Journal of Applied Psychology
2009, Vol. 94, No. 5, 1336–1344

0021-9010/09/$12.00 DOI: 10.1037/a0016476

Evaluating Frame-of-Reference Rater Training Effectiveness Using Performance Schema Accuracy

C. Allen Gorman
Angelo State University

Joan R. Rentsch
The University of Tennessee

Frame-of-reference training has been shown to be an effective intervention for improving the accuracy of performance ratings (e.g., Woehr & Huffcutt, 1994). Despite evidence in support of the effectiveness of frame-of-reference training, few studies have empirically addressed the ultimate goal of such training, which is to teach raters to share a common conceptualization of performance (Athey & McIntyre, 1987; Woehr, 1994). The present study tested the hypothesis that, following training, frame-of-reference–trained raters would possess schemas of performance that are more similar to a referent schema, as compared with control-trained raters. Schema accuracy was also hypothesized to be positively related to rating accuracy. Results supported these hypotheses. Implications for frame-of-reference training research and practice are discussed.

Keywords: frame-of-reference training, rating accuracy, schema accuracy

Despite the recent focus on cognition and rater training (e.g., Roch & O'Sullivan, 2003; Schleicher & Day, 1998; Sulsky & Kline, 2007), little attention has been directed toward how raters cognitively structure performance information presented during rater training or the accuracy of these cognitive structures. The goal of the present study was to examine cognitive structures as outcomes of rater training by examining the efficacy of performance schema accuracy as a measure of frame-of-reference rater training effectiveness.

Frame-of-Reference Training

In reaction to the inconsistent results produced by rater error training, Bernardin and Buckley (1981) proposed frame-of-reference training as an alternative. *Frame-of-reference training* focuses on providing raters with performance standards for each dimension to be rated (Woehr & Huffcutt, 1994). Specifically, this mode of training involves matching ratee behaviors to their appropriate performance dimensions and correctly judging the effectiveness of those behaviors (Sulsky & Day, 1992, 1994). The ultimate goal of frame-of-reference training is to train raters to share and use common conceptualizations of performance when providing their ratings (Athey & McIntyre, 1987; Woehr, 1994). Many studies have demonstrated the effectiveness of frame-of-reference training for improving rating accuracy (Athey & McIntyre, 1987; Bernardin & Pence, 1980; Cardy & Keefe, 1994; D. V. Day & Sulsky, 1995; Hauenstein & Foti, 1989; McIntyre, Smith, & Hassett, 1984; Noonan & Sulsky, 2001; Pulakos, 1984, 1986; Schleicher & Day, 1998; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992, 1994; Woehr, 1994). In a meta-analytic review of the rater training literature, Woehr and Huffcutt (1994) found an average effect size (*d*) of .83 for frame-of-reference training compared with control and no-training groups.

Recently, research on frame-of-reference training has focused on the application of such training for use in the training of assessment center assessors with the aim of improving assessment center construct validity (e.g., Goodstone & Lopez, 2001; Lievens, 2001; Schleicher, Day, Mayes, & Riggio, 2002). This research has yielded generally positive results. For example, interrater reliability, rating accuracy, and discriminant validity were better for assessment center assessors who had received frame-of-reference training than for those who had not (e.g., Lievens, 2001). Likewise, Schleicher et al. (2002) found that frame-of-reference training effectively improved the reliability, accuracy, convergent and discriminant validity, and criterion-related validity of assessment center ratings.

In an effort to explain the underlying theoretical reasons for frame-of-reference training effects, researchers have borrowed from various social–cognitive models of person perception and memory, including Carlston's (1992, 1994) associated systems theory (Schleicher & Day, 1998), Klein and Loftus's (1990) elaboration model (Woehr, 1994), and Wyer and Srull's (1989) model of person memory and judgment (D. V. Day & Sulsky, 1995). Taken together, these models suggest that frame-of-reference training influences how raters process, represent, and remember information. Typically, this empirical work requires participants to recall as many behaviors as they can after watching simulations of

C. Allen Gorman, Department of Psychology, Sociology, and Social Work, Angelo State University; Joan R. Rentsch, Department of Management, The University of Tennessee.

This article is based on a doctoral dissertation, submitted to The University of Tennessee, conducted by C. Allen Gorman under the supervision of Joan R. Rentsch. Portions of this article were presented at the 23rd annual conference of the Society for Industrial and Organizational Psychology.

We thank Dave Woehr, Tom Ladd, Michelle Violanti, Lisa Delise, Carrie Blair, Joy Oliver, John Meriac, Brian Hoffman, Wes Davenport, Abby Mello, Melissa Zullo, Ben Overstreet, and Jennifer Godbey for their constructive and insightful comments on earlier versions of this article.

Correspondence concerning this article should be addressed to C. Allen Gorman, Department of Psychology, Sociology, and Social Work, Angelo State University, San Angelo, TX 76909. E-mail: cgorman@angelo.edu

ratee performance. The organization of recalled information is then examined with indexes, such as the adjusted ratio of clustering (Roenker, Thompson, & Brown, 1971), that assess the extent to which behaviors representing the same performance dimensions are recalled in clusters compared with the amount of clustering expected by chance alone.

Such clustering indexes convey some information regarding knowledge organization. However, cognitive structure measurement techniques, such as multidimensional scaling (MDS; Mohammed, Klimoski, & Rentsch, 2000), are based on raters' direct judgments about the interrelationships among behaviors. Paired comparison ratings have been shown to be effective in eliciting cognitive structures (e.g., Schiffman, Reynolds, & Young, 1981) because respondents make similarity ratings by applying the idiosyncratic relevant constructs they use to understand the stimuli (Mohammed et al., 2000). Billings and Cornelius (1980) noted that "one clear advantage" of paired comparison ratings is "the unstructured nature of the rating task" (p. 152). Participants are not told the characteristics of the stimuli to which they should attend when they make their similarity ratings. Furthermore, research supports the validity of MDS as a method for evaluating paired comparison ratings (e.g., Shoben, 1983) and its use as a method for assessing structural knowledge (e.g., Jones, 1983; Neff, 1983; Pollard-Gott, 1983; Rentsch & Klimoski, 2001). We used this method of direct assessment of structured knowledge in the present study to examine the extent to which frame-of-reference training improves the accuracy of performance knowledge structures. The aim of the present study was to extend the literature by directly assessing the accuracy of performance knowledge structures (or schemas) as outcomes of frame-of-reference training using a pre–post design.

Schemas

Within the expert–novice literature, knowledge structures are referred to as semantic nets (e.g., Leinhardt & Smith, 1985), mental models (e.g., Cannon-Bowers, Tannenbaum, Salas, & Converse, 1991), and schemas (e.g., Howell & Cooke, 1989). A schema is a knowledge structure developed from past experience that is used to organize new information and facilitate understanding (Noble, 1989; Poole, Gray, & Gioia, 1990). Schemas aid and are integral to memory (Bartlett, 1932). With advances in learning and domain-relevant experience, the organization of knowledge, or schemas, changes as knowledge moves from declarative to procedural in nature (Cannon-Bowers et al., 1991; Kozlowski, 1998). As individuals become experts in their domain, their schemas become more pattern oriented and more highly integrated, and information is stored in larger chunks (Cannon-Bowers et al., 1991; Leinhardt & Smith, 1985) relative to novice schemas. Expert schemas enable individuals to recognize the similarity between new and previously experienced situations and to adapt old procedures for new situations (Noble, 1989).

Because a primary aim of frame-of-reference training is to develop expert-like performance schemas, performance schemas are legitimate and necessary training criteria. Training in other domains affects relevant schemas (e.g., Koubek, Clarkston, & Calvez, 1994), and expert schema similarity (or schema accuracy) is a useful measure of learning (e.g., Kraiger, Salas, & Cannon-Bowers, 1995). For example, E. A. Day, Arthur, and Gettman (2001) observed that the similarity of trainees' schemas to an expert schema was correlated with skill acquisition and was predictive of skill retention and transfer.

Hypotheses

Although frame-of-reference training is aimed at developing an expert-like performance schema among raters (Goodstone & Lopez, 2001), researchers have not directly evaluated performance schemas using a structural assessment (Goldsmith & Kraiger, 1997) technique. Conducting such an evaluation was one purpose of the present study. The schema similarity approach suggests that individuals' schemas will become more similar over time with advances in learning (E. A. Day et al., 2001; Rentsch & Hall, 1994). In addition, schema similarity research indicates that individuals with more experience with the task of interest have schemas that are more similar to an expert schema of performance than do those with less experience (e.g., Smith-Jentsch, Campbell, Milanovich, & Reynolds, 2001). Therefore, we proposed the following:

> *Hypothesis 1:* Individuals who receive frame-of-reference training will have performance schemas more similar to an experienced schema (i.e., more accurate) (a) after training than before training and (b) than individuals who receive control training.

Further, previous research has indicated that frame-of-reference training is an effective intervention for improving rating accuracy (e.g., Woehr & Huffcutt, 1994). Thus, we posited the following:

> *Hypothesis 2:* Performance ratings from those who receive frame-of-reference training will be more similar to ratings from experienced raters (i.e., more accurate) than will performance ratings from those who receive control training.

If frame-of-reference training is found to be a successful method of increasing performance schema accuracy, then rating accuracy should be positively related to performance schema accuracy. Hence, we hypothesized the following:

> *Hypothesis 3:* Rating accuracy will be positively related to performance schema accuracy in the frame-of-reference training condition.

Prior research has revealed that frame-of-reference training improves raters' knowledge of performance-related information (e.g., Woehr, 1994). Consequently, we made the following prediction:

> *Hypothesis 4:* Individuals who receive frame-of-reference training will score significantly higher on a measure of declarative knowledge than will those who receive control training.

Finally, declarative knowledge is associated with repetition and an early stage of learning. However, as individuals become more knowledgeable in a particular domain, they move beyond accumulating declarative and procedural knowledge by building meaningful relations among known concepts that provide deep, generalizable (some say adaptable) understanding (Cannon-Bowers et al., 1991; Jonassen, Beissner, & Yacci, 1993; Kozlowski, 1998; Leinhardt & Smith, 1985). Organized cognition is indicative of greater experience, deeper understanding, and more pattern orientation with respect to the domain than declarative knowledge (e.g., Dorsey, Campbell, Foster, & Miles, 1999; Kozlowski, 1998; Leinhardt & Smith, 1985). Therefore, we posited the following:

> *Hypothesis 5:* Performance schema accuracy will account for a unique amount of variance in rating accuracy beyond that of a measure of declarative knowledge.

Method

Participants

Of the 144 undergraduate students enrolled at a public southeastern university who participated in this study for extra course credit, 56% were male, and 44% were female. Among the sample, 90% were Caucasian, 7% were African American, 2% were Asian/Pacific Islander, and 1% were of other ethnicity. A total of 60% held at least a part-time job, and 77% had no experience in rating the job performance of another person. Participants were randomly assigned to a frame-of-reference–training condition ($n = 73$) or a control-training condition ($n = 71$).

Procedure

Participants were informed that the purpose of the study was to examine how people evaluate work performance. They received a brief introduction to the session, and then they completed a pretraining performance schema measure. Frame-of-reference or control training occurred next, followed by participants completing the declarative knowledge and the post-training schema measures. Participants then viewed four videotaped performance episodes that were presented in random order across participants. While viewing the videos, participants recorded observed behaviors on a rating form. After viewing each episode, they recorded their ratings. At the conclusion of the session, each participant completed a demographic questionnaire and was debriefed, thanked, and dismissed.

Stimulus Materials

The performance episodes consisted of four 15-min videotaped scenarios from an executive developmental assessment center. The videotapes depicted a role play exercise in which an assessment center candidate assumed the role of a manager and interacted with a subordinate played by a trained assessor. The exercise was designed to elicit behaviors from the candidate that could be grouped into the following performance dimensions: analysis, decisiveness, leadership, confrontation, and sensitivity. To control for confounding effects due to candidate performance level and sex, we required each participant to view two episodes of above-average performance across most dimensions (one male and one female candidate) and two episodes of below-average performance across most dimensions (one male and one female candidate).

Rating Form and Comparison Scores

Participants recorded candidate behaviors as they observed them and placed a plus sign (positive), a minus sign (negative), or a zero (neutral) next to the behavior. After each videotape, participants rated each dimension using an 11-point Likert-type rating scale (1.0 = *extremely weak*, 1.7 = *very weak*, 2.0 = *weak*, 2.5 = *moderately weak*, 2.7 = *slightly weak*, 3.0 = *satisfactory*, 3.5 = *effective*, 3.7 = *very effective*, 4.0 = *highly effective*, 4.5 = *extremely effective*, 5.0 = *exceptional*). They also rated overall performance using the same rating scale. Each participant made 6 ratings (one for each dimension plus an overall rating) of each videotaped episode, yielding a total of 24 ratings.

Using procedures recommended by Sulsky and Balzer (1988), three experienced raters who had been trained as assessment center assessors and had an average of 3 years of assessment center experience, independently observed and rated the videotaped episodes. The raters then met to achieve consensus. The consensus scores were used as the set of comparison scores for assessing rating accuracy. The experienced raters also completed the performance schema instrument. Following E. A. Day et al.'s (2001) recommendation, we averaged the paired-comparison ratings of these raters to generate a referent schema that served as the comparison standard for evaluating performance schema accuracy.

Although in past research, subject matter experts tended to be selected on the basis of experience alone (e.g., Woehr, 1994), Borman (1987) reported little overlap in the performance schemas of experienced raters. However, training and experience are both important in the development of rating expertise. Therefore, the experienced raters in the present study had previously received intensive 30-hr training over 6 days, followed by annual day-long review training for their roles as assessors. All training, both prior and current, was aimed at developing accurate expert-like (and therefore aligned) performance schemas. Thus, the experienced raters were extremely familiar with the role play exercise and the dimensions being rated.

Rater Training

The frame-of-reference training proceeded according to the following set of procedures outlined by Pulakos (1984, 1986). The trainer read each dimension definition, scale anchors, and example behaviors aloud. Next, participants discussed the information. The trainer then presented and discussed examples of behaviors that represented different levels of performance (i.e., good performance vs. poor performance) on each dimension. Participants then practiced making ratings in response to a practice episode, and they were provided with feedback. Examples of weak and effective performance were highlighted in the practice episode. The entire training session lasted about 45 min.

In the control training condition, the trainer read over each of the dimension definitions, but no other task-specific training was provided to participants in the control training. Instead, they were exposed to a generic video on performance appraisal (e.g., Sulsky & Day, 1992; Woehr, 1994). The control training session also lasted approximately 45 min.

Dependent Variables

Rating accuracy. We assessed rating accuracy using distance accuracy and Borman's (1977) differential accuracy. Distance accuracy indicates the average absolute difference of the 20 participant ratings from the 20 target scores (McIntyre et al., 1984).

Lower scores on this measure represent higher accuracy, whereas higher scores indicate lower levels of accuracy. Borman's differential accuracy was used to assess correlational accuracy between ratings on each dimension and the corresponding target scores across participants. Borman's differential accuracy represents the correlation between participants' ratings for all five dimensions of all four videotapes (i.e., 20 ratings) and the target ratings for all five dimensions of all four videotapes (i.e., 20 ratings). Borman's differential accuracy is the average of the *z*-transformed correlations between each participant's ratings and the target ratings across all five dimensions. Higher scores on the index reflect better rating accuracy. It has been argued that Borman's differential accuracy is an index of rating validity because it provides correlational information and is thus insensitive to distances between ratings and true scores (Sulsky & Balzer, 1988).

We did not consider Cronbach's (1955) four indexes of rating accuracy (elevation, differential elevation, stereotype accuracy, differential accuracy) for use in the present study because these measures decompose overall distance accuracy into ratee and dimension main effects and interactions. Given that the target ratings for our rating stimuli varied only with respect to overall levels of performance (i.e., above-average ratings across all dimensions vs. below-average ratings across all dimensions), we considered overall distance accuracy and correlational accuracy as more appropriate measures of rating accuracy for the present study.

Performance schema accuracy. Performance schema accuracy (PSA) is the degree to which individuals' performance schemas are similar to the performance schema of experienced raters. Participants rated the degree of similarity of 15 behaviors (3 behaviors per dimension) identified as the behaviors representing each dimension most relevant to the role play exercise. Each participant rated the similarity of all of the 105 pairwise comparisons of the 15 behaviors. To eliminate effects due to order of presentation, we presented the 105 pairs in a different random order to each participant. This method for collecting paired comparison data has been used in previous research (e.g., Rentsch, 1990).

Using an 11-point scale ranging from -5 (*very dissimilar*) to $+5$ (*very similar*), participants rated the degree of similarity of the behaviors in each pair. For ease of interpretation in subsequent analyses, we transformed participants' similarity ratings into a scale that ranged from 1 to 11 by uniformly adding 6 to each participant's initial ratings. These data were then subjected to MDS analysis, which is useful for representing knowledge organization (e.g., Forgas, 1981; Rentsch, Heffner, & Duffy, 1994). MDS analysis provides an R^2 value that indicates the variance accounted for by the dimensions produced in the MDS solution (Kruskal & Wish, 1978). R^2 can be interpreted as a goodness-of-fit measure. Values of R^2 range from 0 to 1 with higher values reflecting better fit. To measure PSA, we conducted individual differences Euclidian distance (INDSCAL) MDS analyses using the experienced rater similarity data matrix (representing the referent performance schema) and each participant's similarity data matrix. An individual differences MDS model was chosen because individual difference models provide information on the degree of agreement among raters regarding the organization of stimuli (Schiffman et al., 1981), and this model has been used to assess schema similarity among individuals (Rentsch & Klimoski, 2001). The resulting R^2 value for each participant was operationalized as PSA in subsequent analyses.

To create the referent performance schema, we used MDS to analyze the three similarity data matrices generated by the experienced raters. We applied three decision rules to determine the appropriate dimensionality of the referent schema: (a) No more than five dimensions are likely to be meaningful (Kruskal & Wish, 1978); (b) the number of underlying dimensions should not exceed the number of theoretical dimensions; and (c) at least a 5% increase in variance is required before an additional dimension can be accepted for the solution (Rentsch, 1990). This resulted in a five-dimensional solution that provided the best fit, with a substantial R^2 of .99. Next, consistent with previous research conducted with expert similarity data matrices (e.g., E. A. Day et al., 2001), we averaged the similarity ratings of the three trained and experienced raters to create the experienced rater similarity data matrix used to determine PSA. To provide additional evidence of the validity of the referent data, we conducted a supplementary analysis of the mean differences between similarity ratings of behaviors from the same dimension and ratings of behaviors drawn from different dimensions. As expected, the results of this analysis revealed that the mean similarity rating of behaviors drawn from the same dimension ($M = 10.60$, $SD = 0.74$) was significantly greater than the mean similarity rating of behaviors drawn from different dimensions ($M = 2.46$, $SD = 0.94$), $t(103) = 31.97$, $p < .001$. However, the MDS results, rather than these raw ratings, revealed the underlying dimensions used to cognitively organize the behaviors (Wish & Carroll, 1974).

Declarative knowledge. We assessed declarative knowledge using a behavioral classification measure that required participants to match 15 managerial behaviors to their respective dimensions. Declarative knowledge was operationalized as the total number of correctly classified behaviors (0 to 15) in subsequent analyses. The split-half reliability estimate of .65 was in the acceptable range for research purposes (Murphy & Davidshofer, 2005).

Results and Discussion

Intercorrelations and descriptive statistics for the study variables are reported in Tables 1, 2, and 3. Two-sample tests revealed no significant differences in the two training groups for age, $t(142) = 0.25$, gender, $\chi^2(1) = 3.47$, race, $\chi^2(3) = 0.78$, GPA, $t(142) = 1.12$, or rating experience, $t(142) = 0.02$.

Hypothesis 1a predicted that PSA would be significantly greater after frame-of-reference training than before the training. We tested this using a paired-samples *t* test on the means of the Fisher *z*-transformed square roots of the R^2 values for frame-of-reference–trained participants pre- and post-training. The mean R^2 for the frame-of-reference–trained group was significantly higher after training ($M = .90$, $SD = .06$) than before training ($M = .87$, $SD = .03$), $t(72) = 5.95$, $p < .001$ (one-tailed); Cohen's $d = .90$. There was no significant change in R^2 from pre-training ($M = .87$, $SD = .03$) to post-training ($M = .87$, $SD = .06$) for the control-trained group, $t(70) = .95$, *ns*. Hypothesis 1a was fully supported.

Hypothesis 1b predicted that PSA would be significantly greater for participants in the frame-of-reference training condition than for participants in the control training condition. An independent-samples *t* test was conducted on the means of the Fisher

Table 1
Means, Standard Deviations, and Intercorrelations for Study Variables

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. Gender[a]	1.56	0.50	—							
2. Age	21.44	3.73	.10	—						
3. GPA	3.16	0.41	.05	−.08	—					
4. Rating experience	0.76	1.72	.11	.06	−.04	—				
5. Knowledge score	10.92	2.45	.00	.04	.06	.14	—			
6. Distance accuracy[b]	0.73	0.22	.03	−.07	−.08	.00	−.39**	—		
7. Borman's differential accuracy	0.76	1.72	−.10	−.02	.11	.02	.40**	−.64**	—	
8. Post-training performance schema accuracy	0.89	0.07	−.09	−.02	−.02	.03	.24**	−.26**	.31**	—

Note. $N = 144$. GPA = grade point average; rating experience = total number of times having rated the job performance of another person.
[a] 1 = female; 2 = male. [b] Correlations with distance accuracy are negative because smaller values on this index represent greater accuracy.
** $p < .01$.

z-transformed square roots of the post-training R^2 values for participants in the frame-of-reference and control training conditions. The mean R^2 for the frame-of-reference–trained group ($M = .91$, $SD = .06$) was significantly higher than the mean R^2 for the control-trained group ($M = .87$, $SD = .06$), $t(142) = 4.30$, $p < .001$ (one-tailed); Cohen's $d = .72$. Hypothesis 1b was fully supported.

Hypothesis 2 predicted that frame-of-reference–trained participants would provide more accurate ratings than control-trained participants. A multivariate framework was appropriate for testing this hypothesis (Schleicher et al., 2002). Multivariate analysis of variance, with training (frame of reference vs. control) as the independent variable and the two rating accuracy indexes as the multiple dependent variables, revealed that frame-of-reference–trained participants' ratings were significantly more accurate than those of control-trained participants, $F(2, 141) = 41.75$, $p < .001$; Wilks's $\lambda = .63$; partial $\eta^2 = .37$ (see Table 3). A discriminant analysis revealed one significant eigenvalue ($p < .001$), with training condition accounting for 100% of the variance in the accuracy composite. The structure coefficients from this analysis indicated that distance accuracy and Borman's differential accuracy were driving the discrimination between the different training conditions (.89 and −.83, respectively). Follow-up univariate analyses of variance support the ubiquitous research finding that frame-of-reference training is an effective approach for improving rating accuracy (see Table 4). Hypothesis 2 was fully supported. Frame-of-reference–trained participants ($M = 4.44$, $SD = 0.42$) also used a significantly larger number of performance dimensions to code candidate behaviors on their rating sheets than did control-trained participants ($M = 3.89$, $SD = 0.69$), $t(142) = 5.83$, $p < .001$ (one-tailed); Cohen's $d = .98$. The number of dimensions used by a rater was coded as the number of performance dimensions (0 to 5) for which the rater listed an observed behavior.

Hypothesis 3 predicted that PSA would be positively related to rating accuracy in the frame-of-reference training condition and was fully supported. PSA correlated positively and significantly with both rating accuracy indexes (see Table 2).

Hypothesis 4 predicted that frame-of-reference–trained participants would score higher on a measure of declarative knowledge than control-trained participants. Results indicated that frame-of-reference–trained participants ($M = 11.93$, $SD = 1.89$) scored significantly higher on the declarative knowledge measure than did control-trained participants ($M = 9.89$, $SD = 2.53$), $t(142) = 5.50$, $p < .001$ (one-tailed), Cohen's $d = .92$. Hypothesis 4 was fully supported.

Hypothesis 5 predicted that PSA would account for a unique amount of variance in rating accuracy over and above that of a measure of declarative knowledge. Hierarchical regression analyses on each index of rating accuracy, whereby the declarative knowledge scores were entered into the regression equation as the first step and PSA was entered as the second step, indicated that PSA accounted for a significant amount of unique variance in

Table 2
Intercorrelations for Study Variables by Training Condition

Variable	1	2	3	4	5	6	7	8
1. Gender[a]	—	.01	.16	.26*	.29*	−.12	.03	−.03
2. Age	.18	—	.07	.12	.12	−.17	.17	−.06
3. GPA	−.09	−.17	—	.01	.22	−.24	.21	−.12
4. Rating experience	−.05	.01	−.09	—	.18	.02	.10	.08
5. Knowledge score	−.21	−.05	−.02	.11	—	−.18	.30*	.01
6. Distance accuracy[b]	.02	.03	−.05	−.04	−.27*	—	−.57**	−.02
7. Borman's differential accuracy	−.07	.22	.17	−.09	.12	−.29**	—	.08
8. Post-training performance schema accuracy	−.05	−.01	.08	.01	.21*	−.21*	.25*	—

Note. Data for frame-of-reference participants ($n = 73$) are below the diagonal, and data for control participants ($n = 71$) are above the diagonal. GPA = grade point average; rating experience = total number of times having rated the job performance of another person.
[a] 1 = female; 2 = male. [b] Correlations with distance accuracy are negative because smaller values on this index represent greater accuracy.
* $p < .05$. ** $p < .01$.

Table 3
Means and Standard Deviations of Study Variables by Training Condition

	Frame of reference (*n* = 73)		Control (*n* = 71)		
Variable	*M*	*SD*	*M*	*SD*	*p*
1. Gender[a]	1.48	0.50	1.63	0.49	*ns*
2. Age	21.52	4.33	21.37	3.02	*ns*
3. GPA	3.12	0.41	3.20	0.41	*ns*
4. Rating experience	0.77	1.56	0.76	1.88	*ns*
5. Knowledge score	11.93	1.89	9.89	2.53	<.001
6. Distance accuracy	0.61	0.12	0.86	0.23	<.001
7. Borman's differential accuracy	1.07	0.44	0.44	0.55	<.001
8. Post-training performance schema accuracy[b]	0.91	0.06	0.87	0.06	<.001

Note. GPA = grade point average; rating experience = total number of times having rated the job performance of another person. For distance accuracy, smaller numbers represent greater accuracy. For Borman's differential accuracy, larger numbers represent greater accuracy.
[a] 1 = female; 2 = male. [b] Performance schema accuracy refers to post-training values.

distance accuracy and Borman's differential accuracy over and above that of declarative knowledge (see Table 5). Supplementary analyses also revealed that the reverse was true; declarative knowledge accounted for a significant amount of unique variance in distance accuracy and Borman's differential accuracy over and above that of PSA. Thus, PSA was found to be a unique and meaningful variable in the prediction of rating accuracy; therefore, full support was found for Hypothesis 5.

Contributions of the Present Study

The results of the present study extend previous work that was limited to indirect study of rater schemas. In the present study, we operationalized the organization of recalled ratee behaviors with a structural assessment technique using paired comparisons that allowed for an evaluation of raters' performance schema accuracy relative to a referent model. Previous studies of the cognitive effects of frame-of-reference training also have been limited because they did not directly compare the trainees' and experienced raters' cognitive variables. We addressed rater cognition directly. Perhaps this will prompt rater training researchers to consider experienced raters' cognitive structures as potential resources for evaluating the cognitive effects of training. Furthermore, previous studies inferred change on the basis of training–control differences. The present study provided direct evidence that frame-of-reference training increases PSA.

Implications for Future Research on Frame-of-Reference Training

Future research should examine the long-term effects of frame-of-reference training on rater schemas. For example, Sulsky and Day (1994) found that frame-of-reference–trained raters rated more accurately than control-trained raters after a 48-hr delay, and Roch and O'Sullivan (2003) found no significant decay in rating accuracy 2 weeks after frame-of-reference training. On the basis of the results of the present study, these findings might exist because frame-of-reference training develops relatively stable performance schemas. Future studies should consider the temporal stability of schema accuracy.

As with any study, one must be cautious in generalizing results. The present sample consisted of novice raters in a laboratory setting, who were not accountable, had a low level of psychological involvement in the effects of their ratings, and likely had little rating experience relative to raters in organizations. Although the participants' experience levels were likely lower than those of raters in organizations and their motivation was likely different, our results indicated that they did benefit from the frame-of-reference training. Furthermore, the candidates that were rated in the present study were demographically homogeneous (e.g., age, race) and may not represent the population of ratees on the whole. However, the present results may generalize to organizational settings, because objective evidence has indicated that even expe-

Table 4
Analysis of Variance Results for the Two Rating Accuracy Indexes

	Condition				
Accuracy	Frame of reference	Control	*F*(1, 142)	*p*	R^2
Distance accuracy	0.61	0.86	66.23	<.001	.32
Borman's differential accuracy	1.07	0.44	57.83	<.001	.29

Note. *N* = 144. For distance accuracy, smaller numbers represent greater accuracy. For Borman's differential accuracy, larger numbers represent greater accuracy.

Table 5
Regression Results for the Incremental Validity of Post-training Performance Schema Accuracy

Accuracy index	β	R	R^2	ΔR^2
Distance accuracy				
Step 1				
Declarative knowledge	−.39	.39	.15	
Step 2				
Post-training performance schema accuracy	−.18	.42	.18	.03*
Borman's differential accuracy				
Step 1				
Declarative knowledge	.40	.40	.16	
Step 2				
Post-training performance schema accuracy	.23	.46	.21	.05**

Note. $N = 144$.
* $p < .05$. ** $p < .01$.

rienced raters in an organization possessed performance schemas that showed little overlap in their content (Borman, 1987). These findings suggest that frame-of-reference training may benefit experienced raters in organizations to develop accurate performance schemas. Although laboratory research can contribute to understanding the effects of frame-of-reference training, it should be complemented by field research.

In the present study, the declarative knowledge assessment and the performance schema assessment were designed to contain corresponding content related to behavioral dimensions. PSA was the primary outcome of interest in the present study; therefore, although behavioral effectiveness was presented in the frame-of-reference training, it was not explicitly assessed in either measure. Nevertheless, frame-of-reference training had a significant effect on both measures. Future frame-of-reference training and performance schema research should place additional emphasis on the behavioral effectiveness aspect of the performance domain and measure it directly.

Overall, the results suggested that PSA should be considered a meaningful outcome variable of frame-of-reference training, which is consistent with previous studies (e.g., Davis, Curtis, & Tschetter, 2003; Dorsey et al., 1999; Stout, Salas, & Kraiger, 1997). Frame-of-reference training researchers should consider incorporating performance schema measures as training criteria in addition to traditional indexes of rating accuracy. Moreover, Bernardin and Buckley (1981) originally proposed such training as a method for identifying raters with idiosyncratic frames of reference, a suggestion that has largely been ignored by frame-of-reference training researchers (Hauenstein & Foti, 1989; for an exception, see Uggerslev & Sulsky, 2008). One reason for this apparent oversight may be the lack of a standardized method for identifying idiosyncratic raters. Pre-training PSA could identify such individuals, and the schema measurement technique used in the present study provides a useful approach because paired-comparison ratings are effective in eliciting idiosyncratic understanding of stimuli (Mohammed et al., 2000). Additional research in this area could be directed toward making an empirical connection between rating idiosyncrasy and schema idiosyncrasy.

With respect to the practical issues associated with frame-of-reference training, although research evidence supports the effectiveness of such training, there is little evidence that it has been used extensively in applied settings (Bernardin, Buckley, Tyler, & Wiese, 2001; for an exception, see Noonan & Sulsky, 2001). Researchers have identified some potential reasons for this, arguing that frame-of-reference training is too time-consuming and expensive for organizations (Stamoulis & Hauenstein, 1993) and that the process of developing target scores is likewise too complex and time-consuming for organizations (Bernardin et al., 2001; Ilgen & Favero, 1985). However, Chirico et al. (2004) provided evidence that true score feedback may not be necessary for a frame-of-reference training program to be effective at improving rating accuracy. These authors compared two frame-of-reference training groups: one that received true score feedback and another that received only qualitative feedback to a control group. Chirico et al. found that traditional frame-of-reference training and frame-of-reference training with only qualitative feedback were effective in improving rating accuracy in comparison with a control condition, but there were no significant differences in rating accuracy measures between the two conditions. This finding, along with research on schema training, suggests that frame-of-reference training protocols may be adjusted to suit the needs of an organization while maintaining the integrity of the training principles.

Conclusion

Previous research has found consistently positive effects of frame-of-reference training for improving rating accuracy. Many researchers have recognized the need for a better understanding of the cognitive mechanisms involved in frame-of-reference training, and consequently, numerous studies have been devoted to examining cognitive issues such as rater memory and recall for performance-related information. Despite the encouraging results of these studies, they have neglected to account for the positive effects of frame-of-reference training in situations in which memory and recall are not relied on heavily (e.g., the recent interest in applying frame-of-reference training to assessment center rating situations; Schleicher et al., 2002). The use of direct structural assessment techniques for measuring performance schemas is highly appropriate for these contexts. The results of the present study are only the first step toward attaining a more complete picture of the complex cognitive mechanisms that underlie rating accuracy.

References

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.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, England: Cambridge University Press.

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

Bernardin, H. J., Buckley, M. R., Tyler, C. L., & Wiese, D. S. (2001). A reconsideration of strategies for rater training. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (pp. 221–274). Stamford, CT: JAI Press.

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

Billings, R. S., & Cornelius, E. T. (1980). Dimensions of work outcomes:

A multidimensional scaling approach. *Personnel Psychology, 33,* 151–162.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20,* 238–252.

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an Army office sample. *Organizational Behavior and Human Decision Processes, 40,* 307–322.

Cannon-Bowers, J. A., Tannenbaum, S. L., Salas, E., & Converse, S. A. (1991). Toward an integration of training theory and technique. *Human Factors, 33,* 281–292.

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

Carlston, D. E. (1992). Impression formation and the modular mind: The associated systems theory. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 301–341). Hillsdale, NJ: Erlbaum.

Carlston, D. E. (1994). Associated systems theory: A systematic approach to cognitive representations of persons. *Advances in Social Cognition, 7,* 1–78.

Chirico, K. E., Buckley, 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.

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

Davis, M. A., Curtis, M. B., & Tschetter, J. D. (2003). Evaluating cognitive training outcomes: Validity and utility of structural knowledge assessment. *Journal of Business and Psychology, 18,* 191–206.

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,* 158–167.

Day, E. A., Arthur, W., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology, 86,* 1022–1033.

Dorsey, D. W., Campbell, G. E., Foster, L. L., & Miles, D. E. (1999). Assessing knowledge structures: Relations with experience and post-training performance. *Human Performance, 12,* 31–57.

Forgas, J. P. (1981). Social episodes and group milieu: A study in social cognition. *British Journal of Social Psychology, 20,* 77–87.

Goldsmith, T. E., & Kraiger, K. (1997). Applications of structural knowledge assessment to training evaluation. In J K. Ford (Ed.). *Improving training effectiveness in work organizations* (pp. 73–96). Mahwah, NJ: Erlbaum.

Goodstone, M. S., & Lopez, F. E. (2001). The frame of reference approach as a solution to an assessment center dilemma. *Consulting Psychology Journal: Practice and Research, 53,* 96–107.

Hauenstein, N. M. A., & Foti, R. J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. *Personnel Psychology, 42,* 359–379.

Howell, W. C., & Cooke, N. J. (1989). Training the human information processor: A review of cognitive models. In Goldstein, I. L. (Ed.). *Training and development in organizations,* (pp. 121–182). San Francisco: Jossey-Bass.

Ilgen, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to performance appraisal processes. *Academy of Management Review, 10,* 311–321.

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge.* Hillsdale, NJ: Erlbaum.

Jones, L. E. (1983). Multidimensional models of social perception, cognition, and behavior. *Applied Psychological Measurement, 7,* 451–472.

Klein, S. B., & Loftus, J. (1990). Rethinking the role of organization in person memory: An independent trace storage model. *Journal of Personality and Social Psychology, 59,* 400–410.

Koubek, R. J., Clarkston, T. P., & Calvez, V. (1994). The training of knowledge structures for manufacturing tasks: An empirical study. *Ergonomics, 37,* 765–780.

Kozlowski, S. W. (1998). Training and developing adaptive teams: Theory, principles, and research. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 115–153). Washington, DC: American Psychological Association.

Kraiger, K., Salas, E., & Cannon-Bowers, J. A. (1995). Measuring knowledge organization as a method for assessing learning during training. *Human Factors, 4,* 804–816.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling.* Beverly Hills. CA: Sage.

Leinhardt, G., & Smith, D. A. (1985). Expertise in mathematics instruction: Subject matter knowledge. *Journal of Educational Psychology, 77,* 247–271.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86,* 255–264.

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.

Mohammed, S., Klimoski, R., & Rentsch, J. R. (2000). The measurement of team mental models: We have no shared schema. *Organizational Research Methods, 3,* 123–165.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications.* Upper Saddle River, NJ: Pearson Education.

Neff, C. L. (1983). *Maternal sensitivity to infant signals as related to quality of infant–mother attachment.* Unpublished master's thesis, University of Illinois at Urbana-Champaign.

Noble, D. F. (1989). Schema-based knowledge elicitation for planning and situation assessment aids. *IEEE Transactions on Systems, Man, and Cybernetics, 19,* 473–482.

Noonan, L. E., & Sulsky, L. M. (2001). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance, 14,* 3–26.

Pollard-Gott, L. (1983). Emergence of thematic concepts in repeated listening to music. *Cognitive Psychology, 15,* 66–94.

Poole, P. P., Gray, B., & Gioia, D. A. (1990). Organizational script development through interactive accommodation. *Group and Organization Studies, 15,* 212–232.

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

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.

Rentsch, J. R. (1990). Climate and culture: Interaction and qualitative differences in organizational meanings. *Journal of Applied Psychology, 75,* 668–681.

Rentsch, J. R., & Hall, R. J. (1994). Members of great teams think alike: A model of team effectiveness and schema similarity among team members. In M. M. Beyerlein & D. A. Johnson (Eds.), *Advances in interdisciplinary studies of work teams: Vol. 1. Series on self-managed work teams* (pp. 223–262). Greenwich, CT: JAI Press.

Rentsch, J. R., Heffner, T. S., & Duffy, L. T. (1994). What you know is what you get from experience: Team experience related to teamwork schemas. *Group and Organization Management, 19,* 450–474.

Rentsch, J. R., & Klimoski, R. J. (2001). Why do 'great minds' think alike?

Antecedents of team member schema agreement. *Journal of Organizational Behavior, 22,* 107–120.

Roch, S. G., & O'Sullivan, B. J. (2003). Frame of reference rater training issues: Recall, time and behavior observation training. *International Journal of Training and Development, 7,* 93–107.

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for estimation of clustering in free recall. *Psychological Bulletin, 76,* 45–48.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods and applications.* New York: Academic Press.

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.

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,* 735–746.

Shoben, E. J. (1983). Applications of multidimensional scaling in cognitive psychology. *Applied Psychological Measurement, 7,* 473–490.

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,* 179–194.

Stamoulis, D. T., & Hauenstein, N. M. A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology, 78,* 994–1003.

Stout, R. J., Salas, E., & Kraiger, K. (1997). Role of trainee knowledge structures in aviation team environments. *International Journal of Aviation Psychology, 7,* 235–250.

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.

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.

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.

Sulsky, L. M., & Kline, T. B. (2007). Understanding frame-of-reference training success: A social learning theory perspective. *International Journal of Training and Development, 11,* 121–131.

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,* 711–719.

Wish, M., & Carroll, J. D. (1974). Applications of individual differences scaling to studies of human perception and judgment. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2, pp. 449–491). New York: Academic Press.

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.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67,* 189–205.

Wyer, R. S., & Srull, T. K. (1989). *Memory and cognition in its social context.* Hillsdale, NJ: Erlbaum.

Received March 6, 2008
Revision received May 7, 2009
Accepted May 11, 2009 ■