Industrial and Organizational Psychology (2024), **17**, 206–219
doi:10.1017/iop.2024.6

CAMBRIDGE UNIVERSITY PRESS

FOCAL ARTICLE

Beyond rating accuracy: Unpacking frame-of-reference assessor training effectiveness

C. Allen Gorman[1], Duncan J. R. Jackson[2], John P. Meriac[3], Joseph R. Himmler[4], and Tanya F. Contreras[1]

[1]University of Alabama at Birmingham, Birmingham, USA, [2]King's College London, London, UK, [3]University of Missouri-St Louis. St-Louis, USA and [4]Auburn University, Auburn, USA
Corresponding author: C. Allen Gorman; Email: cagorman@uab.edu

(Received 24 December 2023; accepted 27 December 2023; first published online 12 March 2024)

Abstract
Evidence from previous research suggests that frame-of-reference (FOR) training is effective at improving assessor ratings in many organizational settings. Yet no research has presented a thorough examination of systematic sources of variance (assessor-related effects, evaluation settings, and measurement design features) that might influence training effectiveness. Using a factorial ANOVA and variance components analyses on a database of four studies of frame-of-reference assessor training, we found that (a) training is most effective at identifying low levels of performance and (b) the setting of the training makes little difference with respect to training effectiveness. We also show evidence of the importance of rater training as a key determinant of the quality of performance ratings in general. Implications for FOR training theory and practice are discussed.

Keywords: Assessment centers; assessor training; frame-of-reference training; variance partitioning

Introduction

Training raters, or assessors, is fundamental to a number of HR activities, including job analysis, personnel selection, performance management, and employee development (Aguinis & Kraiger, 2009; Dierdorff et al., 2010). Assessor training has been demonstrated as an effective means of improving rating accuracy in assessment center (AC) and performance appraisal contexts, particularly approaches such as frame-of-reference (FOR) training (Bernardin & Buckley, 1981). However, less is known about the influence of factors associated with assessor training, such as the training setting, trainer characteristics, ratee performance level, training protocols utilized, training stimulus materials, rating instruments utilized, or even training duration. All of these factors are theoretically relevant considerations that may impact rating quality, yet no research has emerged that has made any meaningful comparisons among them. Although there are a number of factors that can potentially impact AC ratings, this study was undertaken as a first step at disentangling some of the common sources of variance associated with assessor training that may influence trainees' performance ratings. Specifically, using a database of four previously conducted studies of FOR training, we examine four factors common to all of the studies in the database: training condition, performance level, performance dimension, and setting.

Research on FOR assessor training

FOR training has emerged as the preferred method of assessor training due to the robust evidence indicating FOR training's effectiveness at improving rating accuracy (Chirico et al., 2004;

Roch et al., 2012). Bernardin and Buckley (1981) originally proposed FOR training as an alternative to rater error training in response to the inconsistent results produced by rater error training. FOR training focuses on providing raters with performance standards for each dimension to be rated (Woehr & Huffcutt, 1994), with the ultimate goal of raters sharing and using common conceptualizations of performance (Athey & McIntyre, 1987; Gorman & Rentsch, 2009, 2017; Woehr, 1994). Although the majority of FOR training research has been conducted in the performance appraisal domain, recent applications of FOR training have been effectively utilized in other domains such as ACs (Jackson et al., 2005), job analysis (Aguinis et al., 2009), competency modeling (Lievens & Sanchez, 2007), and job interviews (Melchers et al., 2011).

Despite the impressive evidence in support of the efficacy of FOR training for increasing rating accuracy, there are theoretical and operational issues with traditional estimates of rating accuracy. Such estimates include Cronbach's (1955) accuracy components (elevation, differential elevation, stereotype accuracy, and differential accuracy), distance accuracy (McIntyre et al., 1984), and Borman's differential accuracy (1977). First, all of these accuracy indices involve a direct comparison between assessor ratings and a set of "true score" or expert ratings. This essentially results in a set of difference scores. Problems inherent in the use of difference scores have been well documented (Edwards, 1995, 2001), foremost being the typical lack of reliability in difference scores. Second, scholars have raised concerns over "true score" development, including: (a) variation in how "true scores" are operationally defined, (b) lack of agreement among expert ratings, and (c) lack of congruence between conceptual and operational definitions of "true scores" (Sulsky & Balzer, 1988). Recognizing issues surrounding true scores, researchers have begun to recognize the benefits of considering alternative indicators of rating quality besides rating accuracy (Greguras et al., 2003; Hoffman et al., 2012). Responding to the call from several researchers for alternative approaches that move beyond traditional accuracy measures, this research incorporates multiple sources of variance in examining the psychometric quality of performance ratings (Murphy & Cleveland, 1995; Roch et al., 2012; Sulsky & Balzer, 1988). Accordingly, we utilize a quasi-experimental design to examine psychometric characteristics associated with AC ratings.

Variance partitioning and FOR training effectiveness

Variance partitioning approaches based on factorial ANOVA and variance components analyses have been applied to assessor training research and have been used to establish, following training, assessors' capacity to distinguish among different AC dimensions and to determine behavioral consistency within AC dimensions across exercises (Lievens, 2001a; 2001b). For example, using generalizability analyses, Lievens (2001a) found that FOR-trained assessors differentiated more effectively among dimensions and provided more reliable ratings than raters who received data-driven or control training. In a similar study utilizing generalizability analyses, Lievens (2001b) found that after training, assessors were able to reliably differentiate AC candidate performance among multiple dimensions.

However, one hallmark of FOR training is the ability of FOR-trained assessors to categorize performance information as either positive (i.e., high level of performance) or negative (i.e., low level of performance; Gorman & Rentsch, 2009, 2017). For example, Melchers et al. (2011) found a large performance level effect in a study of FOR training effects on interview ratings. However, previous studies have not examined performance level[1] as a between-subjects factor, which is important for establishing the extent to which FOR training plays a role in fostering assessor sensitivity to performance variability. Prior research has shown that evaluators may be more attuned to negative performance information (e.g., Fiske, 1980), suggesting that raters may not

[1]Because there were only two ratees in our analyses, one representing high performance and the other representing low performance, it was not reasonable to estimate reliability based on the results of the variance components analyses.

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

evaluate performance in equally effective ways at high and low levels of the performance continuum. Currently, it is not clear whether the effects of FOR training help raters evaluate performance effectively at all levels of performance. Accordingly, we investigate the following research question:

Research Question 1: Does training condition (FOR vs. control) interact with ratee performance level (high vs. low) to influence assessor ratings?

Despite the findings reviewed above, training location may also represent a design consideration that impacts training effectiveness. Some scholars have recognized that training characteristics may differ across settings, even if FOR training is still implemented. As noted by Cheng and Ho (2001), by testing variables in different training settings, "a more consistent view of their functions on training transfer [can] be obtained" (p. 110). According to Orpen (1999), both individual and organizational aspects of the training environment can impact its effectiveness. Even if training is implemented in identical ways, individual differences may be present in motivation or other considerations in how they approach and plan to utilize the training. Although the core elements of FOR training may be present, different aspects of the training environment may potentially impact training effectiveness.

In the present study, we tested for rating differences in training sessions conducted by different trainers at different universities in different regions of the United States. Although these differences in settings could potentially lead to meaningful differences in the effectiveness of the training, meta-analytic studies by both Roch et al. (2012) and Woehr and Huffcutt (1994) found evidence for the effectiveness of FOR training regardless of setting. Based on these empirical findings, we offer the following research question for the present study:

Research Question 2: What proportion of variance in ratee performance is accounted for by the effect of the setting in which the FOR or control training occurred?

In addition, it is commonly found that FOR training increases assessors' ability to recognize patterns (or levels) of performance in the assessor training literature. For instance, Woehr (1994) found that FOR training improves assessors' knowledge of performance-related information, and Schleicher and Day (1998) found that FOR training produced less idiosyncratic representations of performance in assessors. Building on these findings, Gorman and Rentsch (2009) found that FOR-trained assessors possessed performance schemas that were more similar to an expert schema (compared to control-trained assessors), and the accuracy of their schemas accounted for a significant amount of incremental variance in rating accuracy over that of declarative knowledge alone.

In a meta-analysis of methodological factors in AC ratings, Woehr and Arthur (2003) found evidence that assessor training directly enhances assessors' ability to process the vast amount of performance information presented in typical AC exercises. These results are supported by the FOR training literature that suggests that performance information is recalled in accordance with how the information is categorized, such as dimension level and performance level (Day & Sulsky, 1995; Schleicher & Day, 1998; Woehr, 1994). Thus, the evidence indicates that FOR training works by training assessors how to recognize and interpret performance behaviors as either generally positive (high level of performance) or generally negative (low level of performance) and then further categorizing those behaviors into the appropriate performance dimension.

Despite the improvements shown through FOR training, research has established that attention to negative performance information may be weighed more heavily than positive information (Fiske, 1980). Although FOR training aims to assist in the identification of relevant behaviors and scaling of behaviors, it is unclear whether it is equally effective in doing so for both high and low levels of performance. Accordingly, we offer the following research question:

Research Question 3: What proportion of variance in ratee performance is accounted for by performance level in the FOR training condition?

One commonly reported finding in the AC literature is that exercise effects tend to emerge as dominant sources of variance relative to dimension effects (e.g., Lance et al., 2004; Lance, 2008). In part, exercise effects entail large intercorrelations among the same dimensions sampled within

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

exercises. While we do not use exercises in the present study, we do, however, employ the use of differing performance levels, which could offer an explanation for what have historically been labeled exercise effects but could, in part, be performance-level effects (e.g., one exercise might be more challenging than another). Given previous findings on exercise effects, we expect that dimensions are unlikely to vary significantly within performance levels.

It is important to recognize however, that the design of ACs is inherently complex, potentially involving several nested and crossed effects as well as interactions among AC design features. Rather than expecting that the aforementioned AC features operate in the same way across ACs, it is possible that interactions may exist between performance levels and AC settings. In ACs, participants are rated by assessors on the basis of dimensions that could be scored high or low. This evaluation could be affected by the setting in which the evaluation takes place. These factors (participants, dimensions, levels of performance, assessors, evaluation setting) could furthermore interact and affect the ratings generated for each participant. The purpose of the current research is to determine where and how much meaningful variance exists in FOR training situations. Thus, in keeping with the discussion about performance levels, we offer the following research question regarding interactions involving settings:

Research Question 4: What proportion of variance in ratee performance is attributable to (a) dimensions nested within levels, (b) setting by level interactions, (c) setting by dimension interactions nested within levels, and (d) assessor by level interactions nested within setting?

Method

Participants

The database for the present study was compiled from archival data from four previous studies of FOR assessor training conducted between 2007-2012. Three of the datasets have been published (Gorman & Rentsch, 2009, 2017; Hoffman et al., 2012) and one is unpublished (Gorman & Meriac, 2022). The total dataset consisted of 471 participants from 3 different locations (144 undergraduate participants from a large southeastern US university, 240 undergraduate participants from a regional southwestern US university, and 87 undergraduate participants from a regional southeastern US university). All participants completed the study in exchange for extra course credit at their respective university. The same standardized protocol and procedures were used at all locations (see Procedure below). For the large southeastern US university location, the mean age of participants was 21.44 years ($SD = 3.73$) and most held part-time jobs (60%). This sample was predominantly Caucasian (90%) and male (56%), and 77% of the sample reported that they had no experience rating the performance of another person. For the regional southwestern US university location, the mean age of participants was 20.37 years ($SD = 3.66$) and most held part-time jobs (60%). This sample was predominantly Caucasian (64%) and female (61%), and 60% of the sample reported that they had no experience rating the performance of another person. For the regional southeastern US university location, the mean age of participants was 20.00 years ($SD = 4.09$) and 33% held part-time jobs. This sample was predominantly Caucasian (79%) and female (55%), and 75% percent of the sample reported that they had no experience rating the job performance of another person.

Procedure

All training procedures and materials can be located on the Open Science Framework (https://osf.io/8tq5r/?view_only=85a74d720eb6461c827741603a51a8ce). In all studies, participants were randomly assigned to either a FOR training condition ($n = 302$) or a control training condition ($n = 169$). The attendance at each session ranged from 3-10 participants. Before each session, participants received a brief introduction. Next, participants received either FOR training or

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

control training. Participants then viewed two videotaped performance episodes (described below), presented in random order. The participants watched the first video, made notes regarding the behaviors they observed on a rating form, then provided summary ratings for each of the five dimensions. After observing and making ratings for the first video, the process was then repeated for the second video. During the presentation of the videotapes, participants recorded behaviors on a rating form. At the conclusion of each performance episode, participants wrote their ratings on the form.

Stimulus materials

The performance episodes that served as the stimuli in the present study consisted of two videotaped AC exercises featuring actual participants in an operational developmental AC for senior-level executives. These videos have been utilized in prior FOR training research (Gorman & Rentsch, 2009, 2017; Hoffman et al., 2012). One of the videos featured a candidate that had been rated by the AC assessors as below average across most dimensions (low performance), and the other video featured a candidate that had been rated as above average across most dimensions (high performance). Performance was not uniform across all dimensions, but differed somewhat within each candidate's evaluation in the videos, performing differently on each specific dimension. Consistent with other assessor training studies (Sulsky & Day, 1992, 1994; Schleicher & Day, 1998), the videos depicted a scenario in which an executive played the role of a sales manager that is holding a one-on-one meeting with a subordinate. In each video, the executive meets with a role player in the exercise that was designed to elicit behaviors relevant to the following dimensions: analysis, decisiveness, leadership, confrontation, and interpersonal sensitivity. Each of these dimensions are routinely rated across multiple exercises in the operational AC, and each dimension is designed to capture specific candidate behaviors relevant to the scenario (e.g., stating the goals and purposes for the meeting and soliciting input from the employee are behavioral examples indicative of leadership).

The videos were selected for use in prior research (see Gorman & Rentsch, 2009) by a research team based on the ease of observability of specific behaviors, the clarity of the video, the quality of the audio, and clear and unambiguous demonstration of specific positive and negative behaviors relevant to each dimension. Each video was approximately 15 minutes long. Using procedures outlined by Sulsky and Balzer (1988), each video exercise was rated by a team of three upper-level graduate students industrial and organizational psychology who had been trained as AC assessors and had an average of 3 years of AC experience. The experienced raters independently observed and rated each video, and then the raters met to achieve consensus on a final set of scores for each video. The final consensus ratings for the overall low-performance video were as follows: analysis = 2.7, decisiveness = 2.7, leadership = 2.7, confrontation = 2.7, and sensitivity = 3.5. The final consensus ratings for the overall high-performance video were as follows: analysis = 4.0, decisiveness = 3.5, leadership = 3.7, confrontation = 4.0, and sensitivity = 73.7.

Conditions

All sessions were conducted by trained graduate students using a standard written set of procedures.

FOR training

Consistent with previous FOR training research (Gorman & Rentsch, 2009, 2017; Woehr, 1994), the FOR training proceeded according to the following protocol outlined by Pulakos (1984, 1986). First, participants were told that they would evaluate the performance of AC candidates on separate performance dimensions and were given rating scales and instructed to read them as the

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

trainer read the dimension definitions and scale anchors aloud. The trainer discussed ratee behaviors that illustrated different performance levels for each scale. Then, participants were shown a video of a practice vignette featuring a mixed level of performance (some dimensions were above average and some were below average). Participants were asked to evaluate the example ratee using the scales provided, and the ratings were written on a whiteboard and discussed by the group of participants. Finally, the trainer provided feedback to participants explaining why the ratee should receive a particular rating (target score) on a given dimension. The training session lasted approximately 45 minutes.

Control training

Participants in the control training were also instructed that they would be evaluating AC candidate performance on the five performance dimensions. They were also presented with the rating form, and the trainer read over each of the dimension definitions. However, no other specific training was provided. Rather, a broad training video on assessing performance in organizations was shown. The control training session also lasted approximately 45 minutes.

Rating form

Consistent with previous research (Gorman & Rentsch, 2009, 2017; Woehr, 1994), the rating form listed dimensions, provided space for participants to take notes regarding the manager's behavior and make ratings for each dimension. Participants recorded behaviors on their forms as they observed them. For each behavior that was recorded, participants were instructed to place either a +, -, or 0 next to the behavior to indicate whether the behavior was a positive, negative, or neutral behavior. After reviewing their notes for each video, participants recorded their rating for each dimension in the spaces provided. Each dimension was rated using an 11-point scale (1.0 = *extremely weak* to 5.0 = *exceptional*).

Analyses

All analyses were performed in this study using SPSS. In the factorial ANOVA, all effects were treated as fixed, except the effect for assessors, which was treated as random. In the variance components analyses, all effects were treated as random, as is common in this type of analysis (Brennan, 2001b; Putka & Hoffman, 2013; Searle et al., 2006). Variance components were estimated using restricted maximum likelihood procedures (e.g., Marcoulides, 1990). For interested readers, we recommend Brennan (2001a, 2001b), Howell (2007), Searle et al. (2006), and Shavelson and Webb (1991, 2005) for excellent primers on the basics of factorial ANOVA and variance components analyses.

Results

Table 1 shows intercorrelations among high performance versus low performance dimension observations for the FOR-trained group. Table 2 shows the same analyses for the control group. As expected, low performance was associated with lower mean dimension scores and higher performance with higher mean scores. Generally, standard deviations were marginally smaller in the trained versus the control group. Dimensions were intercorrelated to some extent for both the trained ($M_r = .42$, $SD_r = .16$) and the control ($M_r = .39$, $SD_r = .15$) groups, and these estimates were lower than those reported elsewhere. For instance, Bowler and Woehr (2006) reported a meta-analytic estimate of .79, suggesting more discriminability among dimensions here. Coefficients alpha for all low- and high-performance dimensions approached .80. We present the standardized version of coefficient alpha (Kline, 1999), because the overall rating may have

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

Table 1. Dimension Intercorrelations—FOR-Trained Group

	M	*SD*	1	2	3	4	5
Low Performance Observations							
1. *A*	2.79	0.71					
2. *D*	2.66	0.79	.36				
3. *L*	2.44	0.79	.43	.42			
4. *C*	2.90	0.97	.38	.20	.45		
5. IS	3.02	0.79	.34	.17	.33	.23	
6. *O*	2.67	0.61	.68	.51	.70	.56	.46
High Performance Observations							
1. *A*	3.80	0.56					
2. *D*	3.59	0.58	.35				
3. *L*	3.82	0.63	.29	.45			
4. *C*	3.72	0.56	.25	.37	.39		
5. IS	3.74	0.64	.22	.90	.33	.34	
6. *O*	3.86	0.47	.49	.49	.59	.48	.46

Note. Dimensions included: *A* = Analysis, *D* = Decisiveness, L = Leadership, *C* = Confrontation, IS = Interpersonal Sensitivity, *O* = Overall dimension. Coefficients alpha were estimated at .81 (low dimensions) and .79 (high dimensions). All correlations were significant ($p < .05$).

Table 2. Dimension Intercorrelations—Control-Trained Group

	M	*SD*	1	2	3	4	5
Low Performance Observations							
1. *A*	3.07	0.93					
2. *D*	2.97	1.01	.32				
3. L	2.82	0.93	.28	.41			
4. *C*	2.49	0.91	.30	.31	.35		
5. IS	2.96	1.06	.30	.21	.38	.38	
6. *O*	2.94	0.72	.65	.53	.67	.57	.53
High Performance Observations							
1. *A*	3.90	0.73					
2. *D*	3.53	0.84	.37				
3. *L*	3.83	0.80	.31	.45			
4. *C*	3.71	0.91	.26	.22	.24		
5. IS	3.74	0.89	.34	.27	.16	.25	
6. O	3.93	0.63	.63	.59	.62	.43	.51

Note. Dimensions included: *A* = Analysis, *D* = Decisiveness, *L* = Leadership, *C* = Confrontation, IS = Interpersonal Sensitivity, *O* = Overall dimension. Coefficients alpha were estimated at .81 (low dimensions) and .78 (high dimensions). All correlations were significant ($p < .05$).

been conceptualized by assessors in a manner that was slightly different from the other dimensions. Nevertheless, the regular alpha estimates were almost identical to those of the standardized estimates.

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

Table 3. Factorial Analysis of Variance Comparing FOR-Trained versus Control-Trained Assessors

Effect	*df*	MS	*F*	*p*
c (condition)	1	23.66	11.97	<.01
a:*c* (assessor in condition)	168	1.98	4.80	<.01
l (performance level)	1	1059.89	938.64	<.01
d:*l* (dimension in level)	5	13.99	33.98	<.01
cl (condition x level interaction)	1	7.72	18.76	<.01
cd:*l* (condition x dimension interaction in level)	5	1.48	3.61	<.01
al:*c* (assessor x level interaction in condition)	301	1.41	3.43	<.01
ad:*cl*,*e* (highest-order interaction confounded with residual error)	4863	.41		

Note. Conditions = FOR-trained versus control-trained assessors; level = high versus low performance observations.

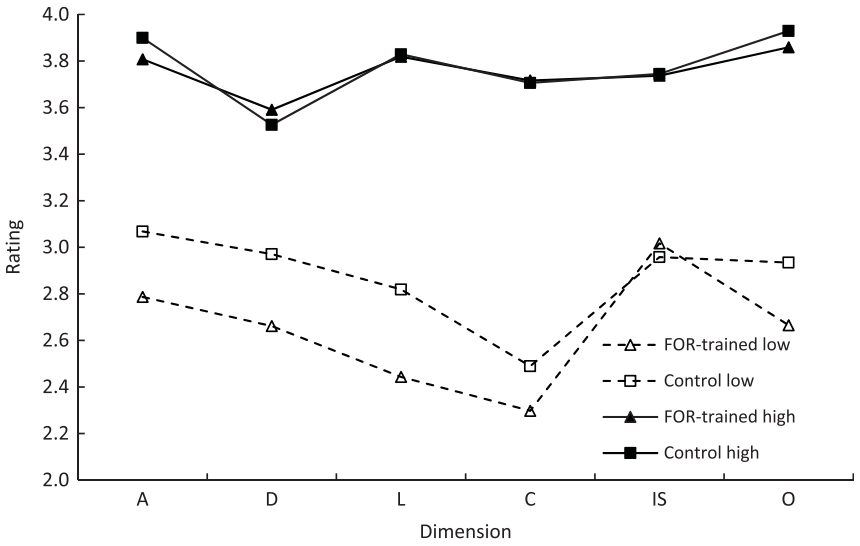

Figure 1. Marginal mean ratings plotted with relation to dimension observations from FOR-trained versus control-trained assessors. Dimension ratings are distinguished by being relevant to videos of low versus high performance. Since dimensions, as presented here, are nominal, lines are shown for clarity only.

To gain an understanding about FOR-trained versus control assessor behavior, Table 3 shows the results of a factorial ANOVA. This analysis included effects for conditions, assessors nested in conditions, high versus low levels of performance, dimensions nested in levels, interaction effects, and residual error. Significant effects ($p < .01$) were observed for all main effects and interactions in the model. Addressing Research Question 1, the results indicated that a significant interaction was present between training condition and performance level. Because the factorial ANOVA suggested that higher-level interactions were present, marginal means were plotted incorporating different levels and conditions with respect to assessment on the dimensions under scrutiny. Figure 1 shows that, in relative terms, high versus low examples of performance were correctly identified by the assessor group. With respect to examples of high performance, the presence of training did not appear to make much difference to the pattern of ratings observed. Both trained and untrained assessors evaluated high performance similarly. However, when considering low performance, ratings from trained versus untrained assessors differed on five of the six dimensions

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

Table 4. Variance Components Analyses Comparing FOR-Trained versus Control-Trained Assessor Subgroups

	Trained			Control		
Effect	*df*	VC	%	*df*	VC	%
s (setting)	2	.004	0.3	1	.017	1.4
a:*s* (assessor in setting)	299	.020	1.7	167	.000	0.0
l (performance level)	1	.627	54.6	1	.401	33.5
d:*l* (dimension in level)	10	.035	3.0	10	.027	2.3
sl (setting x level interaction)	2	.019	1.7	1	.002	0.2
sd:*l* (setting x dimension interaction in level)	20	.002	0.2	10	.003	0.3
al:*s* (assessor x level interaction in setting)	299	.149	13.0	167	.273	22.8
ad:*sl*,*e* (highest-order interaction confounded with residual error)	2990	.292	25.4	1670	.475	39.6

Note. Restricted maximum likelihood (REML) estimates are presented above. VC = variance component, % = percent of total variance explained.

evaluated. Specifically, untrained assessors tended to produce higher ratings than trained assessors for low-performing assessees.

In addition to an overall view of mean differences among effects, the magnitude of various effects within each assessor subgroup (trained versus control) was also of interest. Table 4 shows two separate variance components analyses, for the FOR-trained group and the control group. For each of the conditions, eight effects are presented with associated variance components and the percentage of total variance explained by each effect. To address Research Question 2, the main effect for setting appeared to have little effect on variation among scores (.3% of variance explained in the FOR-trained group and 1.4% of variance explained in the control group). This suggests that the effects of FOR training did not vary across the different settings in which the data were collected. In the trained group, assessor variance also explained a minimal amount of the total variation in scores (1.7%). In response to Research Question 3, the results indicated that assessors distinguished between high and low performance levels, as indicated by the effect for *l*, and the FOR-trained group was considerably more sensitive to performance differences (54.6% of variance) than the untrained control group (33.5%).

Research Question 4a asked what proportion of variance was attributable to dimensions nested within levels, *d*:*l*. As shown in Table 4, a modest proportion of variance was explained by the *d*:*l* effect for trained assessors (3%) and assessors in the control group (2.3%). With regard to interactions, *sl* represents the extent to which distinctions between performance levels are contingent on different settings (Research Question 4b). Ordinarily, such effects would be indicative of the setting interfering with or contributing to fluctuations in measurement and the ultimate aim would be to minimize them (Brennan, 2001a). Here, the effect for *sl* was negligible across both the trained (1.7% of variance explained) and control groups (.2%). The next interaction (Research Question 4c), *sd*:*l*, is similar to that for *sl*, except that it takes error associated with dimensions into account. Again, the effects associated with *sd*:*l* were negligible (.2% for the trained group, .3% for the control group), suggesting that different settings did not contribute substantially to performance level distinctions when dimension error was taken into account.

Finally, to address Research Question 4d, the effect for *al*:*s*, shown in Table 4, is suggestive of whether the distinction between performance levels was bound by error associated with assessors nested in different settings. Thus, the *al*:*s* effect simultaneously takes performance level distinction, assessor error, and different settings into account. The aim here, again, is to minimize this effect as a source of error, because settings and assessors have the potential to interfere with useful distinctions among performance levels. Table 4 shows that this effect was notably lower for

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

the trained (13.0% of variance explained) relative to the control group (22.8% of variance explained). The final effect, that for *ad*:*sl*,*e*, represents the highest order effect confounded with residual error. The *ad*:*sl*,*e* effect was also somewhat lower for the trained group (25.91% of variance explained) relative to the control group (39.32% of variance explained).

Discussion

Organizations rely on assessor training for a variety of purposes (Dierdorff et al., 2010), yet assessor training effectiveness research has historically relied on a limited set of criteria with serious theoretical and operational shortcomings (Sulsky & Balzer, 1988). Moreover, no research has examined the extent to which assessor training generalizes across settings in which the training is conducted. A factorial ANOVA revealed that FOR-trained assessors were more sensitive to distinguishing low levels of performance than control-trained assessors on five of the six dimensions rated (including an overall performance rating; see Figure 1). The variance components analyses indicated that: (a) the setting had little effect on the total variation in ratings in both conditions, (b) performance level accounted for substantially more variation in ratings than assessor idiosyncrasy effects (i.e., assessor and residual error effects) in the FOR training condition (but assessor idiosyncrasy effects were larger than the performance level effect in the control training condition), and (c) assessor x level interaction nested within setting error and random error effects were larger in the control training condition than the FOR training condition.

Does the level of performance matter?

The factorial ANOVA results suggest that FOR training is most effective at helping assessors identify low levels of performance. At high performance levels, differences between FOR-trained and control ratings were negligible. Because assessors in both conditions appeared to have little difficulty distinguishing high levels of performance, FOR training programs might be tailored to focus more on negative behaviors. For example, additional practice examples of negative performance might be given during a training session since the ability to recognize negative behaviors appears to be driving the distinction between FOR and control training. Alternatively, it could also be argued that FOR training should focus additional efforts on recognizing high levels of performance since there was little difference in the ratings at high performance levels. However, we support the former interpretation, as it has been found in previous research that rating accuracy tends to be higher at low levels of performance (Gorman & Rentsch, 2009, 2017). These findings are consistent with the person perception literature, where raters tend to weigh negative information more heavily than positive information (Fiske, 1980). Overall, our results suggest that to judge low-level performance with any degree of sensitivity, assessors must first be trained.

Additionally, the generally higher ratings provided by control-trained assessors at low performance levels could partially explain the widespread problem of rating inflation in performance appraisals (e.g., Jawahar & Williams, 1997). In fact, Bernardin and Buckley (1981) proposed FOR training to help alleviate this and other rating distribution problems. It could be that, in the absence of training, assessors provide higher ratings for low performers because they cannot or do not distinguish negative performance at lower levels. This hypothesis has yet to be tested empirically. Finally, these results could point to an ability factor in the assessment process in that, without training, assessors may lack the ability to distinguish low levels of performance. This ability component could be an important contextual factor in the performance appraisal domain, in addition to other factors such as politics and motivation (Levy & Williams, 2004).

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

Which design elements matter?

Our results showed that the effect of setting was negligible, across different geographic locations and demographics such as age, gender, race/ethnicity, and job status. This finding is consistent with a recent meta-analysis of FOR training which found no significant moderator effects of protocol differences (such as type of participants, purpose of training, and type of training material) on FOR training effectiveness (Roch et al., 2012). We also found that the assessor-nested-within-setting effect accounted for a negligible amount of variance in ratings, where differences among assessors (within settings) had little impact on the total variation in ratings. This is consistent with previous research that has found that FOR training works by imparting a shared schema of performance on trainees (Gorman & Rentsch, 2009, 2017). Our results suggest that the shared schema among FOR-trained assessors acts as a buffer that minimizes the influence of assessor idiosyncrasies on ratings. This result further strengthens the evidence in support of the efficacy of FOR training.

Further support for the suppressing effect of FOR training on assessor idiosyncrasies was evidenced in our finding that the performance level effect accounted for more variance than the combined assessor and random error effects in the FOR training condition. However, in the control training condition, the combination of assessor and random error effects accounted for more variance than the performance level effect. Again, this is consistent with scores of previous studies of FOR training effectiveness and provides additional evidence in support of the shared schema hypothesis.

Limitations and future directions

As with any study, there are possible limitations to consider. It is still possible that unmodeled variance sources may still be worth considering. For instance, the FOR training procedure implemented here were all systematic and contained the same design consistent with the extant research (Pulakos, 1984; 1986). However, it is unclear to what extent complete FOR procedures are implemented in practice, and which components of FOR training have the greatest effect (e.g., examples of high and low performance, practice, etc.). In addition, assessor experience could also be modeled as another factor. It is possible that more experienced assessors could provide very different ratings compared with inexperienced assessors, as were used in the present study.

Moreover, an anonymous reviewer pointed out that because we only used two stimulus episodes (one above average performance and one below average performance), there are other confounding characteristics of the managers in the videos, such as attractiveness or age, that could be responsible for our findings. Additionally, it is possible that different performance scenarios or candidates may facilitate increased sensitivity in identifying high levels of performance. Thus, future studies should model these factors using a wide variety of stimulus episodes. Another anonymous reviewer noted that sensitivity in identifying low levels of performance would be particularly helpful in a developmental AC context, but might be much less useful in a selection or promotion context, where the goal would be to differentiate among candidates with relatively high levels of performance. The implication here would be that FOR training may not be necessary in situations such as promotions, where the focus is on high performing candidates.

We also thank an anonymous reviewer for pointing out that the results of the present study raise several questions that lay the groundwork for a potentially fruitful research agenda on FOR training, including examining how FOR training influences differentiation among performance dimensions, testing whether FOR-trained assessors consistently draw lager distinctions between high and low levels of performance as a result of training, and conducting studies to determine if FOR-trained assessors are better at distinguishing among candidates with generally high levels of performance. We also thank an anonymous reviewer for pointing out that the lack of a practice video in the control condition could potentially explain why assessors in the control condition

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

gave higher ratings for the low performance assessee. Although this procedure was consistent with prior research (Gorman & Rentsch, 2009, 2017; Hoffman et al., 2012), we cannot rule this explanation out, and we recommend that future research carefully and systematically examine the impact of different training procedures and protocols. Finally, the purpose of performance ratings could also be examined along with the aforementioned factors. For instance, administrative ratings have long been shown to be more lenient than ratings made for feedback/developmental purposes (Jawahar & Williams, 1997). However, it is possible that FOR training may serve to minimize the performance appraisal purpose effect.

General conclusions

Despite several robust findings on the effectiveness of FOR assessor training, the time has come for research to move beyond questioning whether FOR training works to uncovering the boundary conditions of its effectiveness. Using a variance partitioning approach, we discovered that FOR training is most effective at helping assessors evaluate low levels of performance. We found very little difference in the ratings made by FOR-trained and control-trained assessors when the performance level of the stimulus ratee was high. This is an important and unique finding because no previous FOR training studies have explicitly modeled performance level as a factor. Our results suggest that performance level is a meaningful factor that should not only be modeled, but training protocols might also need to be tailored to focus more on negative performance and stimulus episodes should likely include a target ratee whose performance was generally negative across all dimensions. Taken together, these findings suggest that recognizing poor performance may be what makes FOR training so effective compared to other interventions.

Author note. C. Allen Gorman, Department of Management, Information Systems and Quantitative Methods and Department of Health Services Administration, University of Alabama at Birmingham. Duncan J. R. Jackson, King's Business School, King's College London. John P. Meriac, Department of Global Leadership and Management, University of Missouri-St. Louis. Joseph R. Himmler, Department of Psychology, Auburn University. Tanya F. Contreras, Department of Management, Information Systems and Quantitative Methods, University of Alabama at Birmingham.

A version of this paper was presented at the 2012 conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

We thank Katy Gaddis, Steven Apodaca, Josh Collins, Lauren Felton, Garolyn Jergins, Ashley McIntyre, Ben Overstreet, Kenneth Smith, Jessica Stoner, Jennifer Thorndike, Soniya Lonkar, Caitlin Nugent, Erin Carroll, Jeanne Donaghy, Megan Poore, and Dave Sharrer for their assistance with data collection.

References

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology*, **60**, 451–474. Doi: 10.1146/annurev.psych.60.110707.163505.

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**, 405–438. Doi: 10.1111/j.1744-6570.2009.01144.x.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Level of processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, **72**, 239–244. Doi: 10.1037/0021-9010.72.4.567.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, **6**, 205–212. Doi: 10.5465/amr.1981.4287782.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology*, **91**, 1114–1124. Doi: 10.1037/0021-9010.91.5.1114.

Brennan, R. L. (2001a). *Generalizability theory*. Springer Verlag.

Brennan, R. L. (2001b). *Manual for urGENOVA*. Iowa Testing Programs, University of Iowa.

Cheng, E. W. L., & Ho, D. C. K. (2001). A review of transfer of training studies in the past decade. *Personnel Review*, **30**, 102–118. Doi: 10.1108/00483480110380163.

Chirico, K. E., Bucley, M. R., Wheeler, A. R., Facteau, J. D., Bernardin, H. J., & Beu, D. S. (2004). A note on the need for true scores in frame-of-reference (FOR) training research. *Journal of Managerial Issues*, **16**, 382–395.

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity. *Psychological Bulletin*, **52**, 177–193. Doi: 10.1037/h0044919.

Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology*, **80**, 001–009.

Dierdorff, E. C., Surface, E. A., & Brown, K. G. (2010). Frame-of-reference training effectiveness: Effects of goal orientation and self-efficacy on affective, cognitive, skill-based, and transfer outcomes. *Journal of Applied Psychology*, **95**, 1181–1191. Doi: 10.1037/a0020856.

Edwards, J. R. (1995). Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes*, **64**, 307–324. Doi: 10.1006/obhd.1995.1108.

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, **4**, 265–287. Doi: 10.1177/109442810143005.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, **38**, 889–906. Doi: 10.1037/0022-3514.38.6.889.

Gorman, C. A., & Meriac, J. P. Nothing but a "g" thing? Toward an integrative model of frame-of-reference training effectiveness. In: *Academy of Management Annual Conference*, 2022.

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology*, **94**, 1336–1344. Doi: 10.1037/a0016476.

Gorman, C. A., & Rentsch, J. R. (2017). Retention of assessment center rater training: Improving performance schema accuracy using frame-of-reference training. *Journal of Personnel Psychology*, **16**(1), 1–11.

Greguras, G. J., Robie, C., Schleicher, D. J., & Goff, I. I. I. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology*, **56**, 1–21. Doi: 10.1111/j.1744-6570.2003.tb00141.x.

Hoffman, B. J., Gorman, C. A., Blair, C. A., Meriac, J. P., Overstreet, B. L., & Atchley, E. K. (2012). Evidence for the effectiveness of an alternative multisource performance rating methodology. *Personnel Psychology*, **65**, 531–563. Doi: 10.1111/j.1744-6570.2012.01252.x.

Howell, D. C. Statistical methods for psychology, 6th Edition edn. (2007).

Jackson, D. J. R., Atkins, S. G., Fletcher, R. B., & Stillman, J. A. (2005). Frame of reference training for assessment centers: Effects on interrater reliability when rating behaviors and ability traits. *Public Personnel Management*, **34**, 17–30. Doi: 10.1177/009102600503400102.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, **50**, 905–925. Doi: 10.1111/j.1744-6570.1997.tb01487.x.

Kline, P. (1999). *Handbook of psychological testing* (2nd edn). Routledge.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **1**, 84–97. Doi: 10.1111/j.1754-9434.2007.00017.x.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology*, **89**, 377–385. Doi: 10.1037/0021.9010.89.2.377.

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, **30**, 881–905. Doi: 10.1016/j.jm.2004.06.005.

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

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior*, **22**, 203–221. Doi: 10.1002/job.65.

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**, 812–819. Doi: 10.1037/0021-9010.92.3.812.

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports*, **66**(2), 379–386. Doi: 10.2466/pr0.1990.66.2.37.

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

Melchers, K. G., Lienhardt, N., Aarburg, M. V., & Kleinman, 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**, 53–87. Doi: 10.1111/j.1744-6570.2010.01202.x.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. SAGE Publications, Inc.

Orpen, C. (1999). The influence of the training environment on trainee motivation and perceived training quality. *International Journal of Training and Development*, **3**, 34–43. Doi: 10.1111/1468-2419.00062.

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

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes*, **38**, 78–91. Doi: 10.1016/0749-5978(86)90027-0.

https://doi.org/10.1017/iop.2024.6 Published online by Cambridge University Press

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology*, **98**, 114–133. Doi: 10.1037/a0030887.

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

Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame-of-reference training: Content and process issues. *Organizational Behavior and Human Decision Processes*, **73**, 76–101. Doi: 10.1006/obhd.1998.2751.

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components*. Wiley.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Shavelson, R. J., & Webb, N. M. (2005). Generalizability theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Complementary methods for research in education* (3rd ed. pp. 599–612). AERA.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, **73**, 497–506. Doi: 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**, 501–510. Doi: 10.1037/0021-9010.77.4.501.

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

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology*, **79**, 525–534. Doi: 10.1037/0021-9010.79.4.525.

Woehr, D. J., & Arthur, W. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management*, **29**, 231–258. Doi: 10.1016/S0149-2063(02)00216-7.

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

Cite this article: Gorman, C. A., Jackson, D. J. R., Meriac, J. P., Himmler, J. R., & Contreras, T. F. (2024). Beyond rating accuracy: Unpacking frame-of-reference assessor training effectiveness. *Industrial and Organizational Psychology* **17**, 206–219. https://doi.org/10.1017/iop.2024.6