4-2019

Restructured frame-of-reference training improves rating accuracy

Ming-Hong TSAI
Singapore Management University, mhtsai@smu.edu.sg

Serena WEE
University of Western Australia

Brandon KOH
Singapore Management University, brandon.koh.2014@phdps.smu.edu.sg

Citation

TSAI, Ming-Hong, WEE, Serena, & KOH, Brandon.(2019). Restructured frame-of-reference training improves rating accuracy. *Journal of Organizational Behavior, 40(6)*, 740-757.
Available at: https://ink.library.smu.edu.sg/soss_research/2824

Received: 3 October 2018 | Revised: 31 March 2019 | Accepted: 5 April 2019

DOI: 10.1002/job.2368

RESEARCH ARTICLE

WILEY Journal of Organizational Behavior

Restructured frame-of-reference training improves rating accuracy

Ming-Hong Tsai[1] | Serena Wee[2] | Brandon Koh[1]

[1] School of Social Sciences, Singapore Management University, Singapore

[2] School of Psychological Science, The University of Western Australia, Perth, Western Australia, Australia

Correspondence
Ming-Hong Tsai, School of Social Sciences, Singapore Management University, 90 Stamford Road, Level 4, Singapore 178903.
Email: mhtsai@smu.edu.sg

Summary

The use of heuristic judgments is prevalent in organizations and negatively impacts accurate employee assessments. To minimize the negative impact of heuristic judgments (i.e., anchoring and adjustment), we aim to improve rating accuracy by restructuring frame-of-reference (FOR) training. We conducted five studies (N = 1,143) using different samples (three including participants with hiring experience), training environments (onsite and online), and rating contexts (evaluations of sales representatives, teachers, contract negotiation specialists, and retail store managers). Across the five studies, the average improvement in rating accuracy was at least twice as large for restructured FOR (vs. control) training as it was for typical FOR (vs. control) training; the difference in rating accuracy between restructured and typical FOR training was statistically significant. Furthermore, minimizing the anchoring effect rather than increasing opportunities for rating adjustments improved rating accuracy (Study 4). Finally, restructured FOR training achieved higher criterion validity (i.e., a higher strength of the association between ratings regarding a target and the target's objective performance) than did typical FOR training (Studies 3 and 5). We discuss implications for improving the effectiveness of diverse training programs and the accuracy of judgments in organizations.

KEYWORDS

anchoring and adjustment heuristic, frame-of-reference, judgment, rating accuracy, subjective evaluation

1 | INTRODUCTION

Assessments are an essential process for organizations to evaluate performance, motivate employees, provide feedback, identify training needs and growth, and distribute rewards fairly. Therefore, understanding and refining rater training methods to enhance assessment accuracy will be a highly valuable endeavor. Organizations benefit from accurate employee assessments, which are associated with a wide range of positive consequences, such as superior job performance (Abbas, 2014), enhanced perceptions of procedural and informational justice (Roberson & Stewart, 2006), increased appraisal satisfaction, and elevated motivation to improve future job performance (Selvarajan & Cloninger, 2012). Despite these benefits, raters often do not provide accurate ratings due to their use of heuristic-based judgments during the evaluation process (Reb, Greguras, Luan, & Daniels, 2014). Rater training aims to mitigate the use of these heuristic judgments, therefore improving rating accuracy (Uggerslev & Sulsky, 2008). In addition, rater training programs help participants to adopt organizational goals, develop skills related to feedback delivery, and increase confidence in performing assessments (Kumar, 2005; Nesbit & Wood, 2002).

In particular, frame-of-reference (FOR) training (Bernardin & Buckley, 1981) is an effective and frequently used rater training approach. This approach uses a *practice-then-feedback* procedure to instill established standards for evaluation (Lievens & Sanchez, 2007; Roch, Woehr, Mishra, & Kieszczynska, 2012; Woehr & Huffcutt, 1994).

Nonetheless, the practice-then-feedback procedure in typical FOR training could—counterintuitively—make it *more difficult* for a rater to provide accurate ratings. As we argue below, the procedure increases a rater's tendency to rely on the anchoring and adjustment heuristic (Tversky & Kahneman, 1974). That is, practice-then-feedback could potentially result in two crucial shortcomings of the typical FOR training method: an initial anchoring effect and subsequent insufficient adjustments. To address these limitations, we restructure the procedures of FOR training, by presenting evaluation standards before practice rating trials and by offering opportunities for sufficient rating adjustments, and investigate whether or not this restructured FOR method improves training effectiveness.

Overall, we attempt to make two contributions to the literature. First, we investigate how susceptible rater trainees are to the anchoring and adjustment heuristic and how rater training procedures may affect this heuristic. For example, it is important to examine the impact of the heuristic on rating accuracy because research in the domain of marketing and consumer behavior has shown that—due to a reliance on the anchoring and adjustment heuristic—people tend to focus on their initial evaluation, even to the point that they ignore other information that could have facilitated more accurate evaluations (Naylor, Lamberton, & Norton, 2011). It, therefore, seems plausible that a similar situation might occur when raters use their first evaluation in practice trials as an anchor to perform subsequent assessments. In addition, when people employ the anchoring and adjustment heuristic, they are not motivated to make large adjustments during subsequent evaluations (Epley & Gilovich, 2006). This phenomenon of insufficient adjustment could affect the accuracy of ratings such that raters may not sufficiently adjust their ratings based on feedback from a training advisor. Thus, we examined whether a rater's behavior could be systematically influenced by training procedures that are designed to mitigate an initial anchoring effect and to resolve the issue of insufficient adjustments. Specifically, we investigated the impacts of restructured information presentation and opportunities for rating adjustment on rating accuracy.

Furthermore, we investigate these impacts both simultaneously and independently. This approach differs from existing research on the anchoring and adjustment heuristic, which examines the accessibility of an anchor (e.g., Naylor et al., 2011) and insufficient adjustment (e.g., Epley, Keysar, Van Boven, & Gilovich, 2004) as separate reasons for why this effect would occur. A simultaneous evaluation of both factors can examine whether these two factors are equally or differentially associated with rating accuracy. Importantly, this approach allows us to investigate whether these factors will amplify or weaken each other's effect on rating accuracy (i.e., the interaction effects of the two factors on rating accuracy). Thus, our approach offers a more integrated examination of the factors underlying the anchoring and adjustment heuristic than does previous research.

By examining the assumptions outlined in our first contribution, our second contribution is an unabashedly practical one: to develop a novel training intervention that yields higher rating accuracy than typical FOR training. Although an enhancement of training effectiveness on rating accuracy is a major goal in rater training research, research in the most recent 5 years has focused on *applying* the FOR training method (e.g., Firth, Hollenbeck, Miles, Ilgen, & Barnes, 2015) rather than on improving training effectiveness. To continue with the pursuit of enhanced training effectiveness, we explore an unexamined intervention that could potentially add new training principles and procedures to the field of rater training. Although practice-then-feedback procedures in a typical FOR training process have been considered as sufficiently effective in the previous research (Roch et al., 2012), we investigate whether restructured FOR training procedures can further improve rating accuracy by minimizing the anchoring and adjustment heuristic. The anchoring and adjustment heuristic has also been demonstrated in a wide range of organizationally relevant contexts, such as negotiation (Gunia, Swaab, Sivanathan, & Galinsky, 2013), selection interviews (Kataoka, Latham, & Whyte, 1997), and team decision-making (Lehner, Seyed-Solorforough, O'Connor, Sak, & Mullin, 1997), where it has been shown to result in suboptimal evaluation outcomes. Despite its relevance and prevalence, relatively little is known about its effects on rater training effectiveness. Given that the restructured FOR training procedures are designed to mitigate an overreliance on the anchoring and adjustment heuristic, significant improvement in rating accuracy from restructured training procedures can implicate the insufficient effectiveness of the practice-then-feedback procedures due to the anchoring and adjustment heuristic. Therefore, our research not only includes a practical and easily implementable solution to rater training but also clarifies how the anchoring and adjustment heuristic affects rater training effectiveness in practice-then-feedback procedures.

To provide background information about our research question, we elaborate on existing FOR training and restructured FOR training and examine how different types of rater training influence rating accuracy and criterion validity in the subsequent sections. Afterward, we explore our research question using five studies with different rating scenarios, performance dimensions, and samples to increase the generalizability of our results. Finally, we discuss the theoretical and practical implications of our findings to present how our research can contribute to organizational management.

2 | FOR TRAINING

The overall premise of FOR training is that individuals have idiosyncratic knowledge (i.e., a personal schema or implicit theory), which differs from the more widely held, and often explicitly stated, institutional knowledge (i.e., the referent schema or espoused theory; Bernardin & Buckley, 1981). Furthermore, the schema-based theory suggests that FOR trainees can replace their personal schema of job performance with the referent schema provided by the organization and therefore improve their rating accuracy (Cardy & Keefe, 1994; Lievens & Sanchez, 2007; Sulsky & Day, 1994).

A schema composes of closely interrelated sets of knowledge about a concept (Marshall, 1995; Piaget, 1997). For example, a person's schema regarding an effective classroom instructor may include beliefs about being detailed, organized, and enthusiastic about course materials. In contrast, the university may prioritize an instructors' rapport with students over their enthusiasm about the course materials. To minimize such a gap between personal and organizational schemas, the FOR training process includes three key elements. For each performance dimension, it provides (a) a standardized definition with behavioral examples, (b) practice evaluations using a rating scale, and (c) feedback from an advisor (or expert) indicating how a rater's evaluation is discrepant from the referent schema and/or standard (Pulakos, 1984). The overall training goal involves educating raters to use common evaluation standards (i.e., the organization's schema of performance) when making evaluations.

The FOR training has been shown to improve rating accuracy across different assessment types: interviews (Melchers, Lienhardt, Von Aarburg, & Kleinmann, 2011), assessment centers (Schleicher, Day, Mayes, & Riggio, 2002), and performance appraisals (Sulsky & Day, 1992). It also improves the evaluation of individual characteristics, such as personality traits (Aguinis, Mazurkiewicz, & Heggestad, 2009), decision-making behavior (Firth et al., 2015), emotions, and social skills (Angkaw, Tran, & Haaga, 2006). Furthermore, meta-analyses indicate that FOR training has a medium-to-large effect on improving rating accuracy (d = 0.50, Roch et al., 2012; d = 0.83, Woehr & Huffcutt, 1994). In summary, FOR training has been found to be a highly versatile and effective rater training method.

However, the FOR training procedure rests on an unexamined assumption that raters can effectively replace their personal schemas with the provided referent schema. Moreover, it is assumed that the practice-then-feedback procedure improves the effectiveness of intuitive judgment (Dane & Pratt, 2007) because this procedure often provides immediate and accurate feedback (Ericsson & Lehmann, 1996). We question this assumption and seek to demonstrate why the practice-then-feedback procedure should be revised. Specifically, in the next section, we draw on specific theoretical perspectives of the anchoring and adjustment heuristic and elaborate on how this heuristic could inhibit the effectiveness of immediate and accurate feedback.

3 | RESTRUCTURED FOR TRAINING

3.1 | Heuristics in judgment and decision-making

Heuristics refer to mental shortcuts that individuals usually rely on in order to make judgments and decisions (Gigerenzer & Gaissmaier, 2010). These straightforward and easily accessible rules typically include a narrow focus on a single aspect of a complex issue and a tendency to ignore other relevant information. The major cause of using heuristics is from a basic human response to reduce effort (Shah & Oppenheimer, 2008). For instance, effort reduction can involve exploring fewer cues or options and integrating less information. Therefore, an overreliance on heuristics could result in inaccurate ratings because raters may not intend to integrate the new information obtained from a training program. The use of heuristics can also lead to a cognitive bias that indicates systematic deviations from a rational process (Tversky & Kahneman, 1974). Therefore, the utilization of heuristics may decrease the likelihood of producing accurate outcomes.

3.2 | The anchoring and adjustment heuristic

We draw on literature regarding the anchoring and adjustment heuristic (Tversky & Kahneman, 1974) to develop an understanding of the key processes that determine how raters perform assessments. Anchoring and adjustment is a mental shortcut that impacts judgment and decision-making. The anchoring and adjustment heuristic is ubiquitous in various settings of judgment (e.g., numerical estimation of geographic features; Simmons, LeBoeuf, & Nelson, 2010) and decision-making (e.g., property pricing decisions; Northcraft & Neale, 1987). On the basis of this heuristic, individuals start with an initial reference point (the "anchor") and make incremental adjustments to arrive at their final judgment based on additional information. The anchor has a significant influence on subsequent assessments, and therefore, the adjustment process from the anchor to the final conclusion is typically insufficient.

Research has also found that the insufficient adjustment from the initial anchor is more severe when the initial anchor is self-generated rather than provided by others (Epley & Gilovich, 2001). In an interpersonal situation, the anchoring and adjustment heuristic simplifies the complicated process of evaluating another's viewpoint by using a person's own viewpoint (which is easily accessible) as a substitute for the other person's viewpoint (which must be inferred before an evaluation; Epley et al., 2004). Other research also supports the fact that self-relevant characteristics constitute a highly accessible anchor for social inference. For instance, business agents apply their own preferences to predict others' preferences when the others' preferences are not revealed (West, 1996). Relatedly, advisors often give recommendations based on their idealistic considerations but fail to evaluate practical constraints that their advice receivers may face (Danziger, Montal, & Barkan, 2012). When inferring the mental states of a similar person based on his or her ambiguous actions, individuals project their personal characteristics onto this person (Ames, 2004). Therefore, the accessibility of a self-anchor may determine whether the anchoring and adjustment heuristic can occur.

Adjustment away from this initial anchor is typically insufficient because raters may stop adjusting their evaluations based on their egocentric interpretation of the established evaluation standards (Epley & Gilovich, 2004). Egocentric viewpoints refer to instances in which people often overestimate the extent to which others share their thoughts and feelings (R. S. Nickerson, 1999; Van Boven, Dunning, & Loewenstein, 2000). Despite a receipt of immediate and accurate feedback from a training advisor, raters may also have a range of ratings they would consider to be plausible based on the advisor's feedback (Quattrone, 1982). Thus, raters may feel that the advisor's ratings are more similar to their own ratings than the actual rating differences between the advisor and the raters and stop adjusting their ratings once their adjustments fall within a range of plausible ratings. To support this proposition, research has demonstrated that individuals often serially adjust from their initial estimate and stop adjusting once they achieve a minimally satisfactory estimate (Epley & Gilovich, 2006). Therefore, an intervention for sufficient adjustments may alleviate the problem of the anchoring and adjustment heuristic.

3.3 | Rationales for restructuring FOR training

To recapitulate, the *anchoring effect* occurs because a rater's first impression or rating tends to exert an excessively strong influence on the rest of the evaluation process (Thorsteinson, Breier, Atwell, Hamilton, & Privette, 2008). Thus, providing an initial rating during the practice phase may make it more difficult for a rater to incorporate the information learned during the feedback phase (i.e., the timing of receiving feedback after the initial ratings) into subsequent judgments. An accessibility to initial ratings may prevent raters from learning information about established performance standards during the feedback phase. The *insufficient adjustment effect* occurs because raters tend to overestimate the similarity between their ratings and the established performance standards (Epley et al., 2004) and consider a range of ratings to be plausible based on the established performance standards (Quattrone, 1982). Then the raters use a satisficing strategy—stop adjusting their ratings once they achieve plausible ratings that are the closest to their initial ratings (Epley & Gilovich, 2006).

To address these limitations of the practice-then-feedback procedure, we propose revising (i.e., restructuring) the typical FOR training procedure in two critical ways. First, in contrast to typical FOR training (where evaluation standards are presented as feedback *after* participants make initial ratings during the practice phase), in the restructured FOR training, we propose a presentation of evaluation standards *before* participants practice rating. We refer to this proposed change as a "restructured presentation"; this procedure is designed to minimize the accessibility of the self-generated anchor (Epley & Gilovich, 2001).

Second, in contrast to typical FOR training (where participants *do not practice* adjusting their ratings), we propose an intervention for restructured FOR training in which participants *practice* adjusting their ratings until these ratings are consistent with the training advisor's ratings. Specifically, participants keep receiving a message and an opportunity to update their ratings until their ratings are the same as the advisor's ratings. We refer to this proposed change as "adjustment opportunities"; this procedure is designed to decrease the likelihood of insufficient adjustment by offering sufficient practice opportunities for adjustment (Epley & Gilovich, 2006). Furthermore, the theory of self-perception posits that participants will observe themselves making multiple adjustments and therefore may infer that they are highly motivated to provide accurate ratings (Bem, 1972), which may strengthen their motivation to seek and provide accurate ratings in subsequent assessment tasks.

In summary, we propose that rating accuracy may be improved by restructuring the presentation of training information (restructured presentation) and by offering opportunities to make sufficient adjustments in the practice task (adjustment opportunities). We refer to FOR training that includes one or both of these revised procedures as restructured FOR training. In Table 1, we summarize the major differences between typical FOR training and restructured FOR training. Therefore, we propose the following hypothesis:

> **Hypothesis 1.** *Rating accuracy is higher in the restructured FOR training condition than in the typical FOR training condition.*

TABLE 1 Theoretical processes manipulated across FOR training methods

	FOR training method	
Theoretical processes	**Typical**	**Restructured**
	Typical presentation: Accessible self-rating anchors	Restructured presentation: Less accessible self-rating anchors
Anchoring: A restructured presentation of information to reduce the accessibility of self-rating anchors	Information on rating evaluation standards (i.e., feedback from the rating advisor) was provided *after* participants completed practice trials. Participants' initial ratings serve as anchors for subsequent assessments.	Information on rating evaluation standards was provided *before* participants completed practice trials. The rating evaluation standards rather than participants' ratings serve as anchors for subsequent assessments.
	Without adjustment	With adjustment
Adjustment: Opportunities for rating adjustment	During the practice trials, participants *received* information comparing their ratings with the evaluation standards. Then participants were requested to *adjust their ratings in the formal assessment task*. However, participants were not allowed to repeat the practice trials.	During the practice trials, participants *did not receive* information comparing their ratings with the evaluation standards. However, they received a notification message when their ratings were inconsistent with these standards. Then participants were requested to *adjust their ratings in the practice trials* until their ratings were consistent with the standards.

Abbreviation: FOR, frame-of-reference.

4 | CRITERION VALIDITY AND RATER TRAINING

Criterion validity is often held as the gold standard for rating accuracy. Criterion validity is defined as the correlation between a predictor variable (e.g., an assessor's ratings regarding a target from an interview) and an outcome variable or criterion (e.g., the target's objective sales performance), that is, the extent to which the assessor's ratings regarding a target positively predict the target's actual performance outcome (Dunnette & Borman, 1979; see also American Psychological Association, National Council on Measurement in Education,,, & American Educational Research Association, 1999). The assumption that increased rating accuracy should positively correlate with increased criterion validity stems from the fact that reliable measurement of a predictor strengthens the observed correlation between the predictor and the criterion (i.e., strengthens criterion validity). Therefore, by improving rating accuracy on the predictor measure, the expectation is that criterion validity should also improve. It is possible, however, that improved rating accuracy does not result in improved criterion validity. This would occur when improved rating accuracy results in better reliability of the predictor measure, but the predictor measure is not strongly related to the criterion. Therefore, it is important to examine whether increased rating accuracy is positively related to increased criterion validity. To date, however, only a limited number of studies on rating accuracy (e.g., Borman, 1979; Schleicher et al., 2002) have examined whether improvements in rating accuracy in a training context can translate into improvements in criterion validity. In a training context, higher criterion validity refers to an instance in which ratings from trained (vs. untrained) evaluators demonstrate stronger relationships with a criterion measure (Schleicher et al., 2002).

We predict that the restructured FOR training will achieve higher criterion validity than will typical FOR training because the former training method restructures the presentation of training information and offers opportunities to make sufficient adjustments. As we discussed previously, these procedures may mitigate the accessibility of self-anchors and increase the likelihood of making sufficient rating adjustments, which may increase rating accuracy. Given that criterion validity is an additional way to assess rating accuracy, we propose the following hypothesis:

> **Hypothesis 2.** *Restructured FOR training leads to higher criterion validity than does typical FOR training.*

5 | THE OVERVIEW OF THE STUDIES

On the basis of the anchoring and adjustment heuristic, we modified existing FOR training by restructuring information presentation to minimize the accessibility of self-rating anchors and by offering opportunities for rating adjustment. The purposes of the present research were to examine the differential effects of restructured versus typical FOR training on rating accuracy[1] (in Studies 1–5) and criterion validity (in Studies 3 and 5). We focused on examining the feedback delivery process during typical FOR training. Therefore, we included a restructured FOR training condition that involved restructuring the feedback delivery process of the typical FOR training and a control training condition without feedback from a training advisor. The current control training condition is similar to those in previous studies (e.g., Athey & McIntyre, 1987) in which participants read an explanation of the rating scales and evaluated practice targets.

In contrast to most of the existing research investigating the effectiveness of rater training, which typically focused on a single context (e.g., Cardy & Keefe, 1994; Uggerslev & Sulsky, 2008), we conducted five studies that varied in the rating context, rating format, hiring experience, and training environment (onsite vs. online) to strengthen the generalizability of our findings. Using a Type I error rate of 5%, a Type II error rate of 20%, and an effect size of d = 0.50 (the meta-analytic effect size reported by Roch et al., 2012), we conducted a power analysis that indicated a minimum of 51 participants per condition. In anticipation of incomplete responses, we aimed to achieve a total sample size of at least 180 for the studies with three conditions (Studies 1–3 and 5) and 300 for the study with five conditions (Study 4). Analyses were conducted only after the completion of data collection.

6 | STUDIES 1–3: EVALUATIONS BASED ON PERSONALITY TRAITS

The purpose of the first three studies was to examine the combined effects of restructured information presentation and rating adjustment on rating accuracy. In addition, Study 3 examined the criterion validity of performance ratings obtained from evaluators receiving restructured (vs. typical) FOR training.

6.1 | Participants and research design

Participants in Studies 1 and 2 were adults residing in the United States recruited using Amazon's Mechanical Turk (MTurk) website (see Behrend, Sharek, Meade, & Wiebe, 2011). They completed the

[1] We focused on rating accuracy rather than other indicators used in the rater training literature (e.g., inter-rater reliability and discriminant validity; Lievens, 2001) because the value of rating accuracy is higher than the value of other indicators. Rating accuracy measures the difference between raters' evaluations and established standards (Sulsky & Balzer, 1988), and therefore, higher rating accuracy can reflect higher inter-rater reliability. However, higher inter-rater reliability does not necessarily indicate higher rating accuracy; raters can have similar ratings that are different from the established standards. Previous research on rater training examined discriminant validity because raters may rely on only a referent schema to evaluate multiple dimensions of a candidate's performance, thus increasing the associations between different dimension ratings (Lievens, 2001). However, our study design involved only two dimensions, and each of the dimensions was associated with a specific definition and particular performance-relevant information. Therefore, our studies demonstrated low average associations of participants' ratings between the two dimensions within a specific candidate. In Studies 1–5, the average correlational coefficients of the dimensional ratings within each candidate ranged from r = .11 to r = .31 (see the average correlational coefficients across the different studies in Section I of the Supporting Information). Thus, the issue of low discriminant validity does not apply to our current studies.

study in exchange for monetary incentives ($0.8). Forty-five percent of participants in Study 1 (N = 180; 56% female; M_{age} = 35.74 years, SD_{age} = 11.71; $M_{\text{work experience}}$ = 15.53 years, $SD_{\text{work experience}}$ = 10.59; 92% currently employed) reported previous experience in hiring employees. Fifty-one percent of participants in Study 2 (N = 174[2]; 49% female; M_{age} = 36.78 years, SD_{age} = 12.01; $M_{\text{work experience}}$ = 15.76 years, $SD_{\text{work experience}}$ = 10.93; 89% currently employed) reported previous experience in hiring employees. Participants in Studies 1 and 2 worked in a wide range of industries, such as telecommunication, health care, education, and finance. Participants in Study 3 (N = 169[3]; 68% female; M_{age} = 21.59 years, SD_{age} = 1.53; $M_{\text{work experience}}$ = 1.86 years, $SD_{\text{work experience}}$ = 1.33) were undergraduate students from a university in Singapore completing the study for course credit. Studies 1–3 were conducted in 2014, 2015, and 2016, respectively. To ensure data quality, we adhered to these two best practice procedures (Peer, Vosgerau, & Acquisti, 2014; Zhou & Fishbach, 2016): (a) We only recruited participants with at least a 96% past approval rate (i.e., the percentage of studies completed by a participant that are approved by the study requesters) in MTurk studies, and (b) we asked participants to provide individualized information (i.e., an MTurk identification number in Studies 1–2 and an email address in Study 3). Using at least a 95% past approval rate can effectively screen against inattentive responders (Peer et al., 2014), and asking for individualized information motivates participants to complete a study (Zhou & Fishbach, 2016).

Each study used a between-subjects design. In Studies 1–3, participants were randomly assigned to one of the three conditions, including control training (N_{Study1} = 60; N_{Study2} = 62; N_{Study3} = 56), typical FOR training (N_{Study1} = 65; N_{Study2} = 57; N_{Study3} = 56), or restructured FOR training (N_{Study1} = 55; N_{Study2} = 55; N_{Study3} = 57). In an attempt to minimize the anchoring and adjustment heuristic, restructured FOR training in these studies included the elements of restructured information presentation and rating adjustment because this training was designed to decrease the accessibility of self-rating anchors and increase the likelihood of sufficient rating adjustments.

6.2 | Procedures and stimulus materials

Studies 1 and 2 were online studies using internet-based rater training (e.g., Aguinis et al., 2009), whereas Study 3 was a laboratory study using traditional onsite training sessions in which an advisor was able to answer questions from participants (e.g., Stamoulis & Hauenstein, 1993). The training advisor was a PhD candidate with expertise in organizational psychology.

Participants were first informed that their task was to evaluate candidates for a given job. They received a brief description of the rating scenario (i.e., a short job description that highlighted the relevant performance dimensions, see Section A in the Supporting Information). Participants evaluated sales representatives on planning and organization, and interpersonal relations in Study 1, teachers on appropriate planning and rapport with students in Study 2, and contract specialists on negotiation tactics and clear objectives in Study 3.

Then participants received rater training. All participants read the definitions and the example behaviors corresponding to the definitions (see examples in the Supporting Information: Section B). The definitions of the dimensions and the examples of behaviors were modified from materials used by staff members in the Indiana State Personnel Department (2015) for Study 1, Beebe (1980) for Study 2, and Numprasertchai and Swierczek (2006) for Study 3. They also answered one comprehension question for each of the performance dimensions to ensure that they understood the definition and the rating categories (see Section B in the Supporting Information). If participants answered the question incorrectly, they were asked the question again until they provided the correct answer.

Afterward, participants received different training procedures depending on the experimental condition to which they were randomly assigned. In the *control training condition*, participants rated training targets (see examples of job candidates in the Supporting Information: Section C) using a 10-point scale (1 = *lowest performance* and 10 = *highest performance*). They did not receive any further instructions or feedback. Each job candidate was described using two behavioral descriptions for each performance dimension. For example, in the rating scenario used in Study 1, the description of a job candidate on the planning and organization dimension might state that the candidate “perseveres until the task is finished” and “likes to think about different ideas” corresponding to a high level of performance on the planning and organization dimension. Participants rated the same training targets for both dimensions. The performance level on each dimension was systematically varied across job candidates (i.e., participants rated job candidates at different levels on each of the performance dimensions). Thus, an evaluated candidate could have a low level of performance on one dimension and a moderate level of performance on the other dimension.

In the *typical FOR training condition*, participants first rated the same job candidates as those in the control condition. Second, they received feedback from their training advisor. For each job candidate, the feedback included (a) the participant's ratings, (b) the advisor's ratings, (c) the differences between the participant's ratings and the advisor's ratings, and (d) the advisor's rationale for the given ratings (see sample feedback in the Supporting Information: Section D). The advisor's evaluation standards and rationales were developed based on an expert panel discussion, including three researchers with postgraduate education degrees in relevant fields. Third, participants received a summary of the differences between their ratings and the advisor's ratings for the candidates.

By contrast, participants first read the advisor's evaluation standards (i.e., the advisor's ratings—and reasons for those ratings—for hypothetical situations with different levels of performance; see Section E in the Supporting Information) in the *restructured FOR training condition*. The content of the advisor's evaluation standards and

[2] In Study 2, 180 adults were recruited to participate in the study and six of them were excluded from the final sample because they already participated in Study 1. To ensure high data quality in our research, we used completely independent samples in the five different studies.

[3] In Study 3, 172 adults were recruited to participate in the study, but three of them did not submit complete responses and therefore were excluded from the final sample.

rationales was the same in the restructured FOR condition and the typical FOR condition. Second, the participants rated the same candidates as in the other training conditions. In addition, participants were asked to "apply their advisor's rating standards" to evaluate candidates. If participants did not apply their advisor's ratings correctly, they were asked to adjust their ratings of the candidates until their ratings were the same as the advisor's ratings. Specifically, after providing an incorrect response, participants would be presented with the same candidate they previously rated and received a message indicating an inconsistency between their own rating and their advisor's rating. This message also included a request for a rating adjustment and a reevaluation of the same candidate.

After receiving the training, participants evaluated 12 job candidates (in random order; see sample candidates in the Supporting Information: Section F) using the previous 10-point scale (Study 1: planning and organization, *M* = 5.26, *SD* = 0.73, and interpersonal relations, *M* = 4.47, *SD* = 0.68; Study 2: appropriate planning, *M* = 5.27, *SD* = 0.58, and rapport with students, *M* = 4.84, *SD* = 0.64; Study 3: negotiation tactics, *M* = 5.22; *SD* = 0.77, and clear objectives, *M* = 5.08; *SD* = 0.59). This allowed for the computation of a measure of post-training rating accuracy. Then participants in the conditions of restructured and typical FOR training rated two statements regarding their training process, which served as manipulation check measures.

Afterward, participants in Study 3 engaged in a second rating task (see relevant information in the Supporting Information: Section G) where they used the previous scales of the two dimensions (i.e., negotiation tactics and clear objectives) to evaluate different individuals based on their negotiation processes. Specifically, participants read the transcripts (in random order) of two actual negotiations in Study 3, where each negotiation occurred between two people. Consistent with existing research on validity in negotiation behavior (e.g., an examination of a relationship between a developed scale and objective negotiation performance; Curhan, Elfenbein, & Xu, 2006), we correlated participants' ratings of performance dimensions with objective negotiation performance, which measured criterion-related validity. Finally, participants reported their demographics and read a debriefing message.

6.3 | Measures

6.3.1 | Manipulation check

Participants in the restructured and typical FOR training conditions rated two statements (1 = *strongly disagree*; 7 = *strongly agree*) that described the features of restructured and typical FOR training, respectively: "My goal was to apply my training advisor's ratings," and "My goal was to modify my original ratings based on my training advisor's feedback."

6.3.2 | Rating accuracy

Rating accuracy was measured using distance accuracy (Bernardin & Pence, 1980; McIntyre, Smith, & Hassett, 1984) and Borman's (1977) differential accuracy (see Gorman & Rentsch, 2009).[4] Distance accuracy provides a measure of *absolute discrepancy*; it indicates the average absolute difference of each participant's ratings on the focal dimensions of the targets (i.e., 12 job candidates during the post-training phase) from the training advisor's ratings (i.e., the referent schema). Lower scores of distance accuracy indicate higher rating accuracy. By contrast, Borman's differential accuracy provides a measure of a *relative discrepancy*; it reflects the correlations between participants' ratings and the training advisor's ratings. Higher scores of Borman's differential accuracy indicate higher rating accuracy. In order to indicate a specific type of rating accuracy, we aggregated multiple ratings (provided by a specific participant) into a single score based on the operational definition from existing research (Sulsky & Balzer, 1988).

6.3.3 | Performance ratings and objective negotiation performance

In Study 3, participants evaluated the performance of different negotiators by reading transcripts that involved these negotiators' interactions. These performance ratings were then compared with the negotiators' objective performance, which was measured using the number of points that each negotiator obtained during the negotiation. Participants rated each of four negotiators using two 10-point scales measuring "negotiation tactics" (*M* = 6.28, *SD* = 1.22) and "clear objectives" (*M* = 6.15, *SD* = 1.32); the four negotiators had an average of 4,000 points (*SD* = 952 points; range 3,000–5,000).

The negotiation processes were based on a separate laboratory study. Four undergraduate students were recruited and paired to form two dyads. Each dyad engaged in an online negotiation using Google Chat software. These recordings served as the negotiation transcripts read by participants in Study 3. In this negotiation case (modified from Dimotakis, Conlon, & Ilies, 2012), the participants served as representatives of two companies (Mountain Enterprises and Pinnacle Services) that were planning to merge into a single company. They negotiated on three issues (i.e., signing bonus, vacation time, and starting date) regarding this merged company. Each issue had five potential options, and each option was associated with a given number of points (see Section H in the Supporting Information). Points differed depending on the role (i.e., a representative from Mountain Enterprises or Pinnacle Services). For example, the Pinnacle Services representative would prefer to offer a large signing bonus, whereas the Mountain Enterprises representative would prefer not to offer any signing bonus for employees in the merged company. Thus, if the dyad agreed to a 10% signing bonus (the largest possible bonus), then the Pinnacle Services representative achieved 1,600 points, whereas the Mountain Enterprises representative achieved 0 point.

[4] We note that other measures of rating accuracy have also been used, such as Cronbach's (1955) squared distance component measures (see a review; Sulsky & Balzer, 1988). We did not use Cronbach's (1955) component measures because these indicators would be unduly affected by large discrepancies between a participant's rating and the referent rating. We chose distance accuracy because it could be easily interpreted in the same metric as the original rating scale. In addition, Borman's (1977) differential accuracy indicator was a complementary measurement reflecting relative discrepancy (i.e., a correlation).

Consistent with existing paradigms (e.g., Beersma & Dreu, 2005), negotiators were given monetary incentives to achieve a consensus on the issues and to maximize their individual benefit in the negotiation. Specifically, objective performance was determined based on whether negotiators could achieve an agreement with the preferred options for each of the three issues. Each option had a corresponding number of points for each representative. Negotiators were not allowed to share the points information with their counterpart during the negotiation process. The points information was used to compute a score of objective negotiation performance.

7 | RESULTS OF STUDIES 1–3

7.1 | Descriptive statistics

Table 2 presents a comparison of the study variables (i.e., rating accuracy) by training conditions for each of the five studies (see the mean levels of rating accuracy across different conditions in Figure 1). As described previously, lower scores of distance accuracy and higher scores of Borman's differential accuracy indicate higher levels of rating accuracy.

7.2 | Manipulation checks

The two manipulation check statements were significantly correlated with each other in Study 1 (r = .20, p < .05) and Study 2 (r = .26, p < .01), but not in Study 3 (p > .05). Analysis of covariance (ANCOVA) was therefore used to examine the effectiveness of the manipulations. That is, when the item describing restructured FOR training (i.e., "My goal was to apply my training advisor's ratings") was the dependent variable, the other item describing typical FOR training (i.e., "My goal was to modify my original ratings based on my training advisor's feedback") was used as the covariate. The training condition was the independent variable. Results demonstrated that the training manipulations of the restructured and typical FOR conditions were effective (all Fs $\geq$ 6.22, all ps < .05). Specifically, participants in the restructured FOR condition ($M_{\text{estimated}}$: 5.66 to 6.21) were more motivated to *apply* their advisor's ratings to the assessment task than participants in the typical FOR condition ($M_{\text{estimated}}$: 4.62 to 5.23). Participants in the typical FOR condition ($M_{\text{estimated}}$: 5.25 to 5.86) were also more motivated to *modify* their ratings based on their advisor's feedback than those in the restructured FOR condition ($M_{\text{estimated}}$: 4.24 to 4.60). These results supported a differentiation between the typical and restructured FOR training conditions.

7.3 | Tests of focal hypotheses

Hypothesis 1. *Training effects*

Hypothesis 1 states that rating accuracy would be higher in the restructured FOR training condition than in the typical FOR training condition. To examine Hypothesis 1, we used analyses of variance (ANOVAs) according to existing research (e.g., Melchers et al., 2011; Stamoulis & Hauenstein, 1993; Uggerslev & Sulsky, 2008). This analytical technique facilitates a comparison between the results in previous research and our research.

In Table 2, the results of the one-way ANOVAs indicated significant differences in rating accuracy across conditions in Studies 1–3 (all ps < .05 for the omnibus F test). Consistent with Hypothesis 1, planned comparisons between training conditions indicated significantly higher rating accuracy (i.e., lower scores of distance accuracy and higher scores of Borman's differential accuracy) in the restructured FOR training condition than in the typical FOR training condition. Specifically, rating accuracy was significantly higher in the restructured FOR (with adjustment) condition than in the typical FOR (no adjustment) condition (Distance accuracy$_{\text{[Studies 1–3]}}$: d = −0.65/−0.46/−0.84; t: −4.42 to −2.36; all ps < .05; Borman's differential accuracy$_{\text{[Studies 1–3]}}$: d = 0.62/0.65/0.75; t: 3.66 to 4.36; all ps < .001). Thus, the results provided consistent support for Hypothesis 1. In addition, rating accuracy was significantly higher in the restructured FOR (with adjustment) condition than in the control condition (Distance accuracy$_{\text{[Studies 1–3]}}$: d = −0.96/−0.81/−1.57; t: −8.82 to −5.00; all ps < .001; Borman's differential accuracy$_{\text{[Studies 1–3]}}$: d = 0.97/0.86/1.40; t: 5.24 to 8.02; all ps < .001).

Hypothesis 2. *Criterion validity of ratings*

Hypothesis 2 states that restructured FOR training leads to higher criterion validity than does typical FOR training. This hypothesis was examined in Study 3, using mixed-effects regression analyses conducted with maximum likelihood estimation (Rabe-Hesketh & Skrondal, 2008). We chose to use mixed-effects regression due to the nonindependence of the data. Specifically, ratings from a specific participant were more related to each other than to ratings from different participants (intra class correlation = .18, p < .001). Similarly, ratings from a specific negotiation dyad were more related to each other than to ratings from different negotiation dyads (intraclass correlation = .04, p < .001; see the full results in the Supporting Information: Section K), indicating that our data violated the assumption of independent observations required for using ordinary least squares regression analyses. Therefore, to control for the random effects resulting from the differences in participants and negotiation dyads, we included participant identification number and negotiation dyad number as random effect variables in the mixed-effects regression models. The dependent variable was the objective performance of the negotiators (i.e., points obtained in the negotiation); the independent variable was derived from participants' ratings (based on the negotiators' behaviors in the transcripts); and the moderator variable was the training condition.

The three training conditions were dummy coded to create two condition variables. Thus, to obtain pairwise comparisons among all three conditions, we ran the regression model twice. Given that the lower order coefficients in each of the regression models differ slightly, only the interaction effect results are presented in Table 3 (i.e., the interaction between ratings and dummy coded training

TABLE 2 Distance accuracy across training conditions

		Restructured FOR		Typical FOR		Control	*F*
Study		**No Adj.**	**Adj.**	**No Adj.**	**Adj.**		
Dependent Variable: Distance accuracy							
Study 1	*M*		0.65_{a}	1.09_{b}		1.34_{c}	16.67***
	SD		0.81	0.51		0.61	
Study 2	*M*		0.55_{a}	0.82_{b}		1.12_{c}	12.55***
	SD		0.74	0.38		0.66	
Study 3	*M*		0.48_{a}	1.01_{b}		1.55_{c}	38.88***
	SD		0.70	0.57		0.66	
Study 4	*M*	0.58_{a}	0.59_{a}	0.70_{ab}	0.73_{ab}	0.78_{b}	2.55*
	SD	0.40	0.44	0.44	0.48	0.42	
Study 5	*M*	0.68_{a}		0.80_{b}		0.85_{b}	6.87**
	SD	0.40		0.33		0.29	
Dependent Variable: Borman's differential accuracy							
Study 1	*M*		2.45_{a}	1.86_{b}		1.55_{c}	17.18***
	SD		1.16	0.67		0.60	
Study 2	*M*		2.60_{a}	2.04_{b}		1.82_{b}	14.34***
	SD		1.06	0.58		0.70	
Study 3	*M*		2.77_{a}	2.10_{b}		1.53_{c}	32.27***
	SD		1.08	0.68		0.64	
Study 4	*M*	1.72_{ac}	1.81_{c}	1.49_{ab}	1.44_{b}	1.31_{b}	4.46**
	SD	0.87	0.95	0.73	0.73	0.52	
Study 5	*M*	1.62_{a}		1.39_{b}		1.35_{b}	7.46***
	SD	0.76		0.51		0.34	
Average effect size			Distance accuracy		Borman's differential accuracy		
			d	95% CI	*d*	95% CI	
Restructured FOR versus typical FOR			−0.46	[−0.60, −0.32]	0.50	[0.35, 0.64]	
Restructured FOR versus control			−0.74	[−0.90, −0.59]	0.77	[0.62, 0.92]	
Typical FOR versus control			−0.36	[−0.51, −0.22]	0.35	[0.20, 0.49]	

Note. N_{Study1} = 180, N_{Study2} = 174, N_{Study3} = 169, N_{Study4} = 300, and N_{Study5} = 320. Smaller values indicate higher distance accuracy and lower Borman's differential accuracy. *F* indicates the one-way analysis of variance test for conditional differences. The different subscript letters (i.e., a, b, and c) indicate significantly different mean values ($p < .05$) within each study. When the correlational coefficients between ratings and true scores were equal to 1, they were replaced with 0.999 and then transformed into a Fisher's *z* score in order to resolve the issue of an infinite number in Fisher's *z* distribution.

Abbreviation: FOR, frame-of-reference.

**p* < .05. ***p* < .01. ****p* < .001.

conditions, obtained from two regression models).[5] To facilitate the interpretation of the interaction effect, we mean-centered each predictor variable before creating the interaction term (Aiken & West, 1991). As can be seen in Table 3 (where the result associated with the Ratings × Restructured vs. typical effect corresponds to the focal hypothesis test), the relationship between performance ratings and objective performance (i.e., criterion validity) was significantly moderated by training condition (B = 51.17, $p < .05$). That is, criterion validity in the restructured FOR training condition differed significantly from criterion validity in the typical FOR training condition. To probe the patterns of this interaction effect, we examined criterion validity for each training condition (i.e., a simple effect of performance rating on objective performance, at each level of training). We found that criterion validity was significantly more positive in the restructured FOR training condition (B = 117.21, $p < .001$) than in the typical FOR training condition (B = 66.04, $p < .001$). Thus, these results provided support for Hypothesis 2.

8 | STUDIES 4 AND 5: EVALUATIONS BASED ON BEHAVIORAL FREQUENCIES

The purpose of Study 4 was to explore the separate effects of restructured information presentation and rating adjustment on rating accuracy. Due to a possible, positive association between the amount of training practice and rating adjustment, we excluded rating adjustment as a predictor in Study 5. Thus, Study 5 was conducted to investigate the effects of restructured information presentation without

[5] The full regression models are presented in the Supporting Information: Section L.

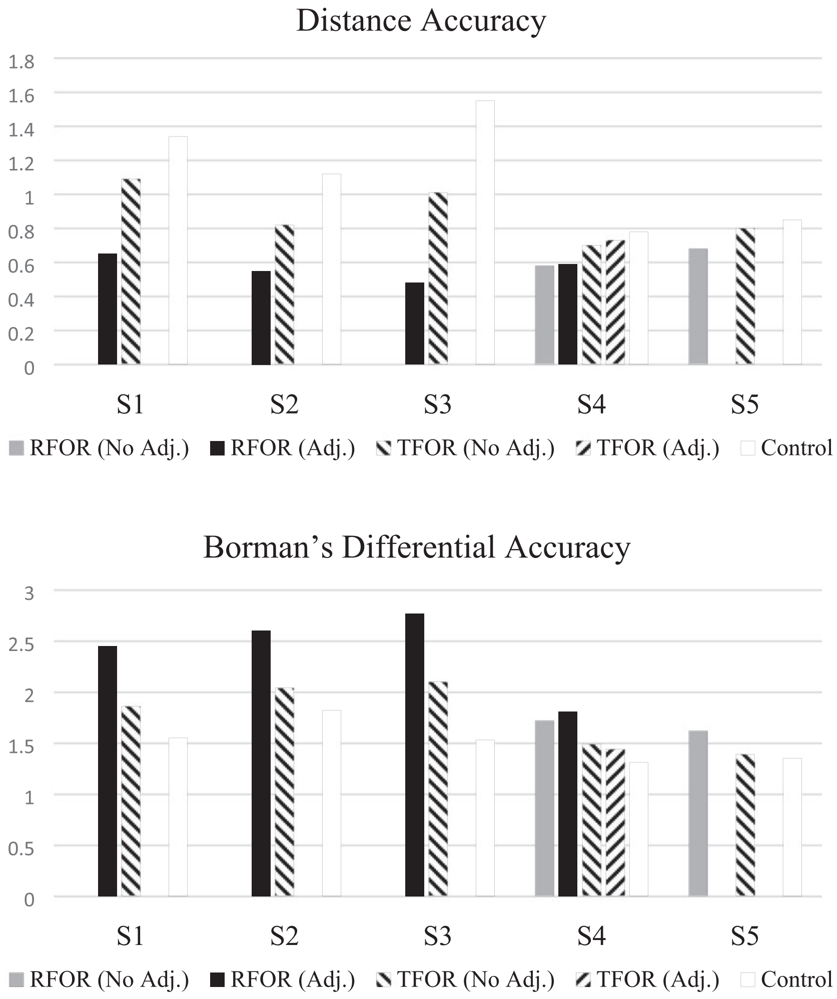

FIGURE 1 Rating accuracy across different conditions in each study. S1–S5 indicate Studies 1–5; RFOR/TFOR refers to restructured/typical frame-of-reference training; the term "Adj/No Adj" denotes the condition involving opportunities/no opportunities for rating adjustment

TABLE 3 Mixed-effects regression results for criterion validity of performance ratings

	Study 3			Study 5		
Criterion validity	***B***	***SE***	**Wald χ^2**	***B***	***SE***	**Wald χ^2**
Interaction effect						
Ratings × Restructured versus control	23.69	21.15		15.63	25.21	
Ratings × Typical versus control	−27.48	21.98		−36.14	24.78	
Ratings × Restructured versus typical	51.17*	20.37		51.77*	24.34	
			125.79***			12.07*
Simple effect						
Restructured	117.21***	13.30	77.63***	50.62**	17.48	8.38**
Typical	66.04***	15.38	18.43***	−1.14	16.93	0.00
Control	93.52***	16.19	33.36***	34.99	18.13	3.72

Note. Unstandardized regression coefficients are reported.

**p* < .05. ***p* < .01. ****p* < .001.

rating adjustment on rating accuracy and criterion validity. To ensure that our findings were generalizable across different rating formats, the information about the rating targets involved the frequencies of specific behaviors in Studies 4 and 5, which was different from the descriptions of personality traits used in Studies 1–3.

8.1 | Participants and research design

Participants in Study 4 (N = 300; 54% female; M_{age} = 37.81 years, SD_{age} = 11.54; $M_{\text{work experience}}$ = 16.89 years, $SD_{\text{work experience}}$ = 11.05; 84% currently employed) were recruited using the same method and procedures as in Studies 1 and 2. Participants were compensated US $1 in exchange for their participation. Fifty-three percent of these MTurk participants reported previous experience in hiring employees. Participants in Study 5 (*N* = 320; 71% female; M_{age} = 21.66 years, SD_{age} = 1.58; $M_{\text{work experience}}$ = 1.19 years, $SD_{\text{work experience}}$ = 1.26) were undergraduate students from a university in Singapore, and they completed the study for course credit or cash (SG$5). Studies 4 and 5 were conducted in 2017 and 2018, respectively.

We further improved data quality by using a motivational filter to determine our final samples in Studies 4 and 5 (see DeSimone, Harms, & DeSimone, 2015; Meade & Craig, 2012). The motivational filter was designed to indicate a level of effort and commitment regarding study completion. Specifically, participants were requested to answer a screening question before the study: Do you commit to providing your thoughtful and honest answers to the questions in this survey? For our data analyses, we used data only from participants who selected the option “I will provide my best answers.” If participants selected the option “I will not provide my best answers” or “I can't promise either way,” they were informed that they were not eligible to participate in the study (i.e., Study 4), or their data were excluded from our statistical analyses (i.e., two of the 322 participants in Study 5).

Study 4 involved a two (restructured presentation vs. typical presentation) by two (rating adjustment vs. no rating adjustment) design (*N* = 54–61 in each of the four conditions). In addition to the four conditions, Study 4 included a control training condition (*N* = 70). Study 5 included three conditions: control training (*N* = 109), typical FOR training (*N* = 105), and restructured FOR training (*N* = 106); in Study 5, the difference between the typical and restructured FOR training conditions involved only restructured information presentation.

8.2 | Procedures and stimulus materials

Study 4 utilized internet-based rater training, whereas Study 5 used traditional onsite training sessions. The basic procedures in Studies 4 and 5 were the same as those in Studies 1–3. The training materials in Studies 4 and 5 are also included in the corresponding sections in the Supporting Information. Participants evaluated retail store managers on active listening and critical thinking in Study 4 and contract specialists on negotiation tactics and clear communication of objectives in Study 5. The definitions of the dimensions and the examples of behaviors were modified from a job description of a retail sales worker used on the website “O-Net Online” (2017) for Study 4 and the dimensions used by Numprasertchai and Swierczek (2006) for Study 5.

In Studies 4 and 5, the behavioral description involved the frequency of a specific behavior, such as “sometimes understands others' perspectives” and “rarely elicits information at appropriate times during a conversation.” Participants in Study 4 rated different training targets for different dimensions, whereas those in Study 5 rated the same training targets for both dimensions. These differences between Studies 4 and 5 allow for increased generalizability of our study results. The advisor's evaluation standards and rationales were developed based on an expert panel discussion. In addition, all the scale points were further differentiated based on the frequency of a specific behavior.

Study 4 participants in the restructured FOR training condition were requested to adjust their ratings only when they were assigned to the two conditions with rating adjustment. In Study 5, there was no request for rating adjustment. After receiving the training, participants evaluated multiple job candidates (*N* = 12 in Study 4 and *N* = 13 in Study 5, in random order) using a 7-point scale of the performance dimensions (Study 4: active listening, *M* = 3.98, *SD* = 0.39, and critical thinking, *M* = 4.01, *SD* = 0.39; Study 5: negotiation tactics, *M* = 4.02, *SD* = 0.41, and clear communication of objectives, *M* = 3.92, *SD* = 0.41). In addition to the manipulation check statements used in Studies 1–3, participants in Studies 4 and 5 completed other manipulation check items to indicate their training process before the formal assessment task in the restructured and typical FOR conditions.

Afterward, participants in Study 5 engaged in a second rating task where they used the previous scales of the two dimensions (i.e., negotiation tactics and clear communication of objectives in Study 5) to evaluate different individuals based on their negotiation processes. In Study 5, participants read the same transcripts (in random order) as those in Study 3. In these two transcripts, the messages regarding the final negotiation agreements were blocked in order to avoid a situation in which participants would infer each negotiator's performance based on the final agreements.[6]

8.3 | Measures

8.3.1 | Manipulation check

In addition to the two manipulation check statements used in Studies 1–3, participants receiving restructured and typical FOR training in Studies 4 and 5 rated other manipulation check statements. In Study 4, they also rated the following three statements (1 = *definitely false*; 7 = *definitely true*): (a) “I read the advisor's rating standards before the practice trials and was requested to apply the advisor's rating standards to the subsequent assessment task,” (b) “I read the advisor's rating standards only after the practice trials and was requested to modify my ratings in the subsequent assessment task,” and (c) “During the practice trials, I adjusted my ratings until my ratings were consistent with my advisor's ratings.” In Study 5, they also rated the statements (a) and (b) but with the modification of “the subsequent assessment task” to “the subsequent formal assessment task.”

8.3.2 | Rating accuracy

The distance accuracy and Borman's differential accuracy indicators computed in Studies 1–3 were also used in Studies 4 and 5.

[6] We conducted a separate study with a manipulation of revealing or blocking the final outcomes in the negotiation transcripts and found that revealing or blocking the final outcomes in the negotiation transcripts did not influence participants' ratings or criterion validity (see the study in Section J of the Supporting Information).

8.3.3 | Performance ratings and objective negotiation performance

In Study 5, participants rated each of four negotiators using two 7-point scales measuring "negotiation tactics" (M = 4.82, SD = 0.77) and "clear communication of objectives" (M = 4.91, SD = 0.81). The same method as in Study 3 (i.e., the associations between participants' ratings and objective negotiation performance) was used to examine criterion validity in Study 5.

9 | RESULTS OF STUDIES 4 AND 5

9.1 | Manipulation checks

The same method as in Studies 1–3 (i.e., ANCOVA) was used to examine the effectiveness of the manipulations. Results demonstrated that the training manipulations of the restructured and typical FOR conditions were effective (all Fs $\geq$ 5.41, all ps < .05). Specifically, participants in the restructured FOR condition(s) ($M_{\text{estimated}}$: 5.63 to 6.23) were more motivated to *apply* their advisor's ratings to the assessment task than those in the typical FOR condition(s) ($M_{\text{estimated}}$: 4.90 to 5.83). Participants in the typical FOR condition(s) ($M_{\text{estimated}}$: 4.38 to 5.40) were more motivated to *modify* their ratings based on their advisor's feedback than those in the restructured FOR condition(s) ($M_{\text{estimated}}$: 3.51 to 4.13).

We also used the same analyses (i.e., ANCOVAs)[7] to demonstrate the results of the other two manipulation check statements in Studies 4 and 5. Results demonstrated that the training manipulations of the restructured and typical FOR conditions were effective (all Fs $\geq$ 15.53, all ps < .001). Participants in the restructured FOR condition(s) ($M_{\text{estimated}}$: 4.69 to 5.66) were more likely to indicate that they read the advisor's rating standards before the practice trials and were requested to apply the advisor's rating standards to the subsequent assessment task than those in the typical FOR condition(s) ($M_{\text{estimated}}$: 3.27 to 4.50). By contrast, participants in the typical FOR condition(s) ($M_{\text{estimated}}$: 4.33 to 4.53) were more likely to indicate that they read the advisor's rating standards only after the practice trials and were requested to modify their ratings in the subsequent assessment task than those in the restructured FOR condition(s) ($M_{\text{estimated}}$: 2.61 to 2.81). In Study 4, participants in the adjustment conditions ($M_{\text{estimated}}$ = 5.43) were more likely to indicate that they adjusted their ratings until their ratings were consistent with their advisor's ratings than those in the no adjustment conditions ($M_{\text{estimated}}$ = 3.41; F = 52.64, p < .001). These results supported differentiation among our training conditions.

9.2 | Tests of focal hypotheses

Hypothesis 1. *Training effects*

In Table 2, the results of the one-way ANOVAs indicated significant differences in rating accuracy across conditions in Studies 4 and 5 (all ps < .05 for the omnibus F test). In Study 4, where restructured and typical FOR training were further differentiated based on whether or not rating adjustment was allowed, the planned comparison between both the restructured FOR training conditions (i.e., with adjustment and no adjustment) and both the typical FOR training conditions (i.e., with adjustment and no adjustment) indicated significantly higher rating accuracy for the restructured than for the typical FOR training conditions (Distance accuracy: d = −0.30, t[297] = −2.30, p < .05; Borman's differential accuracy: d = 0.36, t[297] = 2.99, p < .01). To examine separate effects of restructured information presentation and rating adjustment, we conducted a two-way ANOVA (i.e., restructured information presentation: restructured presentation vs. typical presentation; rating adjustment: with adjustment vs. without adjustment). The results indicated only a significant effect of restructured information presentation (Distance accuracy: F[1, 226] = 5.16, p < .05; Borman's differential accuracy: F[1, 226] = 7.74, p < .01).[8] Neither the main effect of adjustment (Distance accuracy: F[1, 226] = 0.20, p = .658; Borman's differential accuracy: F[1, 226] = 0.04, p = .845) nor the interaction effect (Distance accuracy: F[1, 226] = 0.03, p = .862; Borman's differential accuracy: F[1, 226] = 0.40, p = .526) was statistically significant. In Study 5, where both restructured and typical FOR training sessions were conducted without adjustment, rating accuracy was significantly higher in the restructured FOR training condition than in the typical FOR training condition (Distance accuracy: d = −0.33, t[317] = −2.59, p < .001; Borman's differential accuracy: d = 0.36, t[177] = 3.01, p < .01). Thus, the results consistently supported Hypothesis 1 regarding the positive impact of a restructured presentation of training information rather than rating adjustments on rating accuracy. In addition, rating accuracy was significantly higher in the restructured FOR condition(s) than in the control condition (Distance accuracy$_{\text{[Studies 4 and 5]}}$: d = −0.47/−0.48; t = −2.97/−3.60; both ps < .01; Borman's differential accuracy$_{\text{[Studies 4 and 5]}}$: d = 0.62/0.47; t = 3.94/3.60; both ps < .001).[9]

Hypothesis 2. *Criterion validity of ratings*

The same method as in Study 3 was used to examine criterion validity of ratings in Study 5. In Table 3, the relationship between performance ratings and objective performance (i.e., criterion validity) was significantly moderated by training condition (i.e., restructured FOR training vs. typical FOR training, B = 51.77, p < .05). Specifically, the results of criterion validity were significantly more positive in the

[7] When one of the manipulation check statements was used as the dependent variable, the other two statements were used as covariates in Study 4, whereas the other one statement was used as a covariate in Study 5.

[8] To examine whether the length of time between typical and restructured FOR training was comparable, we used the data available for the length of the entire survey time in the online training studies (i.e., Studies 1, 2, and 4). We did not find any significant differences in the length of the survey time between typical and restructured FOR training in these studies (all |t|s $\leq$ 0.72, all ps $\geq$ .48), which suggests a comparable amount of time between restructured and typical FOR training conditions.

[9] We explored the possibilities of estimating the results using mixed-effects regression for Hypothesis 1. We found consistent, significant patterns between the results of ANOVAs and mixed-effects regression for distance accuracy. However, we failed to produce an accurate correlational coefficient for each evaluation target due to a small sample size. Thus, we cannot use a mixed-effects regression model to estimate the results for Borman's differential accuracy. Please see our detailed discussion in the Supporting Information: Section I.

restructured FOR training condition ($B = 50.62$, $p < .01$) than in the typical FOR training condition ($B = -1.14$, $p = .946$), which supported Hypothesis 2.

10 | GENERAL DISCUSSION

FOR training is a frequently used method for training raters to make accurate subjective evaluations. However, the practice-then-feedback procedure on which it is based assumes that raters can effectively replace their idiosyncratic schema of job performance with the referent schema of performance espoused by the organization. In order to enhance the understanding of how cognitive heuristics may influence the schema replacement process in a subjective evaluation (i.e., ratings) situation, we conducted empirical studies to examine the influence of a cognitive heuristic. Specifically, we restructured the typical FOR training procedure to mitigate the effects of the anchoring and adjustment heuristic. Our results clearly demonstrate that rating accuracy could be improved through the novel and cost-effective method, even beyond the effectiveness of typical FOR training. To elaborate, the results of our first three studies demonstrated that rating accuracy was higher after the restructured FOR training session involving both restructured information presentation and rating adjustment than after the typical FOR session or the control training session. In Study 4, we found that restructured information presentation rather than rating adjustment of restructured FOR training significantly improved rating accuracy. In Study 5, we replicated the findings in Study 4 by demonstrating the positive effects of restructured information presentation without adjustment on rating accuracy.

Our results supported that the *overall* effectiveness of restructured FOR training exceeded that of typical FOR training (see Table 2). The average *d*s for restructured FOR versus control training on rating accuracy were more than twice as large as the average *d*s for typical FOR versus control training.[10] Specifically, comparing restructured FOR training with typical FOR training, distance accuracy was 2.06 times as large (i.e., −0.74/−0.36), and Borman's differential accuracy was 2.20 times as large (i.e., 0.77/0.35). Further, the average *d*s (for typical FOR vs. control training) on rating accuracy (Distance accuracy: $d = -0.36$; Borman's differential accuracy: $d = 0.35$) were of a similar magnitude to those reported in a recent meta-analysis (i.e., *d* from 0.20 to 0.51; Roch et al., 2012). In addition, we contribute to the limited number of rater training studies that examined criterion validity; Studies 3 and 5 established that restructured FOR training led to higher criterion validity than did typical FOR training. Overall, the results supported a novel method that can improve rating accuracy and enhance criterion validity by restructuring the existing FOR training method.

10.1 | Theoretical implications

Our work advances understanding of how training procedures can influence the impact of heuristics on subjective evaluations. As we noted earlier, previous research on FOR training focused on a schema-based theory and developed interventions to replace a personal schema of performance with an organizational schema of performance. For example, consistent with the schema-based theory, Melchers et al. (2011) proposed that behaviorally anchored rating scales reduce rater idiosyncrasies by offering a common evaluation standard during rating processes. However, other researchers have found that raters can only pay attention to information that is consistent with their personal schemas (e.g., Kulik, 1989), thus implying the suboptimal effectiveness of behaviorally anchored rating scales. Similar to Kulik's (1989) findings, our results also show that raters have difficulty in focusing on more than their personal schemas.

Furthermore, our results extend this previous research by offering one plausible explanation for the causes of limited attention. That is, our current study demonstrates that a narrow focus on personal schemas occurs—at least in part—because raters are susceptible to the anchoring effect (i.e., accessibility of a self-generated anchor hinders rating accuracy and criterion validity in the context of rater training). Other researchers have also proposed accessibility as the main reason for an adoption of an anchor (Mussweiler & Strack, 1999). Moreover, we provide a novel way to restrict the accessibility of a self-anchor by restructuring the presentation of training information in which participants are exposed to an advisor's evaluation standards before practice trials. Existing research also limits the accessibility of a self-anchor to lessen the anchoring and adjustment heuristic, such as priming other-related thoughts and emphasizing the inappropriateness of the self-anchor (Naylor et al., 2011). Therefore, our research expands the theoretical foundation of rater training by considering an important mental shortcut: the anchoring and adjustment heuristic.

Our research also complements the perspectives of heuristics utilization in organizational management. Business organizations are motivated to use heuristics due to the uncertain market environment and the pressure to respond spontaneously (Gigerenzer & Gaissmaier, 2010). Researchers have regarded a reliance on heuristics as an adaptive strategy in organizational decision-making processes (Artinger, Petersen, Gigerenzer, & Weibler, 2015). For example, experienced managers who used heuristics (i.e., a simple decision rule that only considers part of the available information) outperformed a complex, computational model (involving multiple parameters) in identifying repeat customers (Wübben & von Wangenheim, 2008). Nonetheless, our research suggests that the use of heuristics does *not always* provide an adaptive strategy. That is, raters who rely on the anchoring and adjustment heuristic provide inferior evaluation outcomes due to a difficulty in learning and applying new information. Future research efforts could examine the boundary conditions for when and how heuristics utilization can enhance or reduce the likelihood of generating accurate estimation outcomes.

Our research also contributes to an understanding of the practice-then-feedback approach. The typical FOR method may assume that trained raters can fully adopt the reference standards through the practice-then-feedback approach. This approach is consistent with the notion that immediate feedback can reinforce correct responses and eliminate incorrect responses (Skinner, 1954). However, some

[10] To estimate the effects of restructured FOR consistently, we focused on the manipulations involving restructured information presentation across the five studies.

research has demonstrated that the typical FOR training is ineffective (e.g., A. B. Nickerson & Nagle, 2001) and that the initial incorrect responses will interfere with the learning of correct responses from immediate feedback (Kulhavy, 1977). The opportunities for rating adjustment in our restructured FOR training also involved immediate feedback and were not effective at improving rating accuracy. One disadvantage of an answer-until-correct feedback approach during the rating adjustment process is the potential for raters to produce more than one incorrect response before they discover the correct rating (Butler, Karpicke, & Roediger, 2007). Relatedly, existing research has demonstrated that a selection of incorrect options on a multiple-choice test leads individuals to acquire and retain these incorrect responses on subsequent memory tests (Roediger & Marsh, 2005). Thus, our findings demonstrate the shortcomings of the practice-then-feedback approach and highlight the importance of a new training procedure—a presentation of the exemplary responses *before* rather than after individuals have developed their self-anchors for evaluation tasks.

Our research also illuminates the impact of feedback on work performance outcomes. For instance, research has found that when restaurant staff members are given corrective feedback on how to improve their customer service, their service quality remains similar (Waldersee & Luthans, 1994). Our research suggests that these employees may have already formed a self-anchor regarding how they should serve their customers, which leads to the ineffectiveness of the corrective feedback. Furthermore, research has focused on investigating ideal characteristics of feedback on employees' work performance, such as a delivery of timely, nonpersonal, accurate, and evidence-based feedback (Mayfield & Mayfield, 2011). However, this approach neglects an important psychological process: Employees may have already anchored themselves on their preferred way to handle various issues in the workplace. To address this limitation, our research identifies employees' original preferences as potentially powerful anchors and offers a different perspective on the effectiveness of feedback.

10.2 | Practical implications

Our restructured FOR method offers practical implications not only in rating settings but also in other training settings. Our restructured FOR method increases rating accuracy and criterion validity more significantly than does the typical FOR method. Business leaders and practitioners can use this restructured method to improve the quality of ratings. Furthermore, the same basic method (i.e., having trainees emulate a role model rather than modify their own behavior based on a trainer's feedback) of restructured FOR training can also be applied to other types of training, such as teamwork skills (Hughes et al., 2016), news delivery (Richter, König, Koppermann, & Schilling, 2016), managerial decision-making (Goodman & Wood, 2004), and principles regarding effective and ineffective teamwork (Smith-Jentsch, Campbell, Milanovich, & Reynolds, 2001).

Our findings also offer useful suggestions on the important features of a training program. Given the significant effects of a restructured information presentation rather than an opportunity for sufficient adjustment, our findings implicate that the feedback delivery process may be ineffective due to an interference of inconsistent information (Kulhavy, 1977). Therefore, managers should give feedback on organizational members' incorrect responses at an appropriate timing to avoid an interference of inconsistent information. Trainers should also eliminate the opportunities for learning incorrect responses at the initial stage of a program to mitigate the formation of self-anchors. More importantly, managers should understand the importance of initial responses and allow organizational members to emulate a benchmark as the first step of a training process.

Given that our new solutions to rater training are developed based on the anchoring and adjustment heuristic, effective interventions in previous research may also offer practical suggestions on how to improve the accuracy of judgment and decision-making in organizations. For instance, organizational decision-makers should be requested to think critically about an anchor and to identify reasons for dissimilarity between the anchor and the target estimate (Chapman & Johnson, 1994). Organizations can also establish a real-time feedback system to give a warning to managers or organizational members who participate in an evaluation process when their judgments are very close to an initial anchor (George, Duffy, & Ahuja, 2000). In addition, organizational managers and members should be educated to recognize the anchoring and adjustment heuristic in their judgments and modify their behavior accordingly to avoid anchoring (Gick & Holyoak, 1980). We hope that these de-biasing practices can help organizational decision-makers to achieve a high accuracy of evaluation outcomes.

10.3 | Limitations and future research

Although our research provides a novel way to improve rating accuracy, some limitations must be acknowledged and addressed through future research efforts. For instance, the training advisor for the onsite studies was a PhD candidate with expertise in organizational psychology rather than an experienced, professional trainer. The PhD candidate might also have little experience in delivering accurate feedback, and therefore, we took the following precautions: We predetermined the content of the feedback based on a panel of experts, preprogrammed, and standardized the delivery of feedback. Our investigation thus constituted a conservative examination of rater training effectiveness. Future research could use an expert in relevant fields as a training advisor to replicate and extend the findings in the present investigation.

Moreover, even though we investigated five independent samples (three of which involved working adults with hiring experience) with different rating scenarios, we recognize that performance evaluation typically occurs in circumstances where raters have more information about the ratees than mere behavioral descriptions of the ratees. Nevertheless, our investigation served as an internally valid examination given the effectiveness of the restructured FOR training method across different studies. Studies 3 and 5 also included evaluations based on negotiation processes in addition to behavioral descriptions. Future research could continue to examine the effects of the restructured FOR training on rating accuracy in organizational settings or situations that resemble evaluation processes in organizations (e.g., video vignettes of an interview). In addition, future research can control the length of training time across different conditions by adding filter tasks to eliminate the possible influence of training time on the findings.

We also derive several lines of future research from the current research. Researchers could examine moderators that can distinguish the effects of the restructured and typical FOR training in future studies. A rater's ability may also moderate the training effectiveness of the two FOR methods. Research demonstrated that low performers were not receptive to concrete feedback by discrediting either the accuracy or the relevance of the feedback whereas top performers were motivated to improve themselves based on the feedback (Sheldon, Dunning, & Ames, 2014). This suggests that restructured FOR training may be more beneficial for low performers than for top performers. In addition, goal orientation may moderate the differential association between the two methods and rating quality. People with a learning goal orientation tend to seek feedback to improve themselves, whereas people with a performance goal orientation avoid negative feedback to preserve a positive image in front of others (VandeWalle, Cron, & Slocum, 2001). In this case, a performance rather than learning goal orientation may enhance the differences in rating quality between these two training methods. Thus, researchers can investigate raters' ability and goal orientation as potential moderators of how the two methods affect the quality of ratings.

In addition, researchers can investigate the mediating processes of why the restructured FOR training method leads to a higher quality of rating outcomes than does the typical FOR training method. For instance, the typical FOR method involves raters' initial judgment during the training process. This judgment may reduce the effectiveness of the typical FOR training because raters' initial judgment may interfere with subsequent assessments (Athey & McIntyre, 1987). Furthermore, competition among several associated responses to a rating target may interfere with the ability to learn the correct response (Anderson, 1974). In addition to thought interference, negative emotional responses may serve as another potential mediator. Research has found that negative feedback elicits trainees' negative emotions and therefore increases turnover intention (Belschak & Den Hartog, 2009). This research suggests that when feedback is inconsistent with one's preference, the individual may resist the feedback due to negative emotions. This suggestion may also explain why rating adjustment in restructured FOR training did not significantly increase rating accuracy. Given that the typical FOR training emphasizes corrective feedback from an advisor, this method may elicit stronger negative emotions than the restructured FOR method without rating adjustment. Consequently, the former method leads to lower training effectiveness than does the latter method. Future research can focus on these potential mediators.

Moreover, future research can examine how different types of FOR training influence outcomes in nontraining contexts. Past research has found the effectiveness of typical FOR in a nontraining situation, such as coordination between different teams (Firth et al., 2015). Restructured FOR training may also outperform typical FOR training regarding outcomes in nontraining contexts. For instance, the restructured information presentation facilitates an adoption of a common evaluation standard without an effortful process of feedback and correction, which may conserve organizational leaders' energy to regulate their behavior and subsequently increase an enactment of interactional justice between these leaders and their followers (Whiteside & Barclay, 2016). Compared with typical FOR training, restructured FOR training may also significantly increase perceptions of procedural justice regarding performance appraisal through the increased accessibility of the information about the appraisal criteria (Erdogan, Kraimer, & Liden, 2001). Therefore, rater training can impact a wide array of behaviors in organizations.

11 | CONCLUSION

Our work underscores the impact of heuristic judgments on an evaluation process and demonstrates a novel and cost-effective method for increasing rating accuracy. We emphasize the benefit of restructuring typical FOR training to minimize the anchoring and adjustment heuristic. This new intervention significantly improves rating accuracy through a presentation of an expert's responses to possible rating scenarios *before* rather than after practice trials. Furthermore, restructured FOR training achieves higher criterion validity (i.e., ratings regarding a target are more strongly associated with the target's objective performance) than does typical FOR training. Our findings also offer practical implications on how to improve the effectiveness of diverse training programs and the accuracy of judgment in organizations. We also provide future research directions on rater training. We hope that our work can facilitate the reduction of biased and subjective assessments in organizational decision-making processes.

REFERENCES

Abbas, M. Z. (2014). Effectiveness of performance appraisal on performance of employees. *IOSR Journal of Business and Management*, *16*(6), 173–178. https://doi.org/10.9790/487X-1662173178

Aguinis, H., Mazurkiewicz, M. D., & Heggestad, E. D. (2009). Using web-based frame-of-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology*, *62*(2), 405–438. https://doi.org/10.1111/j.1744-6570.2009.01144.x

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA, US: Sage Publications, Inc. PsycINFO (1991-97932-000)

American Psychological Association, National Council on Measurement in Education, & American Educational Research Association (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ames, D. R. (2004). Inside the mind reader's tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology*, *87*(3), 340–353. https://doi.org/10.1037/0022-3514.87.3.340

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451–474. https://doi.org/10.1016/0010-0285(74)90021-8

Angkaw, A. C., Tran, G. Q., & Haaga, D. A. F. (2006). Effects of training intensity on observers' ratings of anxiety, social skills, and alcohol-specific coping skills. *Behaviour Research and Therapy*, *44*(4), 533–544. https://doi.org/10.1016/j.brat.2005.04.002

Artinger, F., Petersen, M., Gigerenzer, G., & Weibler, J. (2015). Heuristics as adaptive decision strategies in management. *Journal of Organizational Behavior*, *36*(S1), S33–S52. https://doi.org/10.1002/job.1950

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, *72*(4), 567–572. https://doi.org/10.1037/0021-9010.72.4.567

Beebe, R. J. (1980). Use of the behaviorally anchored rating scale in evaluating teacher performance. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno = ED194567

Beersma, B., & De Dreu, C. K. W. (2005). Conflict's consequences: Effects of social motives on postnegotiation creative and convergent group functioning and performance. *Journal of Personality and Social Psychology*, *89*(3), 358–374. WOS:000233092000007

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800. https://doi.org/10.3758/s13428-011-0081-0

Belschak, F. D., & Den Hartog, D. N. (2009). Consequences of positive and negative feedback: The impact on emotions and extra-role behaviors. *Applied Psychology: An International Review*, *58*(2), 274–303. https://doi.org/10.1111/j.1464-0597.2008.00336.x

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6) (pp. 1–62). San Diego, CA, Academic Press. https://doi.org/10.1016/S0065-2601(08)60024-6

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, *6*(2), 205–212. Retrieved from bah

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, *65*(1), 60–66. https://doi.org/10.1037/0021-9010.65.1.60

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance*, *20*(2), 238–252. https://doi.org/10.1016/0030-5073(77)90004-6

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, *64*(4), 410–421. https://doi.org/10.1037/0021-9010.64.4.410

Butler, A. C., Karpicke, J. D., & Roediger, H. L. I. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Cardy, R. L., & Keefe, T. J. (1994). Observational purpose and evaluative articulation in frame-of-reference training: The effects of alternative processing modes on rating accuracy. *Organizational Behavior and Human Decision Processes*, *57*(3), 338–357. https://doi.org/10.1006/obhd.1994.1019

Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making*, *7*(4), 223–242. https://doi.org/10.1002/bdm.3960070402

Cronbach, L. (1955). Processes affecting scores on “understanding of others” and “assumed similarity”. *Psychological Bulletin*, *52*(3), 177–193. https://doi.org/10.1037/h0044919

Curhan, J. R., Elfenbein, H. A., & Xu, H. (2006). What do people value when they negotiate? Mapping the domain of subjective value in negotiation. *Journal of Personality and Social Psychology*, *91*(3), 493–512. https://doi.org/10.1037/0022-3514.91.3.493

Dane, E., & Pratt, M. G. (2007). Exploring intuition and its role in managerial decision making. *Academy of Management Review*, *32*(1), 33–54. WOS:000243182200003

Danziger, S., Montal, R., & Barkan, R. (2012). Idealistic advice and pragmatic choice: A psychological distance account. *Journal of Personality and Social Psychology*, *102*(6), 1105–1117. https://doi.org/10.1037/a0027013

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171–181. https://doi.org/10.1002/job.1962

Dimotakis, N., Conlon, D. E., & Ilies, R. (2012). The mind and heart (literally) of the negotiator: Personality and contextual determinants of experiential reactions and economic outcomes in negotiation. *Journal of Applied Psychology*, *97*(1), 183–193. https://doi.org/10.1037/a0025706

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification systems. *Annual Review of Psychology*, *30*(1), 477. Retrieved from a9h

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*(5), 391–396. https://doi.org/10.1111/1467-9280.00372

Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Personality and Social Psychology Bulletin*, *30*(4), 447–460. https://doi.org/10.1177/0146167203261889

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Erdogan, B., Kraimer, M. L., & Liden, R. C. (2001). Procedural justice as a two-dimensional construct: An examination in the performance appraisal context. *The Journal of Applied Behavioral Science*, *37*(2), 205–222. https://doi.org/10.1177/0021886301372004

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence and maximal adaptation to task constraints. *Annual Review of Psychology*, *47*(1), 273. Retrieved from a9h

Firth, B. M., Hollenbeck, J. R., Miles, J. E., Ilgen, D. R., & Barnes, C. M. (2015). Same page, different books: Extending representational gaps theory to enhance performance in multiteam systems. *Academy of Management Journal*, *58*(3), 813–835.

George, J. F., Duffy, K., & Ahuja, M. (2000). Countering the anchoring and adjustment bias with decision support systems. *Decision Support Systems*, *29*(2), 195–206. https://doi.org/10.1016/S0167-9236(00)00074-9

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

Gigerenzer, G., & Gaissmaier, W. (2010). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Goodman, J. S., & Wood, R. E. (2004). Feedback specificity, learning opportunities, and learning. *Journal of Applied Psychology*, *89*(5), 809–821. https://doi.org/10.1037/0021-9010.89.5.809

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of*

Applied Psychology, *94*(5), 1336–1344. https://doi.org/10.1037/a0016476

Gunia, B. C., Swaab, R. I., Sivanathan, N., & Galinsky, A. D. (2013). The remarkable robustness of the first-offer effect: Across culture, power, and issues. *Personality and Social Psychology Bulletin*, *39*(12), 1547–1558. https://doi.org/10.1177/0146167213499236

Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., Marlow, S. L., Lacerenza, C. N., … Salas, E. (2016). Saving lives: A meta-analysis of team training in healthcare. *Journal of Applied Psychology.*. https://doi.org/10.1037/apl0000120

Indiana State Personnel Department. (2015). Behaviorally anchored ratings scale. Retrieved from Retrieved from www.in.gov/spd/files/bars.doc

Kataoka, H. C., Latham, G. P., & Whyte, G. (1997). The relative resistance of the situational, patterned behavior, and conventional structured interviews to anchoring effects. *Human Performance*, *10*(1), 47–63. https://doi.org/10.1207/s15327043hup1001_3

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, *47*(2), 211–232. https://doi.org/10.3102/00346543047002211

Kulik, C. T. (1989). The effects of job categorization on judgments of the motivating potential of jobs. *Administrative Science Quarterly*, *34*(1), 68–90. Retrieved from bsu

Kumar, D. (2005). Performance appraisal: The importance of rater training. *Journal of the Kuala Lumpur Royal Malaysia Police College*, *4*, 1–17.

Lehner, P., Seyed-Solorforough, M.-M., O'Connor, M. F., Sak, S., & Mullin, T. (1997). Cognitive biases and time stress in team decision making. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *27*(5), 698–703. https://doi.org/10.1109/3468.618269

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*(2), 255–264. https://doi.org/10.1037/0021-9010.86.2.255

Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology*, *92*(3), 812–819. https://doi.org/10.1037/0021-9010.92.3.812

Marshall, S. P. (1995). *Schemas in problem solving*. New York, NY: Cambridge University Press.

Mayfield, M., & Mayfield, J. (2011). Effective performance feedback for learning in organizations and organizational learning. *Development and Learning in Organizations: An International Journal*, *26*(1), 15–18. https://doi.org/10.1108/14777281211189128

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, *69*(1), 147–156. https://doi.org/10.1037/0021-9010.69.1.147

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Melchers, K. G., Lienhardt, N., Von Aarburg, M., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. *Personnel Psychology*, *64*(1), 53–87. https://doi.org/10.1111/j.1744-6570.2010.01202.x

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*(2), 136–164. https://doi.org/10.1006/jesp.1998.1364

Naylor, R. W., Lamberton, C. P., & Norton, D. A. (2011). Seeing ourselves in others: Reviewer ambiguity, egocentric anchoring, and persuasion. *Journal of Marketing Research*, *48*(3), 617–631. https://doi.org/10.1509/jmkr.48.3.617

Nesbit, P. L., & Wood, R. E. (2002). Improving confidence and accuracy in performance appraisals. *Journal of Management & Organization*, *8*(2), 40–51. https://doi.org/10.1017/S1833367200005010

Nickerson, A. B., & Nagle, R. J. (2001). Interrater reliability of the devereux behavior rating scale-school Form: The influence of teacher frame of reference. *Journal of Psychoeducational Assessment*, *19*(4), 299–316. https://doi.org/10.1177/073428290101900401

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes*, *39*(1), 84–97. https://doi.org/10.1016/0749-5978(87)90046-X

Numprasertchai, H. P., & Swierczek, F. W. (2006). Dimensions of success in international business negotiations: A comparative study of Thai and international business negotiators. *Journal of Intercultural Communication*, (11). 3–3. Retrieved from ufh. (20787866)

O-Net Online. (2017). First-line supervisors of retail sales workers. Retrieved from https://www.onetonline.org/link/summary/41-1011.00

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Piaget, J. (1997). *The moral judgement of the child*. New York, Simon and Schuster. Retrieved from https://books.google.com/books?id = n7Gyqmp6cU8C

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, *69*(4), 581–588. https://doi.org/10.1037/0021-9010.69.4.581

Quattrone, G. A. (1982). Overattribution and unit formation: When behavior engulfs the person. *Journal of Personality and Social Psychology*, *42*(4), 593–607. https://doi.org/10.1037/0022-3514.42.4.593

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. TX: Stata Press: College Station.

Reb, J., Greguras, G. J., Luan, S., & Daniels, M. A. (2014). Performance appraisals as heuristic judgments under uncertainty. In S. Highhouse, R. Dalal, & E. Salas (Eds.), *Judgment and decision making at work* (pp. 13–36). New York, NY: Routledge.

Richter, M., König, C. J., Koppermann, C., & Schilling, M. (2016). Displaying fairness while delivering bad news: Testing the effectiveness of organizational bad news training in the layoff context. *Journal of Applied Psychology*, *101*(6), 779–792. https://doi.org/10.1037/apl0000087

Roberson, Q. M., & Stewart, M. M. (2006). Understanding the motivational effects of procedural and informational justice in feedback processes. *British Journal of Psychology*, *97*(3), 281–298. Retrieved from asn

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*(2), 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Roediger, H. L. I., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *31*(5), 1155–1159.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology*, *87*(4), 735–746. https://doi.org/10.1037/0021-9010.87.4.735

Selvarajan, T. T., & Cloninger, P. A. (2012). Can performance appraisals motivate employees to improve performance? A Mexican study. *The International Journal of Human Resource Management*, *23*(15), 3063–3084. https://doi.org/10.1080/09585192.2011.637069

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222. https://doi.org/10.1037/0033-2909.134.2.207

Sheldon, O. J., Dunning, D., & Ames, D. R. (2014). Emotionally unskilled, unaware, and uninterested in learning more: Reactions to feedback about deficits in emotional intelligence. *Journal of Applied Psychology*, *99*(1), 125–137. https://doi.org/10.1037/a0034138

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology*, *99*(6), 917–932. https://doi.org/10.1037/a0021540

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review*, *24*, 86–97.

Smith-Jentsch, K. A., Campbell, G. E., Milanovich, D. M., & Reynolds, A. M. (2001). Measuring teamwork mental models to support training needs assessment, development and evaluation: Two empirical studies. *Journal of Organizational Behavior*, *22*(2), 179–194. Retrieved from ABI/INFORM Collection; Research Library. (224898921)

Stamoulis, D. T., & Hauenstein, N. M. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, *78*(6), 994–1003. https://doi.org/10.1037/0021-9010.78.6.994

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*(3), 497–506. https://doi.org/10.1037/0021-9010.73.3.497

Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology*, *77*(4), 501–510. https://doi.org/10.1037/0021-9010.77.4.501

Sulsky, L. M., & Day, D. V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology*, *79*(4), 535–543. https://doi.org/10.1037/0021-9010.79.4.535

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology*, *93*(3), 711–719. https://doi.org/10.1037/0021-9010.93.3.711

Van Boven, L., Dunning, D., & Loewenstein, G. (2000). Egocentric empathy gaps between owners and buyers: Misperceptions of the endowment effect. *Journal of Personality and Social Psychology*, *79*(1), 66–76. https://doi.org/10.1037/0022-3514.79.1.66

VandeWalle, D., Cron, W. L., & Slocum, J. W. J. (2001). The role of goal orientation following performance feedback. *Journal of Applied Psychology*, *86*(4), 629–640. https://doi.org/10.1037/0021-9010.86.4.629

Waldersee, R., & Luthans, F. (1994). The impact of positive and corrective feedback on customer service performance. *Journal of Organizational Behavior*, *15*(1), 83–95. https://doi.org/10.1002/job.4030150109

West, P. M. (1996). Predicting preferences: An examination of agent learning. *Journal of Consumer Research*, *23*(1), 68–80. Retrieved from bsu

Whiteside, D. B., & Barclay, L. J. (2016). When wanting to be fair is not enough: The effects of depletion and self-appraisal gaps on fair behavior. *Journal of Management*, *44*(8), 3311–3335. https://doi.org/10.1177/0149206316672531

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189–205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Wübben, M., & von Wangenheim, F. (2008). Instant customer base analysis: Managerial heuristics often "get it right". *Journal of Marketing*, *72*(3), 82–93. Retrieved from JSTOR

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

AUTHOR BIOGRAPHIES

Ming-Hong Tsai is an Assistant Professor in the School of Social Sciences at Singapore Management University. His research examines judgment and decision-making, group dynamics, and conflict management.

Serena Wee is a Senior Lecturer in the School of Psychological Science at the University of Western Australia. Her research interests include recruitment and selection, organizational diversity, and research methods.

Brandon Koh is a PhD candidate in the School of Social Sciences at Singapore Management University. His research interests include creativity, culture, and performance assessment.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Tsai M-H, Wee S, Koh B. Restructured frame-of-reference training improves rating accuracy. *J Organ Behav.* 2019;1–18. https://doi.org/10.1002/job.2368