in performance evaluations) is less well explored.

However, research in psychology has already approached performance evaluation from a cognitive processing perspective, suggesting possible mechanisms underlying bias in evaluations during the distinct cognitive stages of information acquisition, encoding, retrieval, and weighting (e.g., Robbins and DeNisi 1994; Feldman 1981). But due to substantial differences between business contexts and the more generic setups of most of these studies, it remains an open question whether those findings can be generalized to business contexts. In practice, subjectivity is often induced to performance evaluations by allowing managers to discretionary weight and aggregate *multiple performance measures* to overall performance evaluations.[1] Yet, accounting research on the cognitive processes underlying such subjective performance evaluations is scarce. First steps have been taken by Kaplan et al. (2007), who suggest that likeability may lead to bias in information seeking and weighting of performance measures in performance reports. In this paper, we complement this line of research

[1] As we will outline in more detail in Sect. 2, psychology research usually entails participants directly observing behavior during evaluation tasks, while in the business context, managers often have to rely on multiple performance measures to evaluate subordinates. This may trigger different cognitive processes.

by taking a cognitive processing perspective and addressing the research question of how likeability causes bias in multi-measure-based performance evaluations.

We focus on a dyadic setting, in which a manager evaluates the performance of a subordinate based on her subjective weighting of multiple performance measures.[2] Although often theorized (e.g., Kramer and Maas 2019; Kaplan et al. 2007), to the best of our knowledge, no study has focused directly on manager–subordinate-relationship-driven performance evaluation bias as an outcome of a biased weighting of multiple performance measures yet. In line with the affect-consistency heuristic (Robbins and DeNisi 1994), we posit that managers weight performance measures indicating positive or negative performance differently to yield overall evaluations consistent with their relationship to the subordinate, i.e., managers attach greater weight to likeability-consistent performance measures than to likeability-inconsistent measures. This is reflected in an asymmetric interaction between subordinate likeability and subordinate performance: In the presence of positive subordinate likeability, managers inflate their subjective performance evaluations given positive (negative) performance information to a greater (lesser) extent than in the absence of such likeability. In contrast, in the presence of negative subordinate likeability, managers deflate their subjective performance evaluations given negative (positive) performance information to a greater (lesser) extent than in the absence of such likeability.

To test our predictions, we conduct a 3×3 factorial between-subjects experiment with student participants. Similar to prior research, participants in our experiment assume the role of managers and subjectively evaluate the performance of a subordinate, who is presented to them as likeable, dislikeable, or unknown (control condition). The subordinate's performance is manipulated by altering one of the performance measures within the subordinate's performance report, which indicates negative, neutral, or positive performance. We observe participants' performance evaluations.

Our results support our predictions. In line with prior research, we find evidence of likeability bias: Regardless of the subordinate's performance, positive likeability leads to upward-shifted evaluations, while negative likeability leads to downward-shifted evaluations in comparison to the control condition. Our results indicate that this likeability bias is driven by a differential weighting of likeability-consistent and likeability-inconsistent performance measures. That is, our results also show the predicted interactions. Moreover, additional analyses further reveal that subordinates' performance level affects their perceived likeability: Using data from the post-experimental questionnaire, our results suggest that perceived likeability partially mediates the effect of subordinate performance on managers' subjective performance evaluations.

This study makes three contributions to the accounting literature. First, we add to the growing number of studies on managers' cognitive processes during subjective performance evaluations (e.g., Kramer and Maas 2019; Sohn et al. 2019; Dai et al. 2018; Fehrenbacher et al. 2018; Chen et al. 2016). We complement prior work, which has mainly examined the cognitive stage of information acquisition, by focusing on information weighting. Though a biased weighting of multiple, differently valenced

[2] Other, non-dyadic settings include calibration committees (e.g., Demeré et al. 2018), where multiple managers are involved in the evaluation of a single employee, or team settings (e.g., Arnold and Tafkov 2019) in which a manager evaluates multiple employees simultaneously and allocates bonuses.

(i.e., positive and negative) performance measures has been proposed before, empirical evidence has been absent yet. We show that likeability affects managers' weighting of likeability-consistent and likeability-inconsistent performance measures.

Second, we thereby also add to accounting research on affect in general. Extant literature has explored the influence of affect in a variety of settings, including capital budgeting decisions (e.g., Fehrenbacher et al. 2019; Farrell et al. 2014; Kida et al. 2001), risky choices (Moreno et al. 2002), whistleblowing (e.g., Robertson et al. 2011), and investor decision-making (Elliott et al. 2014). Our results suggest that in decision-making based on multiple information cues, affect may influence the weight decision-makers place upon differently valenced information. For example, in an investment scenario, affect towards a CEO may drive investors' weighting of ambiguous financial statement information (cf. Kaplan et al. 2018).

Finally, research has also been concerned with bias in performance evaluations due to subordinates' prior performance (e.g., Woods 2012; Reilly et al. 1998). While this body of literature has long been distinct from the literature on likeability in performance evaluations (Robbins and DeNisi 1994), more recent research has investigated the interplay between these two topics (e.g., Varma and Pichler 2007). We extend this line of research by showing that subordinates' *current* performance also affects their perceived likeability and may thus exacerbate bias in performance evaluations. Intuition suggests that while subordinate likeability and prior level of performance might cause bias in performance evaluations, subordinates' current level of performance should be directly linked to managers' performance evaluations. However, our results suggest that this relationship is more complex: current subordinate performance may increase (or decrease) their perceived likeability and thus unduly affect performance evaluations. Hence, from a practical perspective, our results should alert organizations to the possible consequences of managers' susceptibility to affective influences during performance evaluations. However, since it is likely not possible to prevent likeability from arising in manager–subordinate relationships, organizations that want to maintain such performance evaluation systems might consider inducing further elements of control. For example, providing feedback and reminding managers of the crucial role of fairness in performance evaluations seems to be a promising path (Kang and Fredin 2012; Maas et al. 2012).

We have organized the remainder of this paper as follows: In the next section, we review the relevant literature and develop the hypotheses. In the third section, we describe the research design, while in the fourth section, we present our results. In the final section of the paper, we conclude with a discussion of our results.

2 Literature review and hypothesis development

2.1 Subjective performance evaluation using accounting information

From a practical perspective, performance evaluation is implemented in many organizations by one person subjectively assigning an overall evaluation (expressed as one key indicator) to another person, based on a selection of measures listed in a performance report (e.g., Maas and Verdoorn 2017). Thus, one person discretionar-

ily interprets and weights several performance measures and subjectively determines an evaluation. While most organizations use such subjective performance evaluation systems to increase the perceived fairness of evaluations, subjectivity also makes performance evaluations susceptible to cognitive biases (e.g., Kramer and Maas 2019; Cardinaels and van Veen-Dirks 2010; Moers 2005; Lipe and Salterio 2000). For example, research has found that managers tend to place little weight on unique (as opposed to common) performance measures (Lipe and Salterio 2000) or non-financial measures (Cardinaels and van Veen-Dirks 2010) when conducting performance evaluations.

Such managerial judgments do not happen in temporal and personal isolation but within an organizational context. Hence, such behavior might be triggered by elements of the interpersonal relationship between managers and subordinates (e.g., Bol 2011; Xu and Tuttle 2005). Yet, accounting researchers investigating multi-dimensional performance evaluation systems have only recently begun to address the role of interpersonal relationships in subjective performance evaluations (e.g., Carmona et al. 2014; Kaplan et al. 2007).

More research is needed on this topic since the unique features of the accounting environment usually preclude generalizing findings from psychology to business contexts (Haynes and Kachelmeier 1998). Psychology research on performance evaluations typically examines scenarios in which a decision-maker directly observes behavior (often operationalized through the use of videotapes) or is provided with behavioral incidents without target values and then forms an evaluation (e.g., Robbins and DeNisi 1998; DeNisi et al. 1997; Foti and Hauenstein 1993). However, in most business contexts, managers do not have an opportunity to directly monitor subordinates' performance.[3] Instead, despite frequently interacting personally with subordinates, managers usually have to base their evaluations on a combination of various performance measures (and their corresponding target and actual values) presented in a report. This is a different task, especially given the underlying cognitive mechanisms and the possibilities for biases to arise.[4]

2.2 Interpersonal relationships and bias in performance evaluations

Still, the psychology literature does provide evidence of the possible impact of various aspects of the manager–subordinate relationship on performance evaluations. A large body of research shows that particularly the manager–subordinate likeability may cause bias in performance evaluations (e.g., Tsui and Barry 1986; Cardy and Dobbins 1986). Research by Robbins and DeNisi (1994, 1998) suggests that likeability triggers an *affect-consistency heuristic*, according to which likeable subordinates receive favorable evaluations, whereas dislikeable subordinates receive unfavorable evaluations. A related stream of psychology research has shown similar biases in performance evaluations resulting from subordinates' prior performance (e.g., Reilly

[3] This is one of the initial reasons why performance evaluation as an element of management control is warranted.

[4] For example, in such a case, it is not necessary to first encode performance as 'good' or 'bad' since a comparison of target and actual values directly classifies performance. In this regard, the literature also suggests that likeability should have less influence in the presence of clear performance targets (Kaplan et al. 2007; Baltes and Parker 2000).

et al. 1998; Steiner and Rain 1989; Balzer 1986). This line of research typically refers to an assimilation effect, when a subordinate who has performed well (poor) in the past receives a more favorable (unfavorable) evaluation for an average performance, than a subordinate who has performed poorly (well) in the past.

On the one hand, given the apparent similarity between assimilation effects (evaluation bias due to prior performance) and the affect-consistency heuristic (evaluation bias due to likeability), it is puzzling that the two streams of research have been treated as distinct by psychology research (Robbins and DeNisi 1994). On the other hand, it is necessary, from a theoretical perspective, to distinguish between these two phenomena. To some extent, prior performance may be of informational value for the evaluation of current performance. For example, if a subordinate who has consistently performed well in the past suddenly performs poorly, there might be other factors at stake than in the case of a constantly underperforming subordinate. Likeability, by contrast, is irrelevant for the process of performance evaluation in all cases (Antonioni and Park 2001).[5] However, Robbins and DeNisi (1994) note that in real organizational settings, there may be a high correlation between prior performance and likeability (e.g., a subordinate who has performed well in the past may be perceived as more likeable). In this vein, two strands of literature have emerged: According to the "rater bias perspective" (e.g., Cardy and Dobbins 1986; Feldman 1981), likeability affects performance evaluations "through either biased information processing and/or intentional distortion" (Sutton et al. 2013, p. 411). On the contrary, studies that take a "true performance interpretation" (e.g., Lefkowitz 2000; Varma et al. 1996) argue that subordinates' "true" performance affects both their performance evaluations (directly) and their likeability or claim that at least third variables (e.g., intelligence or diligence) are associated with both likeability and "true" performance (Sutton et al. 2013).

2.3 Likeability bias in subjective performance evaluations based on accounting information

We are aware of only a small number of studies in the management accounting literature that are closely related to this topic. These studies investigate settings where managers have to form overall evaluations of subordinate performance based on performance reports containing multiple performance measures and their results suggest that managers' evaluations display similar biases to those detected in the psychology literature.

Using the theoretical background provided by Robbins and DeNisi (1994), Kaplan et al. (2007) study the effects of subordinate likeability in a managerial performance evaluation setting, where information is presented either in an unstructured manner or organized in a Balanced Scorecard. They propose that the Balanced Scorecard presentation format mitigates the influence of likeability on performance evaluations but find that participants provide biased evaluations regardless of the presentation format. More recently, Carmona et al. (2014) confirm these findings. Besides providing further

[5] There are numerous possible examples of factors that might cause managers to perceive subordinates as likeable that are irrelevant to performance evaluations, such as when a manager and a subordinate favor the same sports club or their children attend the same school.

evidence to show that likeability leads to biased evaluations in business contexts, they also used two samples from different countries and found regional differences. Furthermore, Kramer and Maas (2019) find that managers who recommended a subordinate for promotion escalate their commitment and provide more favorable evaluations than managers who advised against a subordinate's promotion. Finally, Xu and Tuttle's (2005) results indicate that manager–subordinate similarity fosters a manager's perception of a subordinate as likeable, which, in turn, affects managers' performance evaluations based on accounting information.

The present paper adds to this stream of research by examining the effects of likeability in settings where a manager is presented with a performance report consisting of multiple performance measures and is asked to subjectively evaluate a subordinate's performance based on this information. We do not aim to examine the effects of prior performance. However, we vary the information on *current* performance as we are interested in how managers' use of positive and negative performance information is affected by likeability. First, in line with prior research, we expect likeability to affect evaluations *regardless* of the favorability of the presented performance information. That is, we expect that likeable subordinates receive favorable ratings, while dislikeable subordinates receive unfavorable ratings. Hence, we first state the following two baseline hypotheses:

H1a Managers' subjective performance evaluations of likeable subordinates will be inflated.

H1b Managers' subjective performance evaluations of dislikeable subordinates will be deflated.

While serving as a necessary baseline, we also provide these hypotheses to replicate and validate the findings of prior research in a different setting, an issue that is of vital importance (e.g., Shields 2015). Scholars have only recently stressed the necessity of replication to demonstrate the reliability and validity of research findings outside the context in which the findings were established, which applies especially to experimental research (Kaplan et al. 2018).[6]

2.4 The biased weighting mechanism underlying likeability bias

Next, we focus on the cognitive mechanism underlying likeability bias and, thus, on how likeability bias differs in the presence of diverging performance information. Hence, while various explanations have been proposed of why managers' evaluations might be biased,[7] our argument is based on the mechanism that underlies the specific setting of forming overall evaluations based on multiple performance measures.

[6] Salterio (2014) also stresses the importance of replication in accounting research. In particular, he outlines that the paper he co-authored with Marlys Lipe on the common measure bias (Lipe and Salterio 2000), which, like the present study, deals with managers' weighting of performance measures in subjective performance evaluation, has been replicated at least 18 times. Prominent examples which have been published in major accounting journals include (but are not limited to) Banker et al. (2004), Dilla and Steinbart (2005), and Libby et al. (2004).

[7] For example, some authors suggest that managers generally provide inflated ratings to avoid confrontations (e.g., Bol et al. 2016).

Referring to Robbins and DeNisi (1994), Kaplan et al. (2007) argue that the affect-consistency heuristic operates in the stages of information acquisition and information weighting. That is, managers cognitively process performance information in a biased fashion.

In a recent study, Kramer and Maas (2019) experimentally address biased attention as a cause of escalation bias in subjective performance evaluation. They suggest that biased performance evaluations can be explained by selective exposure, which builds on the theory of cognitive dissonance (Festinger 1957). Kramer and Maas (2019) assume that managers who made positive or negative recommendations regarding the promotion of an employee pay different degrees of visual attention to favorable and unfavorable information presented in an ambiguous Balanced Scorecard. However, contrary to their expectations, their eye-tracking data does not support the prediction. Hence, this seminal evidence implies that biased attention to performance measures is likely not the mechanism underlying biases in performance evaluation due to interpersonal relationships.[8]

However, besides a possible mechanism located in information acquisition, Kaplan et al. (2007) argue that the mechanism in information weighting leading to affect-consistency may be similar to motivated reasoning. According to Kunda (1990), motivated reasoning is the behavior of evaluating and weighting information in a way that fits one's own preferences. This means that when managers are confronted with various divergent pieces of information, they will discount information that is inconsistent with their preferences and over-emphasize information that supports their preferences. Applied to the domain of subjective performance evaluation, Kaplan et al. (2007) suggest that managers are motivated to inflate their evaluations of likeable subordinates and deflate their evaluations of dislikeable subordinates. Thus, when managers evaluate subordinate performance based on multiple measures, they tend to place more weight on likeability-consistent performance measures than on likeability-inconsistent ones. In the case of a likeable subordinate, positive information will thus be over-weighted and negative information under-weighted, whereas in the case of a dislikeable subordinate, positive information will be underweighted and negative information taken into account more severely (Kaplan et al. 2007; Robbins and DeNisi 1994).

In this paper, we set out to directly test this mechanism. We argue that this mechanism is reflected in a specific interaction between subordinates' likeability and subordinates' performance information. In particular, we expect managers to place greater (lesser) weight on positive (negative) performance information in the presence of positive subordinate likeability than in the absence of such likeability.[9] Therefore, we predict that in the presence of positive subordinate likeability, managers will inflate their subjective performance evaluations given positive (negative) performance information to a greater (lesser) extent than in the absence of such likeability. By contrast,

[8] Regarding likeability, Robbins and DeNisi's (1994) results also imply that affect-consistency is not associated with information acquisition.

[9] While not the focus of our paper, we note that the consciousness of this behavior is ambiguous. For example, Luft and Shields (2009) elaborate on motivated reasoning and outline that it affects individuals' cognitive processes "…in ways of which individuals are not fully conscious." (p. 234). We revisit this issue in our supplementary analyses.

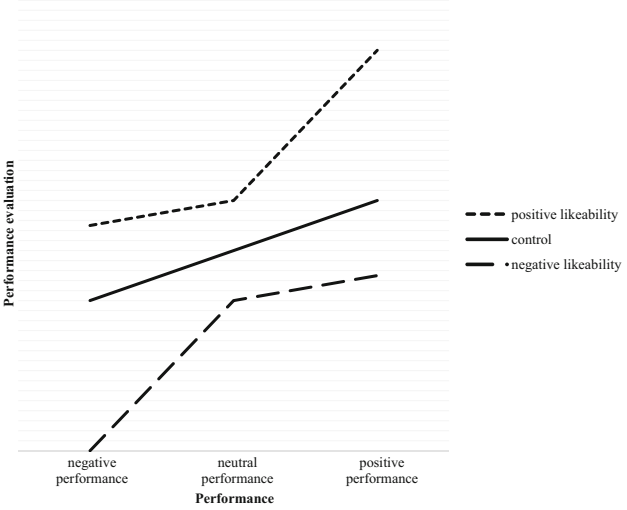

Fig. 1 Predicted interaction effect between likeability and performance. This figure presents the predicted interaction effects for the experiment, where the dependent variable is performance evaluation. The experiment manipulates likeability (negative likeability, control, positive likeability) and performance (negative performance, neutral performance, positive performance) between subjects

we expect managers to place greater (lesser) weight on negative (positive) performance in the presence of negative subordinate likeability compared to the absence of such likeability. Thus, we conversely predict that in the presence of negative subordinate likeability, managers will deflate their subjective performance evaluations given negative (positive) performance information to a greater (lesser) extent than in the absence of such likeability. This leads to the following formal hypotheses:

H2a In the presence of positive subordinate likeability, managers will inflate their subjective performance evaluations given positive (negative) instead of neutral performance information to a greater (lesser) extent than in the absence of such likeability.

H2b In the presence of negative subordinate likeability, managers will deflate their subjective performance evaluations given negative (positive) instead of neutral performance information to a greater (lesser) extent than in the absence of such likeability.

Figure 1 illustrates the predicted pattern of means. The expected biased weighting of performance information is contrasted with the weighting of performance information for an unknown subordinate (control group), where likeability is absent and managers are assumed to weight positive and performance deviations similar to equal-in-magnitude negative performance deviations (i.e., in a linear manner).

3 Research method

3.1 Experimental design and task overview

We conduct an experiment with a 3×3 full-factorial design to test our predictions.[10] We manipulate subordinate likeability and subordinate performance between subjects. Similar to Carmona et al. (2014) and Kaplan et al. (2007), we provide a description of a hypothetical subordinate to manipulate likeability. However, while Kaplan et al. (2007) base their design on the seminal work of Lipe and Salterio (2000), asking participants to evaluate a likeable and a dislikeable subordinate simultaneously,[11] we manipulate likeability between subjects. Hence, each participant is only asked to evaluate one subordinate. In the positive likeability condition, our case description outlines a likeable subordinate, while in the negative likeability condition, the subordinate is described as dislikeable. Furthermore, in order to test the proposed interactions, we add a control condition as a baseline, whereby the hypothetical subordinate is not described as either likeable or dislikeable.[12] To manipulate performance, we vary the performance report for the hypothetical subordinate to indicate a negative, neutral, or positive overall performance by altering one performance measure between subjects, while keeping the remaining performance measures constant. We observe participants' subjective performance evaluations. There was no deception of any kind.

Based on prior research (e.g., Kramer and Maas 2019; Kaplan et al. 2007; Lipe and Salterio 2000), we ask participants to assume the role of a regional manager in a retail clothing company called TrueDenim Clothing Group (TDC). Participants are told that it is part of the job to evaluate subordinate-managers' performance and that the management accounting department of TDC provides relevant information on four performance measures relevant for this task.[13] Furthermore, they are informed about their discretion in weighting these measures. After studying the performance report containing the target and actual values for these four measures for a subordinate named Michael Schmitz,[14] participants are asked to provide their evaluation of him.

[10] The research design initially featured two positive likeability treatments. By intention, they should affect performance evaluations differently. However, the second treatment did not significantly differ in its effect on performance evaluation. Furthermore, regarding the likeability manipulation check, both treatments yielded inferentially identical results. In order to retain a balanced sample, we refrained from pooling those treatment conditions but omitted this second likeability condition.

[11] Likewise, Carmona et al. (2014) presented two subordinates (one likeable, one dislikeable) simultaneously to each participant.

[12] This design choice follows related psychology research, which emphasizes the necessity of a control condition in settings such as ours (Kravitz and Balzer 1992).

[13] As we acknowledge that experiments should not strive for unnecessary mundane realism, we refrained from implementing real-world performance measures (e.g., customer satisfaction) but instead labeled the measures A, B, C, and D, respectively, to avoid that participants' weighting of favorable and unfavorable performance information is confounded with their perceived importance of various performance measures (Kadous and Zhou 2018).

[14] "Michael" and "Schmitz" are among the most common German first and family names, respectively. We, therefore, expect that any positive or negative connotations would be non-systematic and, due to experimental randomization, would not affect our results.

3.2 Participants and procedures

In total, 267 undergraduate students participated in the experiment.[15] Due to missing answers to variables on which the analyses are based, 26 participants had to be excluded. Hence, the net sample consists of 241 participants.[16] Participants are on average slightly below 21 years old and 49.6% are females. A Kruskal–Wallis test indicates that there are no significantly different means between the experimental conditions for age ($p = 0.63$, two-tailed). Likewise, a Chi square test indicates no differences across conditions with regard to gender ($p = 0.32$, two-tailed). We, thus, conclude that experimental randomization was successful.

The experiment was conducted during a cost accounting lecture at a large German university.[17] Students were told that participation was voluntary, and participating students were advised to work separately and to follow the predefined order. There were no financial incentives.[18] Instruments were distributed randomly. After completing the experimental task, participants were asked to complete a post-experimental questionnaire, which included a manipulation check, several questions regarding the experimental task and demographics. Participants were debriefed directly after participation. In total, procedures took approximately 20 min.

3.3 Variables

3.3.1 Dependent variables

After studying the performance report for Michael Schmitz, participants are asked to provide an overall evaluation on an 11-point Likert scale ranging from 0 to 10. Participants' evaluations on this scale represent our main dependent variable labeled *performance evaluation*. Furthermore, we asked participants to indicate their level of agreement with the statement "I liked store manager Michael Schmitz" on a scale

[15] A stream of methodologically oriented studies addresses the topic of using students in accounting-related judgment and decision-making experiments (e.g., Elliott et al. 2007; Libby et al. 2002; Ashton and Kramer 1980). The results obtained by Elliott et al. (2007) suggest that as long as the cognitive complexity of the task does not exceed the capabilities of the students, the results can be transferred to real-world decision-makers. Libby et al. (2002) even conclude that researchers should refrain from using professionals unless necessary. Schwering (2017) argues that students should not be used as surrogates for managers if managers' experience is important to the task but that in tasks that do not require such experience, real managers' reliance on experience may indeed be a confounding factor. As our task does not necessarily require expertise and students' cognitive processes are assumed not to differ from practitioners' cognitive processes in the experimental task, using a student sample is deemed suitable for answering our research question.

[16] All analyses have been replicated using the full sample where possible; effects stay inferentially identical.

[17] We conducted the experiment paper-based and the original language of the materials was German.

[18] In practice, firms are usually unable to incentivize managers to provide accurate performance evaluations as this would imply the possibility of determining objectively what constitutes an accurate performance evaluation (Ding and Beaulieu 2011). However, subjective performance evaluations are especially well-suited mechanisms in cases where such objective performance evaluations are not determinable.

from 1 (does not apply at all) to 7 (fully applies). This variable is labeled *perceived likeability* and is subject to additional analyses.

3.3.2 Independent variables

Our first independent variable is *likeability*. We manipulate likeability by introducing Michael Schmitz in the case description prior to the evaluation task. In the positive likeability condition, Michael Schmitz is introduced as a sane, considerate, unassuming, and unobtrusive employee whereas in the negative likeability condition, he is described as boastful, gossipy, self-centered, and superficial. In the control condition, participants proceed to the performance evaluation directly after the instructions.

Our second independent variable is *performance*, which is manipulated at three levels by providing different types of performance reports: a negative performance condition (one performance measure is diagnostic, indicating poor performance; the remaining three measures are average), a neutral performance condition (all performance measures are average) and a positive performance condition (one performance measure is diagnostic, indicating good performance; the remaining three measures are average).[19] Hence, we keep three out of four performance measures constant across conditions. Any deviation in the evaluation (within a specific likeability condition) must, therefore, reflect a different weighting of the diagnostic performance measure.

4 Results

4.1 Preliminary analyses and descriptive statistics

Using perceived likeability as a manipulation check, a one-way ANOVA and follow-up contrasts (untabulated) reveal that participants in the positive likeability condition perceived the subordinate as more likeable ($t = 4.29$, $p < 0.01$, one-tailed) than participants in the control condition. Participants in the negative likeability conditions show significantly lower values ($t = -6.53$, $p < 0.01$, one-tailed).

Table 1 reports the means and standard deviations for the performance evaluations by experimental condition. The descriptive statistics are in line with our hypotheses: On average, participants in the positive likeability condition (mean = 5.25, sd = 1.67) provided inflated evaluations in comparison to the control group (mean = 4.72, sd = 1.82), while participants in the negative likeability condition (mean = 4.09, sd = 1.75) provided deflated evaluations. Overall, evaluations in the presence of negative performance information (mean = 4.08, sd = 1.84) were lower than those in the presence of neutral performance information (mean = 4.83, sd = 1.71) or positive performance information (mean = 5.39, sd = 1.61), indicating that participants considered differences in performance information in the expected manner.

[19] The deviation from the neutral condition within the diagnostic performance measure was equal-in-magnitude for the positive and negative performance conditions.

Table 1 Descriptive statistics for performance evaluation across conditions

Likeability[b]	Performance[a]			
	Negative	Neutral	Positive	Total
Negative	3.67 (1.37) n = 24	4.28 (1.93) n = 25	4.33 (1.91) n = 21	4.09 (1.75) n = 70
Control	3.64 (1.73) n = 25	4.96 (1.66) n = 23	5.52 (1.58) n = 27	4.72 (1.82) n = 75
Positive	4.69 (2.05) n = 35	5.16 (1.51) n = 32	6.03 (0.91) n = 29	5.25 (1.67) n = 96
Total	4.08 (1.84) n = 84	4.83 (1.71) n = 80	5.39 (1.61) n = 77	4.75 (1.80) n = 241

Performance evaluation is our dependent variable and is measured on an 11-point Likert-scale. It ranges from 0 to 10. The standard deviations are in brackets

[a]*Performance* is manipulated between-subjects on three levels. In the negative conditions, the presented performance report contains one poor performance measure, while in the neutral conditions, all performance measures are average. Participants in the positive performance conditions receive a report containing one good performance measure

[b]*Likeability* is manipulated between-subjects on three levels: in the negative likeability condition, a dislikeable profile is presented. The control conditions receive no profile for Michael Schmitz. In the positive likeability condition, participants receive a likeable profile

4.2 Hypotheses testing

4.2.1 Tests of hypotheses 1a and 1b

In H1a and H1b, we predict that positive likeability will lead to inflated evaluations while negative likeability will cause deflated evaluations. Table 2 shows the results of an ANOVA with the two manipulations as factors. In line with H1a and H1b, we find a significant main effect for likeability ($F = 10.56$, $p<0.01$, two-tailed). Follow-up contrasts (untabulated) comparing the treatments with the control condition reveal that positive likeability led to significantly inflated performance evaluations ($t = 2.29$, $p = 0.01$, one-tailed) while negative likeability led to deflated evaluations ($t = -2.21$, $p = 0.01$, one-tailed). Thus, the results support H1a and H1b. While we also find a significant main effect for performance ($F = 12.29$, $p<0.01$, two-tailed) in the ANOVA, the interaction term between likeability and performance is insignificant ($F = 1.12$, $p = 0.35$, two-tailed).

4.2.2 Tests of hypothesis 2a

However, for analyzing the interaction, we do not rely on ANOVA, as conventional ANOVA is not best suited to detect ordinal interactions, especially when variables are varied on more than two levels (Ravenscroft and Buckless 2018). Since we vary both our independent variables on more than two levels, and predict an asymmetric pattern of cell means a priori, custom contrast coding is used to provide an appropriate test of H2a and H2b, while also maximizing statistical power (Guggenmos et al. 2018; Buckless and Ravenscroft 1990). With H2a, we predict that in the presence of positive

Table 2 Test of H1 ANOVA with treatments as between factors

Source of variation	SS	df	MS	F-stat	p value[a]
Likeability	58.19	2	29.10	10.56	< 0.01
Performance	67.73	2	33.87	12.29	< 0.01
Likeability × performance	12.33	4	3.08	1.12	0.35
Error	639.22	232	2.76		

Dependent variable: performance evaluation (n = 241), $R^2 = 0.18$
[a]All reported p values are two-tailed

subordinate likeability, managers will inflate their subjective performance evaluations given positive (negative) performance information to a greater (lesser) extent than in the absence of such likeability.

To test H2a, we compare participants in the positive likeability condition to those in the control condition. For the first part of the hypothesis test, we restrict the analysis to positive and neutral performance information. We employ the following contrast weights: − 2 for the neutral performance/control condition, − 1 for the neutral performance/positive likeability condition, − 1 for the positive performance/control condition, and + 4 for the positive performance/positive likeability condition.[20] The results are presented in panel A of Table 3. The hypothesized pattern is supported (F = 7.76, p < 0.01, two-tailed). We also perform the accompanying semi-omnibus *F* test, which tests for the residual between-groups variance and, therefore, indicates whether the contrast model is a good fit for the data. The test is not significant (F = 0.60, p = 0.55, two-tailed), which indicates that the contrast model explains all between-groups variance in the data. The final step, which is recommended by Guggenmos et al. (2018), is to test the relative contrast variance residual (q^2), which indicates how much of the variance is not explained by the employed set of contrasts. A q^2 of 0.15 indicates that our contrast model explains a reasonable amount of the variance in the data.

In a second step, we compare participants who received negative or neutral performance information. We employ the following contrast weights: − 4 for the negative performance/control condition, + 1 for the neutral performance/control condition, + 1 for the negative performance/positive likeability condition, and + 2 for the neutral performance/positive likeability condition. As panel B of Table 3 shows, the contrast model is again highly significant (F = 11.14, p < 0.01, two-tailed). The residual between-groups variance test is not significant (F = 0.18, p = 0.83, two-tailed), and a

[20] Note that the sets of contrast weights used to test H2a and H2b all test for patterns that represent a combination of a likeability main effect and an ordinal interaction between likeability and performance information as our theory predicts (cf. Guggenmos et al. 2018). For example, in the case of our first contrast test, both the contrast weights for the neutral performance/positive likeability condition (− 1) and those for the positive performance/positive likeability condition (+ 4) are greater than the respective contrast weights for the neutral performance/control condition (− 2) and the positive performance/control condition (− 1), thus representing a main effect of likeability. However, the greater difference in contrast weights within the positive performance condition (+ 4 vs. − 1) than within the neutral performance condition (− 1 vs. − 2) tests the predicted ordinal interaction. Use of such contrast weights is in line with extant accounting literature (e.g., Tan et al. 2019; Koonce et al. 2019; Lambert and Agoglia 2011; Kadous et al. 2003).

Table 3 Tests of H2a and H2b

Source of variation	df	MS	F-stat	p value[e]
Panel A: planned contrasts to test H2a—positive vs. neutral performance information[a]				
Model contrast	1	15.87	7.76	< 0.01
Source of variation	df	MS	F-stat	p value[e]
Panel B: planned contrasts to test H2a—negative vs. neutral performance information[b]				
Model contrast	1	34.77	11.14	< 0.01
Source of variation	df	MS	F-stat	p value[e]
Panel C: planned contrasts to test H2b—positive vs. neutral performance information[c]				
Model contrast	1	21.15	6.77	0.01
Source of variation	df	MS	F-stat	p value[e]
Panel D: Planned Contrasts to test H2b—negative vs. neutral performance information[d]				
Model contrast	1	10.89	3.82	0.05

Dependent variable: performance evaluation

[a]The contrast weights are − 2 for neutral performance/control, − 1 for neutral performance/positive likeability, − 1 for positive performance/control, and + 4 for the positive performance/positive likeability n = 111

[b]The contrast weights are − 4 for negative performance/control, + 1 for neutral performance/control, + 1 for negative performance/positive likeability, and + 2 for neutral performance/positive likeability, n = 115

[c]The contrast weights are − 2 for neutral performance/negative likeability, − 1 for positive performance/negative likeability, − 1 for neutral performance/control, and + 4 for the positive performance/control, n = 96

[d]The contrast weights are − 4 for negative performance/negative likeability, + 1 for neutral performance/negative likeability, + 1 for negative performance/control, and + 2 for neutral performance/control, n = 97

[e]All reported p values are two-tailed

q^2 of 0.03 implies that only 3% of the systematic variance is not explained by the contrast model. Taken together, these results support H2a, indicating that in the presence of positive subordinate likeability, managers will inflate their subjective performance evaluations given positive (negative) performance information to a greater (lesser) extent than in the absence of such likeability.

4.2.3 Tests of hypothesis 2b

With H2b, we predict that in the presence of negative subordinate likeability, managers will deflate their subjective performance evaluations given negative (positive) performance information to a greater (lesser) extent than in the absence of such likeability. To test H2b, we compare participants in the negative likeability condition to those in the control condition. Similar to our test of H2a, we restrict the analysis to positive and neutral performance information for the first part of the hypothesis test. We employ

the following contrast weights: − 2 for the neutral performance/negative likeability condition, − 1 for the positive performance/negative likeability condition, − 1 for the neutral performance/control condition, and + 4 for the positive performance/control condition. The results are shown in panel C of Table 3. The contrast model is significant ($F = 6.77$, $p = 0.01$, two-tailed). The residual between-groups variance test is not significant ($F = 0.74$, $p = 0.47$, two-tailed), and q^2 amounts to 0.20.

To compare participants who received negative or neutral performance information, we employ the following contrast weights: − 4 for the negative performance/negative likeability condition, + 1 for the neutral performance/negative likeability condition, + 1 for the negative performance/control condition, and + 2 for the neutral performance/control condition. While the contrast model presented in panel D of Table 3 marginally reaches a conventional level of significance ($F = 3.82$, $p = 0.05$, two-tailed), the semi-omnibus *F* test for the residual variance shows a similar level of significance ($F = 2.90$, $p = 0.06$, two-tailed). Moreover, a q^2 of 0.61 suggests that our contrast model explains less than half of the systematic variance. In sum, we thus conclude that our results provide mixed support for H2b: the evidence suggests that in the presence of negative subordinate likeability, managers deflate their subjective performance evaluations given positive performance information to a lesser extent than in the absence of such likeability. However, results do not indicate that in the presence of such negative subordinate likeability, managers deflate their subjective performance evaluations given negative performance information to a greater extent than in the absence of likeability.

Figure 2 displays the interaction effect of likeability and performance information in graphic form (cf. Guggenmos et al. 2018). It can be observed that, in comparison to the control group, participants in the positive likeability condition put more weight on positive information and less weight on negative information. Participants in the negative likeability condition appear to have put little weight on positive information. However, in comparison to the control group, participants in the negative likeability condition do not appear to have put greater weight on negative information.

4.3 Supplementary analyses

4.3.1 Additional tests of biased weighting of performance information

As the inspection of visual fit suggests, mixed support for H2b might stem from participants in the control group not weighting negative performance deviations similar to equal-in-magnitude positive performance deviations (i.e., in a linear manner). Hence, to further investigate our proposed biased weighting mechanism, we analyze the positive and negative likeability conditions separately. First, we turn to the positive likeability condition. The observed means in Table 1 show that participants inflated their performance evaluations in the presence of positive performance information (mean = 6.03)—compared to neutral performance information (mean = 5.15)—to a greater extent than they downward adapted their performance evaluations in the presence of equal-in-magnitude negative performance information (mean = 4.68). To provide a formal test, we employ the following contrast weights: − 2 for negative

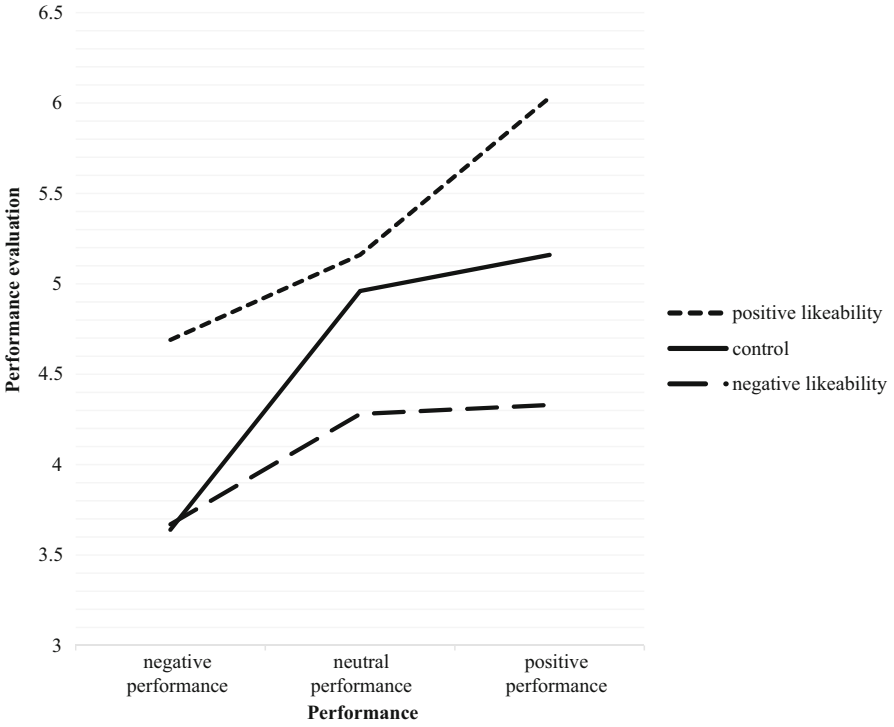

Fig. 2 Observed interaction effect between likeability and performance. This figure presents the observed pattern of cell means for participants' performance evaluation on whether (1) the subordinate was presented likeable (positive likeability), not familiarized (control) or dislikeable (negative likeability) and (2) the subordinate's performance being negative, neutral or positive

performance information, − 1 for neutral performance information, and + 3 for positive performance information. As untabulated results show, the predicted pattern is significant ($t = 3.32$, $p < 0.01$, one-tailed). The residual variance is insignificant ($F = 0.48$, $p = 0.49$, two-tailed), and q^2 amounts to 0.03.

Turning to the negative likeability condition, the observed means also suggest that participants deflated their performance evaluations in the presence of negative performance information (mean = 3.67)—compared to the neutral performance information (mean = 4.28)—to a greater extent than they upward adapted their performance evaluations in the presence of equal-in-magnitude positive performance information (mean = 4.33). We employ a similar (but reversed) set of contrast weights: − 3 for negative performance information, + 1 for neutral performance information, and + 2 for positive performance information. In line with our expectations, untabulated results show that the predicted pattern is marginally significant ($t = 1.43$, $p = 0.08$, one-tailed). The residual variance is again insignificant ($F = 0.06$, $p = 0.81$, two-tailed), and a q^2 of 0.01 indicates that only 1% of the systematic variance is not explained by the contrast model. Taken together, these results further corroborate our prediction that a biased weighting mechanism underlies likeability bias.

Table 4 Descriptive statistics for perceived likeability across conditions

Likeability	Performance			
	Negative	Neutral	Positive	Total
Negative	1.79 (0.78) n = 24	2.24 (1.30) n = 25	2.29 (1.38) n = 21	2.10 (1.18) n = 70
Control	3.16 (1.34) n = 25	3.65 (1.56) n = 23	3.74 (0.86) n = 27	3.52 (1.28) n = 75
Positive	4.11 (1.43) n = 35	4.34 (1.38) n = 32	4.75 (1.41) n = 29	4.38 (1.42) n = 96
Total	3.17 (1.57) n = 84	3.49 (1.65) n = 80	3.73 (1.57) n = 77	3.45 (1.61) n = 241

Perceived likeability is measured during the post-experimental questionnaire by asking participants to indicate their level of agreement to the statement "I liked store manager Michael Schmitz." on a 7-point Likert-scale

4.3.2 The effects of performance on perceived likeability

Our full-factorial design allows us to conduct supplementary analyses of the influence of performance on perceived likeability and subsequent performance evaluations. We first present descriptive statistics for perceived likeability in Table 4.

In line with our intended manipulation and preliminary analysis, Table 4 reveals that participants in the positive likeability condition perceived Michael Schmitz as more likeable (mean = 4.38, sd = 1.42) than participants in the control condition (mean = 3.52, sd = 1.28) or the negative likeability condition (mean = 2.10, sd = 1.18). However, as Table 4 further reveals, Michael Schmitz' perceived likeability appears to be affected by his performance as participants in the positive performance condition perceived him to be more likeable (mean = 3.73, sd = 1.57) than participants in the neutral performance condition (mean = 3.49, sd = 1.65) or the negative performance condition (mean = 3.17, sd = 1.57). Hence, in panel A of Table 5, we run an ANOVA with perceived likeability as the dependent variable and the two manipulations as factors. We find a significant main effect for both the likeability manipulation ($F = 63.00$, $p<0.01$, two-tailed) and the performance manipulation ($F = 4.01$, $p = 0.02$, two-tailed) on perceived likeability.

Panel B of Table 5 shows the results of an ANCOVA including perceived likeability as a covariate in our main model. The results show that while perceived likeability significantly affects performance evaluations ($F = 15.91$, $p<0.01$, two-tailed), the effect of our likeability manipulation is no longer significant ($F = 1.17$, $p = 0.31$, two-tailed). Hence, as expected, the results suggest that perceived likeability fully mediates the effect of the likeability manipulation. However, our performance manipulation is still significant ($F = 9.26$, $p<0.01$, two-tailed) even though the effect is of a smaller magnitude than in the unmediated model ($F = 12.29$, $p<0.01$, two-tailed), which indicates partial mediation. Figure 3 visualizes these results.

Table 5 Additional mediation analysis of perceived likeability

Source of variation	SS	df	MS	F-stat	p value[c]
Panel A: ANOVA with mediator as dependent variable and treatments as between factors[a]					
Likeability	213.24	2	106.62	63.00	< 0.01
Performance	13.57	2	6.79	4.01	0.02
Likeability × performance	1.22	4	0.31	0.1895	0.35
Error	392.64	232	1.69		
Source of variation	SS	df	MS	F-stat	p value[c]
Panel B: ANCOVA with treatments as between factors and mediator as covariate[b]					
Likeability	6.05	2	3.02	1.17	0.31
Performance	47.92	2	23.96	9.26	< 0.01
Likeability × performance	10.78	4	2.69	1.04	0.39
Perceived likeability	41.18	1	41.18	15.91	< 0.01
Error	598.05	231	2.59		

[a]Dependent variable: perceived likeability (n = 241), $R^2 = 0.37$
[b]Dependent variable: performance evaluation (n = 241), $R^2 = 0.23$
[c]All reported p values are two-tailed

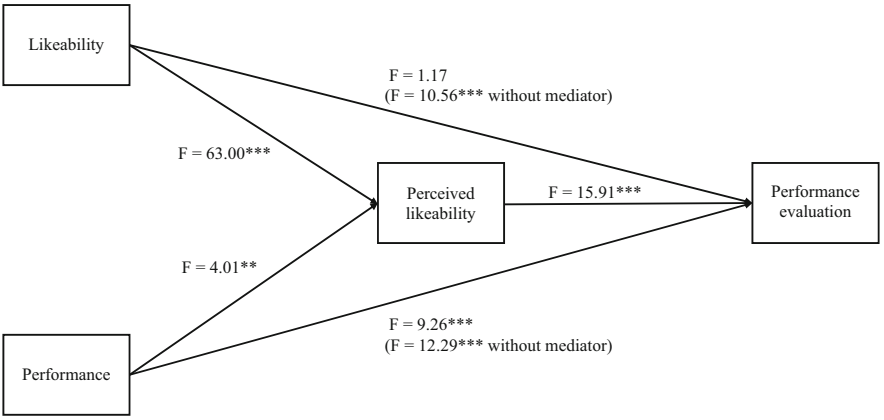

Fig. 3 The mediating effect of perceived likeability. This figure shows the results of a series of ANOVAs depicted in Tables 2 and 4. *, **, **Represent significance at the 0.10, 0.05, and 0.01 levels, respectively (two-tailed)

4.3.3 Post-experimental questionnaire

Finally, information on certain variables collected in the post-experimental questionnaire provides further insights into the proposed biased weighting mechanism. We argued that participants place greater weight on likeability-consistent than likeability-inconsistent performance information, a behavior which might be both conscious and unconscious. Thus, we first asked participants to indicate, on a scale from 1 (does not

apply at all) to 7 (fully applies), whether they placed equal weight on all performance measures presented. An untabulated ANOVA shows that there are no significant differences between conditions ($p = 0.65$) and participants' mean score on this item (mean $= 4.58$, sd $= 1.94$) is significantly above the scale midpoint of 4 ($t = 4.60$, $p < 0.01$, two-tailed).

Furthermore, we asked participants to indicate, using the same scale, whether they placed greater weight on performance measures for which Michael Schmitz achieved a low value than on performance measures for which he achieved a high value. Participants in the negative likeability condition indicate higher agreement compared to participants in the positive likeability condition (mean $= 3.49$, sd $= 0.21$ vs. mean $= 3.03$, sd $= 0.17$, $t = 1.70$, $p = 0.05$, one-tailed). However, participants in both conditions seem to rather disagree with the statement as the mean scores of both participants in the positive likeability condition ($t = -5.64$, $p < 0.01$, two-tailed) and participants in the negative likeability condition ($t = -2.37$, $p = 0.02$, two-tailed) are significantly below the scale midpoint.

Hence, although we acknowledge that self-reported measures of decision processes should be treated with some caution, these results indicate that the mechanism of biased weighting of likeability-consistent and likeability-inconsistent performance information causing bias in performance evaluations seems to operate predominantly unconscious.

5 Discussion and conclusion

Subjective performance evaluations are often assessed in a dyadic manager–subordinate setting, entailing managers' subjective weighting of multiple performance measures to determine overall performance evaluations. Prior research suggests that subordinates' likeability may affect performance evaluations in such settings because managers engage in an affect-consistency heuristic, placing greater weight on likeability-consistent performance information than on likeability-inconsistent information (e.g., Kaplan et al. 2007; Robbins and DeNisi 1994). In this paper, we examine the interaction of likeability and valence of subordinate performance information. Using an experiment, we find that managers asked to subjectively evaluate a likeable subordinate inflate their evaluations in the presence of positive performance information to a greater extent than they downward-adapt their evaluations in the presence of equal-in-magnitude negative information. When asked to evaluate a dislikeable subordinate, they deflate their evaluations in the presence of negative performance information to a greater extent than they upward-adapt their evaluations in the presence of equal-in-magnitude positive information.

Since we vary only one performance measure between the different performance conditions in our experiment, our results indicate a biased weighting mechanism. Hence, our results are in line with the affect-consistency heuristic and imply that managers place greater weight on likeability-consistent performance measures than on likeability-inconsistent measures (Robbins and DeNisi 1994). Our study thus extends prior research implying a biased weighting of performance measures as the mechanism underlying bias in performance evaluations (e.g., Kaplan et al. 2007) by directly exam-

ining this mechanism. It also contributes to the accounting literature on managers' cognitive processing of multiple performance measures in subjective performance evaluations in general (e.g., Lipe and Salterio 2000). We thereby also add to the general discussion on affect in accounting contexts (e.g., Fehrenbacher et al. 2019; Elliott et al. 2014) by showing that affect may cause decision-makers to place different weights on positively and negatively valenced information.

Additional analyses of our experiment further reveal that as well as directly affecting evaluations, subordinates' performance also indirectly affects managers' performance evaluations through perceived likeability, which may exacerbate likeability bias. While prior research has already shown that managers' impressions of subordinates' prior performance may cause bias in subjective performance evaluations (e.g., Reilly et al. 1998; Balzer 1986), the influence of current performance has, to the best of our knowledge, been largely unexplored. Our results suggest that while performance information is appropriately incorporated into managers' performance evaluations, it may also cause affective reactions: Managers perceive poorly performing subordinates as less likeable and rate them therefore unduly severe. Favorable performance, in turn, results in the opposite behavior. In consequence, this might induce a feedback loop, in particular at the beginning of a manager-subordinate relationship: initial performance of an employee may cause likeability, which then affects how the further performance of this employee is perceived. These findings extend the stream of research that investigates the interplay between performance information and likeability in subjective performance evaluations (e.g., Varma and Pichler 2007). Hence, we contribute to recent discussions on "rater bias" versus "true performance" perspectives (e.g., Sutton et al. 2013) by showing that subordinates' *current* performance affects their performance evaluations both directly and indirectly through managers' perceptions of subordinates as likeable or dislikeable.

In the light of the concern, among both researchers and practitioners, with bias in performance evaluations, our results also have important practical implications. A sound understanding of the possible adverse effects of subjectivity in performance evaluation is crucial for informing organizations about possible interventions. When managers' evaluations of more or less likeable subordinates are not solely vertically 'shifted', but instead likeability interacts with subordinates' performance level (i.e., likeability causes managers to engage in a biased weighting and rate the same performance differently for more or less likeable subordinates), the perceived fairness of organizations' performance evaluation systems can be particularly impaired. This can reduce employee motivation and, thus, be hazardous for organizations. However, since organizations are most likely not capable of controlling employees' emotions (i.e., their interpersonal relationships and their affective reactions triggered by certain levels of performance), organizations should consider complementing the evaluation procedure with other (informal) controls to increase the accuracy and employees' fairness perceptions of performance evaluations. Possible controls include (but are not limited to) accountability structures, feedback, and training (Kang and Fredin 2012; Dilla and Steinbart 2005; Libby et al. 2004).

Like all research, this study has limitations, which have to be taken into account when interpreting its results while also offering opportunities for future research. First, it is possible that participants perceived part of our likeability manipulation as

performance-relevant: For example, it could be argued that "self-centered" can be hindering in some tasks but also helpful in others. Thus, without a clear description of the requirements of the subordinate's tasks, a performance-related (positive or negative) interpretation of the wording of our manipulation is highly subjective. Due to experimental randomization, we therefore expect this not to affect our results. Furthermore, while this might only affect the level of bias between the likeability conditions, it cannot explain the obtained interactions.

Second, with our likeability manipulation, we aim to trigger interpersonal affect through a written description of a fictional subordinate. While this procedure is in line with experimental accounting research on affect (e.g., Fanning and Piercey 2014; Bhattacharjee et al. 2012; Moreno et al. 2002), we acknowledge that it is difficult to trigger an affective reaction through written text. However, we reason that this would work against finding any effects. While our results show the direction of the effects, we thus believe that the real-world effects of likeability on the weighting of performance measures in subjective performance evaluations may be of greater magnitude.[21]

Third, to be able to directly test biased weighting behavior, we provided a limited amount of performance measures in our experiment. Future research could examine the effects of likeability at different levels of information load and, thus, explore whether exceeding managers' cognitive capacities exacerbates likeability-caused bias (Schick et al. 1990; Miller 1956). In particular, it is possible that in such a scenario, managers might engage in a non-exhaustive information search and shift their attention predominantly to likeability-consistent performance measures (Sohn et al. 2019; Kaplan et al. 2007).

Finally, we only consider a one-period scenario in our study. In practice, manager—subordinate dyads often persist over multiple periods. Future research might address this fact and examine whether the effect of likeability varies when subordinates constantly show favorable (or unfavorable) performance and whether current performance exerts a greater influence on likeability and bias in performance evaluations than prior performance impressions.

Acknowledgements The authors greatly appreciate the helpful comments and suggestions from Hans-Ulrich Küpper, Thorsten Knauer, Philipp Schreck, Friedrich Sommer, and Arnt Wöhrmann (editors) as well as two anonymous reviewers. We also thank Markus Arnold, Stephan Kramer, Matthias Sohn as well as participants at the 2019 AAA MAS midyear meeting, the 2018 EAA annual meeting, and the 2018 VHB annual meeting for helpful comments.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

[21] In line with this argument, the literature acknowledges that experiments are usually not well-suited to detect effect sizes which can be extrapolated to real-world settings but rather aim to test the direction of effects (Kadous and Zhou 2018).

References

Antonioni D, Park H (2001) The relationship between rater affect and three sources of 360-degree feedback ratings. J Manag 27:479–495

Arnold MC, Tafkov ID (2019) Managerial discretion and task interdependence in teams. Contemp Account Res 36:2467–2493

Ashton RH, Kramer SS (1980) Students as surrogates in behavioral accounting research. Some evidence. J Account Res 18:1–15

Baltes BB, Parker CP (2000) Reducing the effects of performance expectations on behavioral ratings. Organ Behav Hum Decis Process 82:237–267

Balzer WK (1986) Biases in the recording of performance-related information: the effects of initial impression and centrality of the appraisal task. Organ Behav Hum Decis Process 37:329–347

Banker RD, Chang H, Pizzini MJ (2004) The balanced scorecard: judgmental effects of performance measures linked to strategy. Account Rev 79:1–23

Bhattacharjee S, Moreno KK, Riley T (2012) The interplay of interpersonal affect and source reliability on auditors' inventory judgments. Contemp Account Res 29:1087–1108

Bol JC (2011) The determinants and performance effects of managers' performance evaluation biases. Account Rev 86:1549–1575

Bol JC, Kramer S, Maas VS (2016) How control system design affects performance evaluation compression: the role of information accuracy and outcome transparency. Account Organ Soc 51:64–73

Buckless FA, Ravenscroft SP (1990) Contrast coding: a refinement of ANOVA in behavioral analysis. Account Rev 65:933–945

Cardinaels E, van Veen-Dirks PM (2010) Financial versus non-financial information: the impact of information organization and presentation in a Balanced Scorecard. Account Organ Soc 35:565–578

Cardy RL, Dobbins GH (1986) Affect and appraisal accuracy. Liking as an integral dimension in evaluating performance. J Appl Psychol 71:672–678

Carmona S, Iyer G, Reckers PM (2014) Performance evaluation bias. A comparative study on the role of financial fixation, similarity-to-self and likeability. Adv Account 30:9–17

Chen Y, Jermias J, Panggabean T (2016) The role of visual attention in the managerial judgment of Balanced-Scorecard performance evaluation: insights from using an eye-tracking device. J Account Res 54:113–146

Dai NT, Kuang X, Tang G (2018) Differential weighting of objective versus subjective measures in performance evaluation: experimental evidence. Eur Account Rev 27:129–148

Demeré BW, Sedatole KL, Woods A (2018) The role of calibration committees in subjective performance evaluation systems. Manag Sci 65:1562–1585

DeNisi AS, Robbins TL, Summers TP (1997) Organization, processing, and use of performance information: a cognitive role for appraisal instruments. J Appl Soc Psychol 27:1884–1905

Dilla WN, Steinbart PJ (2005) Relative weighting of common and unique Balanced Scorecard measures by knowledgeable decision makers. Behav Res Account 17:43–53

Ding S, Beaulieu P (2011) The role of financial incentives in Balanced Scorecard-based performance evaluations: correcting mood congruency biases. J Account Res 49:1223–1247

Elliott WB, Hodge FD, Kennedy JJ, Pronk M (2007) Are M.B.A. students a good proxy for nonprofessional investors? Account Rev 82:139–168

Elliott WB, Jackson KE, Peecher ME, White BJ (2014) The unintended effect of corporate social responsibility performance on investors' estimates of fundamental value. Account Rev 89:275–302

Fanning K, Piercey MD (2014) Internal auditors' use of interpersonal likability, arguments, and accounting information in a corporate governance setting. Account Organ Soc 39:575–589

Farrell AM, Goh JO, White BJ (2014) The effect of performance-based incentivecontracts on system 1 and system 2 processing in affective decision contexts: fMRI and behavioral evidence. Account Rev 89:1979–2010

Fehrenbacher DD, Schulz AK-D, Rotaru K (2018) The moderating role of decision mode in subjective performance evaluation. Manag Account Res 41:1–10

Fehrenbacher DD, Kaplan SE, Moulang C (2019) The role of accountability in reducing the impact of affective reactions on capital budgeting decisions. Manag Account Res. https://doi.org/10.1016/j.mar.2019.100650

Feldman J (1981) Beyond attribution theory: cognitive processes in performance appraisal. J Appl Psychol 66:127–148

Festinger L (1957) A theory of cognitive dissonance. Stanford University Press, Radwood City
Foti RJ, Hauenstein NM (1993) Processing demands and the effects of prior impressions on subsequent judgments: clarifying the assimilation/contrast debate. Organ Behav Hum Decis Process 56:167–189
Guggenmos RD, Pierce MD, Agoglia CP (2018) Custom contrast testing: current trends and a new approach. Account Rev 93:223–244
Haynes CM, Kachelmeier SJ (1998) The effects of accounting contexts on accounting decisions: a synthesis of cognitive and economic perspectives in accounting experimentation. J Account Lit 17:97–136
Kadous K, Zhou Y (2018) Maximizing the contribution of JDM-style experiments in accounting. In: Libby T, Thorne L (eds) The Routledge companion to behavioural accounting research. Routledge, London, pp 175–192
Kadous K, Kennedy SJ, Peecher ME (2003) The effect of quality assessment and directional goal commitment on auditors' acceptance of client-preferred accounting methods. Account Rev 78:759–778
Kang G, Fredin A (2012) The balanced scorecard: the effects of feedback on performance evaluation. Manag Res Rev 35:637–661
Kaplan SE, Petersen MJ, Samuels JA (2007) Effects of subordinate likeability and Balanced Scorecard format on performance-related judgments. Adv Account 23:85–111
Kaplan SE, Petersen MJ, Samuels JA (2017) Further evidence on the negativity bias in performance evaluation: when does the evaluator's perspective matter? J Manag Account Res 30:169–184
Kaplan SE, Samuels JA, Sawers KM (2018) Social psychology theories as applied to behavioural accounting research. In: Libby T, Thorne L (eds) The Routledge companion to behavioural accounting research. Routledge, London, pp 497–506
Kida TE, Moreno KK, Smith JF (2001) The influence of affect on managers' capital-budgeting decisions. Contemp Account Res 18:477–494
Koonce L, Leitter Z, White BJ (2019) Linked balance sheet presentation. J Account Econ 68:1–16
Kramer S, Maas VS (2019) Selective attention as a determinant of escalation bias in subjective performance evaluation judgments. Behav Res Account. https://doi.org/10.2308/bria-18-021
Kravitz DA, Balzer WK (1992) Context effects in performance appraisal: a methodological critique and empirical study. J Appl Psychol 77:24–31
Kunda Z (1990) The case for motivated reasoning. Psychol Bull 108:480–498
Lambert TA, Agoglia CP (2011) Closing the loop: review process factors affecting audit staff follow-through. J Account Res 49:1275–1306
Lefkowitz J (2000) The role of interpersonal affective regard in supervisory performance ratings: a literature review and proposed causal model. J Occup Organ Psychol 73:67–85
Libby R, Bloomfield R, Nelson MW (2002) Experimental research in financial accounting. Account Organ Soc 27:775–810
Libby T, Salterio SE, Webb A (2004) The Balanced Scorecard: the effects of assurance and process accountability on managerial judgment. Account Rev 79:1075–1094
Lipe MG, Salterio SE (2000) The Balanced Scorecard: judgmental effects of common and unique performance measures. Account Rev 75:283–298
Luft J, Shields MD (2009) Psychology models of management accounting. Found Trends Account 4:199–345
Maas VS, Torres-González R (2011) Subjective performance evaluation and gender discrimination. J Bus Ethics 101:667–681
Maas VS, Verdoorn N (2017) The effects of performance report layout on managers' subjective evaluation judgments. Account Bus Res 47:731–751
Maas VS, van Rinsum M, Towry KL (2012) In search of informed discretion: an experimental investigation of fairness and trust reciprocity. Account Rev 87:617–644
Miller G (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63:81–97
Moers F (2005) Discretion and bias in performance evaluation: the impact of diversity and subjectivity. Account Organ Soc 30:67–80
Moreno KK, Kida TE, Smith JF (2002) The impact of affective reactions on risky decision making in accounting contexts. J Account Res 40:1331–1349
Ravenscroft SP, Buckless FA (2018) Contrast coding in ANOVA and regression. In: Libby T, Thorne L (eds) The Routledge companion to behavioural accounting research. Routledge, London, pp 349–372

Reilly SP, Smither JW, Warech MA, Reilly RR (1998) The influence of indirect knowledge of previous performance on ratings of present performance: the effects of job familiarity and rater training. J Bus Psychol 12:421–435
Robbins TL, DeNisi AS (1994) A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations. J Appl Psychol 79:341–353
Robbins TL, DeNisi AS (1998) Mood vs. interpersonal affect: identifying process and rating distortions in performance appraisal. J Bus Psychol 12:313–325
Robertson JC, Stefaniak CM, Curtis MB (2011) Does wrongdoer reputation matter? Impact of auditor-wrongdoer performance and likeability reputations on fellow auditors' intention to take action and choice of reporting outlet. Behav Res Account 23:207–234
Salterio SE (2014) We don't replicate accounting research—or do we? Contemp Account Res 31:1134–1142
Schick AG, Gordon LA, Haka S (1990) Information overload: a temporal approach. Account Organ Soc 15:199–220
Schwering A (2017) The influence of peer honesty and anonymity on managerial reporting. J Bus Econ 87:1151–1172
Shields MD (2015) Established management accounting knowledge. J Manag Account Res 27:123–132
Sohn M, Hirsch B, Schulte-Mecklenbeck M (2019) The effect of information search and attention distribution on the common measure bias in performance evaluations. Working paper. https://ssrn.com/abstract=3240457. Accessed 12 Mar 2020
Steiner DD, Rain JS (1989) Immediate and delayed primacy and recency effects in performance evaluation. J Appl Psychol 74:136–142
Sutton AW, Baldwin SP, Wood L, Hoffman BJ (2013) A meta-analysis of the relationship between rater liking and performance ratings. Hum Perform 26:409–429
Tan HT, Wang EY, Yoo GS (2019) Who likes jargon? The joint effect of jargon type and industry knowledge on investors' judgments. J Account Econ 67:416–437
Tsui AS, Barry B (1986) Interpersonal affect and rating errors. Acad Manag J 29:586–599
Varma A, Pichler S (2007) Interpersonal affect: does it really bias performance appraisals? J Lab Res 28:397–412
Varma A, DeNisi AS, Peters LH (1996) Interpersonal affect and performance appraisal: a field study. Pers Psychol 49:341–360
Voußem L, Kramer S, Schäffer U (2016) Fairness perceptions of annual bonus payments. The effects of subjective performance measures and the achievement of bonus targets. Manag Account Res 30:32–46
Woods A (2012) Subjective adjustments to objective performance measures: the influence of prior performance. Account Organ Soc 37:403–425
Xu Y, Tuttle BM (2005) The role of social influences in using accounting performance information to evaluate subordinates: a causal attribution appoach. Behav Res Account 17:191–210