5-2008

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

Charles A. Gorman
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Charles A. Gorman entitled "Evaluating Frame-of-Reference Rater Training Effectiveness via Performance Schema Accuracy." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial and Organizational Psychology.

Joan R. Rentsch, Major Professor

We have read this dissertation and recommend its acceptance:

David J. Woehr, R. Tom Ladd, Michelle Violanti

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Charles Allen Gorman entitled "Evaluating Frame-of-Reference Rater Training Effectiveness via Performance Schema Accuracy." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial/Organizational Psychology.

Joan R. Rentsch
Major Professor

We have read this dissertation
And recommend its acceptance:

David J. Woehr

R. Tom Ladd

Michelle Violanti

Acceptance for the Council:

Carolyn R. Hodges
Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

EVALUATING FRAME-OF-REFERENCE RATER TRAINING EFFECTIVENESS VIA PERFORMANCE SCHEMA ACCURACY

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Charles Allen Gorman
May 2008

Dedication

This dissertation is dedicated to the two loves of my life: my baby boy Jackson and his momma Annie. Without the two of you, none of this would be worth it.

Acknowledgments

I wish to thank all those who helped me complete my Doctor of Philosophy degree in Industrial/Organizational Psychology. I would like to thank Joan Rentsch for her help and guidance on this project. I would also like to thank the other members of my committee: Dave Woehr, Tom Ladd, and Michelle Violanti. Special thanks also go to my fellow graduate students for all their help and encouragement, including Lisa Delise, Carrie Blair, Joy Oliver, John Meriac, Josh Ray, Wes Davenport, and Melissa Zullo.

Abstract

Frame-of-reference (FOR) 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 FOR training, few studies have empirically addressed the ultimate goal of FOR training, which is to train raters to share a common conceptualization of performance (Athey & McIntyre, 1987; Woehr, 1994). The present study tested the hypothesis that FOR-trained raters would possess schemas of performance after training that are more similar to an expert schema than would control-trained raters. It was also hypothesized that schema accuracy would be positively related to rating accuracy. Results supported these hypotheses. Implications for FOR training research and practice are discussed.

Table of Contents

List of Tables

Table Page

I. Introduction

The evaluation of human performance in work settings has long been an interest of psychological researchers (Arvey & Murphy, 1998). Typically, human performance in organizations is evaluated using subjective performance ratings provided by the employee's supervisor(s), peers, and/or subordinates. The accuracy of these ratings is important to the success of a performance rating system, and some researchers have suggested that rating accuracy is the primary goal of performance evaluation (e.g., Werner & Bolino, 1997). Rating accuracy is typically evaluated by comparing an individual's ratings across dimensions to ratings made by expert raters (i.e., "true" scores). The closer these ratings are to the true score, the more accurate they are believed to be (Sulsky & Balzer, 1988).

Two general strategies have been advanced as ways of improving rating accuracy: rating scale development and rater training (Woehr & Huffcutt, 1994). With regard to rating scale development, the general finding from this literature was that the type of rating scale used made little difference in terms of improving ratings (Landy & Farr, 1980). Thus, recent research has tended to focus more on rater training as an intervention for improving the accuracy of performance ratings.

Purpose of Investigation

Despite the recent focus on the cognitive operations involved in rater training (e.g., Roch & O'Sullivan, 2003; Schleicher & Day, 1998; Sulsky & Kline, 2007), there has been surprisingly little attention paid to how raters cognitively structure performance information presented during training, or more importantly, the accuracy of these cognitive structures. The goal of the present study was to gain an improved

understanding of the cognitive changes that occur as a result of rater training by examining the efficacy of performance schema accuracy as a measure of frame-of-reference rater training effectiveness.

Overview of Manuscript

This manuscript will begin with a brief discussion of the concept of rater training in general, followed by an introduction to a specific type of rater training, frame-of-reference training. Next, the idea of performance schema accuracy will be discussed as a tool for examining the cognitive changes that have been hypothesized to occur as a result of such training. Furthermore, specific and testable hypotheses will be offered with respect to the effects of frame-of-reference training on performance schema accuracy and rating accuracy. Then, the methods utilized in the present study will be addressed, followed by a summary of the study results. Finally, an interpretation of the study results will be presented, in addition to study limitations and future research directions.

Rater Training

Training raters to improve the accuracy of their ratings has been a major focus of research on performance ratings (Smith, 1986). In general, rater training has been shown to be effective (Spool, 1978) and has shown some promise for improving the accuracy of performance ratings (Woehr & Huffcutt, 1994). One of the first references to rater training in the literature is credited to Bittner (1948), who noted that training provided to American army officers on the performance dimensions of the military evaluation scale improved officers' ratings of their soldiers' performance. McIntyre, Smith, and Hassett (1984) identified two major benefits of rater training: (a) to enhance raters' knowledge and skills for carrying out evaluations, and (b) to motivate raters to use the knowledge

and skills learned in the training program. Researchers have also found that employee perceptions of fairness, accuracy, and credibility of the performance rating process and the rater were positively affected by rater training (e.g., Bannister, 1986: Fulk, Brief, & Barr, 1985). Perhaps of even greater benefit, Werner and Bolino (1997) found that court judges showed some preference for performance rating systems that included rater training programs.

Woehr and Huffcutt's (1994) quantitative review identified four general approaches to rater training based on the content of the training: (a) rater error training, (b) performance dimension training, (c) behavioral observation training, and (d) frame-of-reference training. Of these four approaches, frame-of-reference (FOR) training has received a considerable amount of recent research attention due to its relative effectiveness at improving rating accuracy.

Frame-of-Reference Training

FOR training is one of several training approaches that developed as a reaction to the inconsistent results of rater error training. Rater error training requires raters to recognize leniency, halo, and central tendency errors and avoid making these errors in future ratings. However, although rater error training resulted in fewer leniency and halo errors, it inadvertently lowered levels of rating accuracy (Bernardin & Pence, 1980; Landy & Farr, 1980; Smith, 1986). Others have suggested that rater error training actually produces a meaningless redistribution of ratings (Smith, 1986) and that rater errors may not be errors, but rather rater effects that reflect true variance (Arvey & Murphy, 1998; Hedge & Kavanagh, 1988). Moreover, Arvey and Murphy (1998)

suggested that rater errors are relatively unimportant and trivial when it comes to rating accuracy.

In response, Bernardin and Buckley (1981) proposed FOR training as an alternative to rater error training. FOR training focuses on providing raters with performance standards for each dimension to be rated (Woehr & Huffcutt, 1994). Specifically, FOR 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 FOR training is to train raters to share and use common conceptualizations of performance when providing their ratings (Athey & McIntyre, 1987; Woehr, 1994). Accordingly, an abundant number of studies have demonstrated the effectiveness of FOR training for improving rating accuracy (Athey & McIntyre, 1987; Bernardin & Pence, 1980; Cardy & Keefe, 1994; Day & Sulsky, 1995; Hauenstein & Foti, 1989; McIntyre et al, 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 FOR training compared to control or no training groups.

More recently, research on FOR training has focused on the application of FOR training methods for use in the training of assessment center (AC) assessors in the hopes of improving AC construct validity (e.g., Goodstone & Lopez, 2001; Lievens, 2001; Schleicher, Day, Mayes, & Riggio, 2002). This line of research has found generally positive results. For example, Lievens (2001) found that interrater reliability, rating accuracy, and discriminant validity were better for AC assessors in a FOR training

condition than assessors in control or data-driven training conditions. Likewise, in a study of 58 assessees and 122 assessors, Schleicher et al. (2002) found the FOR training was effective at improving the reliability, accuracy, convergent and discriminant validity, and criterion-related validity of AC ratings.

Cognitive Models of FOR Training Effectiveness

In an effort to explain why FOR training increases rating accuracy, many 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 (Day & Sulsky, 1995). Taken together, these models suggest that FOR training works by influencing how ratee information is processed and represented in raters' memories. The primary evidence pointed to by FOR researchers in support of these models has been based on analyses of recalled performance information. Typically, participants are asked to recall as many behaviors as they can remember after having watched a number of simulations of ratee performance. The organization of recalled information can then be examined using various indexes such as the adjusted ratio of clustering (ARC; Roenker, Thompson, & Brown, 1971), which assesses the extent to which behaviors representing the same performance dimensions are recalled in clusters compared to the amount of clustering expected by chance alone. ARC scores of 1.0 represent perfect clustering, scores of 0.0 represent chance clustering, and negative scores indicate a clustering scheme other than the one being assessed.

Despite the continued use of clustering indexes such as the ARC, these indexes convey a limited amount of information regarding knowledge organization. For example, ARC scores are based only on the order in which behaviors are recalled. More sophisticated data reduction techniques, such as multidimensional scaling, allow raters to make their own judgments about the interrelationships among behaviors regardless of the order in which they recall the information. The central premise of FOR training is to train raters to share and use common conceptualizations of performance when making ratings. This has only been partially tested. Furthermore, no research has examined the extent to which FOR training improves the *accuracy* of performance knowledge structures. The present study seeks to extend the FOR training literature by highlighting how measuring the accuracy of performance knowledge structures (or schemas) can be instrumental in evaluating the success of FOR training. The following section will introduce the concept of a schema and detail its utility as a training outcome variable.

Schemas

The study of knowledge structures is nothing new to the expert-novice literature. Within this literature, several terms for knowledge structures have been used, including 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 used to organize new information and facilitate understanding (Noble, 1989; Poole, Gray, & Gioia, 1990). With advances in learning and domain-relevant experience, the organization of the schema 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). Expert schemas enable individuals to recognize the similarity between new and previously experienced situations and to adapt old procedures for new situations (Noble, 1989).

The schema perspective has been especially influential in the team cognition literature. The team schema approach developed from simultaneous research on organizational climate, culture, and sense making. Simultaneously, research was developing related to shared mental models based on human factors research. A shared mental model is an organized mental representation of knowledge that is shared among a group of individuals (Cannon-Bowers, Salas, & Converse, 1993; Klimoski & Mohammed, 1994). The shared mental models approach proposes that greater similarity of individuals' mental models leads to greater shared expectations within a team, which in turn leads to superior team performance (Rouse, Cannon-Bowers, & Salas, 1992).

Rentsch and Hall (1994) recognized that the term *shared mental models* inadvertently suggests that individuals' mental models must be identical to be shared. Thus, the authors introduced the term *schema similarity*, which refers to the degree to which individuals have similar knowledge structures for organizing and understanding concepts (Rentsch & Klimoski, 2001). The schema similarity approach proposes that individuals' knowledge structures will become more similar over time with relevant experience, which then leads to greater team effectiveness (Rentsch & Hall, 1994). There is some research evidence that supports the notion of schema similarity. Rentsch, Heffner, and Duffy (1994), for example, found that more experienced team members

conceptualized teamwork more precisely and in more abstract terms than less experienced team members. Similar results were reported in a study by Smith-Jentsch, Campbell, Milanovich, and Reynolds (2001), who noted that more experienced navy personnel had more similar schemas than did less experienced navy personnel. Mathieu, Heffner, Goodwin, Salas, and Cannon-Bowers (2000) observed that team schema similarity was related to subsequent team process and performance. Moreover, Rentsch and Klimoski (2001) reported that demography, team experience, team member recruitment, and team size were significantly related to team member schema agreement, which in turn was related to team effectiveness.

Schema Accuracy and Training

If the goal of training is to create experts in the domain of interest, then it would seem beneficial to utilize schemas as training criteria (Cannon-Bowers et al., 1991). To this end, researchers have demonstrated that individual schemas can be manipulated through training (e.g., Koubek, Clarkston, & Calvez, 1994), and that expert schema similarity (or schema accuracy) can be used as a measure of learning during training. For example, in a training program for computer programming and naval decision making, Kraiger, Salas, and Cannon-Bowers (1995) found that trainees' schemas were significantly more similar to an expert schema after training than before. Moreover, using a card sorting technique, Smith-Jentsch et al., (2001) noticed that higher ranking navy personnel held mental models of teamwork that were more similar to an empirically derived model of expert team performance than lower ranking personnel. Furthermore, in a study of college students, Day, Arthur, and Gettman (2001) observed that similarity

of trainees' schemas to an expert schema was correlated with skill acquisition and was predictive of skill retention and transfer.

Not only have schemas been shown to be useful as training criteria, but there is some evidence that schema measures convey unique information related to training not available in traditional measures of learning (Stout, Salas, & Kraiger, 1997). A study by Davis, Curtis, and Tschetter (2003), for example, indicated that schema assessment predicted performance self-efficacy over and above declarative knowledge. Likewise, Dorsey, Campbell, Foster, and Miles (1999) found that schema measures contain unique variance that does not overlap with traditional measures of declarative knowledge.

The Present Study

The concept of schema accuracy holds great promise for the study and application of FOR training. Although the hallmark of FOR training is the development of a common view of performance that is shared by all raters (Goodstone & Lopez, 2001), there is little evidence that researchers have attempted to measure this shared view of performance. Thus, the research findings discussed in the previous section have clear implications for the current research proposal.

First, the schema similarity approach suggests that individuals' schemas will become more similar over time with advances in learning (Rentsch & Hall, 1994). Therefore,

> *Hypothesis 1a*: Individuals who receive FOR training will have performance schemas more similar to an expert schema (i.e., more accurate) after training than before training.

In addition, schema similarity research indicates that individuals with more experience on 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 et al., 2001). Hence,

Hypothesis 1b: Individuals who receive FOR training will possess performance schemas that are more similar to an expert schema (i.e., more accurate) than will individuals who receive control training.

Second, previous research indicates that FOR training is an effective intervention for improving rating accuracy (e.g., Woehr & Huffcutt, 1994). Thus,

Hypothesis 2: Performance ratings from those who receive FOR training will be more similar to expert ratings (i.e., more accurate) than will performance ratings from those who receive control training.

Third, if FOR 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,

Hypothesis 3: Five measures of rating accuracy will be positively related to performance schema accuracy.

Fourth, prior research has revealed that FOR training improves raters' knowledge of performance-related information (e.g., Woehr, 1994). Consequently,

Hypothesis 4: Individuals who receive FOR training will score significantly higher on a measure of declarative knowledge than will those who receive control training.

Finally, research suggests that schema measures contain unique variance that does not overlap with traditional measures of declarative knowledge (e.g., Dorsey et al., 1999). Therefore,

> *Hypothesis 5*: Performance schema accuracy will account for a unique amount of variance in all five measures of rating accuracy over and above that of a measure of declarative knowledge.

II. Method

Participants

One hundred forty-four undergraduate students at a large southeastern university were solicited to participate in this study. Fifty-six percent of the participants were male, and 90 percent of the participants identified themselves as Caucasian. Sixty percent of participants held at least a part-time job, and 77 percent of participants had no experience rating the job performance of another person. Participants were randomly assigned to either a FOR-training condition (n = 73) or a control-training condition (n = 71). All participants were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 2002).

Procedure

Participants who volunteered to take part in the study were telephoned the evening prior to their scheduled date of participation to remind them of the time and place of the study. Sessions were randomly divided into FOR training or control training conditions, and the attendance of each session ranged from 3 to 10 participants. The videotaped episodes were presented at individual computer terminals. Participants were informed that the purpose of the study was to examine the way people evaluate work performance. Before training, participants received a brief introduction to the session, after which they completed a pre-training performance schema measure. Next, participants received either FOR or control training. After training, participants completed a measure of declarative knowledge and a post-training schema measure. Participants then viewed four videotaped performance episodes (described below) that were presented in random order across individual participants. During the presentation of

the videotapes, participants recorded specific behaviors as they observed them on a rating form. At the conclusion of each performance episode, participants recorded their ratings in the spaces provided on the form. Upon viewing and rating all of the episodes, participants completed a demographic questionnaire (extracted and adapted from Organizational Research Group, 1998). (See Appendix A). At the conclusion of the session, each participant was debriefed, thanked, and dismissed. See Figure 1 on page 14 for a timeline of the study methods.

Stimulus Materials

The performance episodes that served as the stimuli in the present study consisted of videotaped performance episodes from a previously conducted developmental assessment center at a large southeastern university. The videotapes depicted a role play exercise in which an assessment center candidate assumes the role of a manager and interacts with a subordinate, played by a trained assessor. See Appendix B for a character sketch of the role player in the assessment center exercise (extracted and adapted from Tennessee Assessment Center, 2002). The exercises were designed to elicit behaviors from the candidate that can be grouped into the following performance dimensions: Analysis, Decisiveness, Leadership, Confrontation, and Sensitivity. See Appendix C for dimension definitions and behavioral examples (extracted and adapted from Tennessee Assessment Center, 2002). These videotapes were rated by subject matter experts in order to develop comparison performance ratings (described below).

The candidates that appeared in the videotapes were executives enrolled in the same class of a professional MBA program at a large southeastern university. These candidates participated in the developmental assessment center as part of their first year

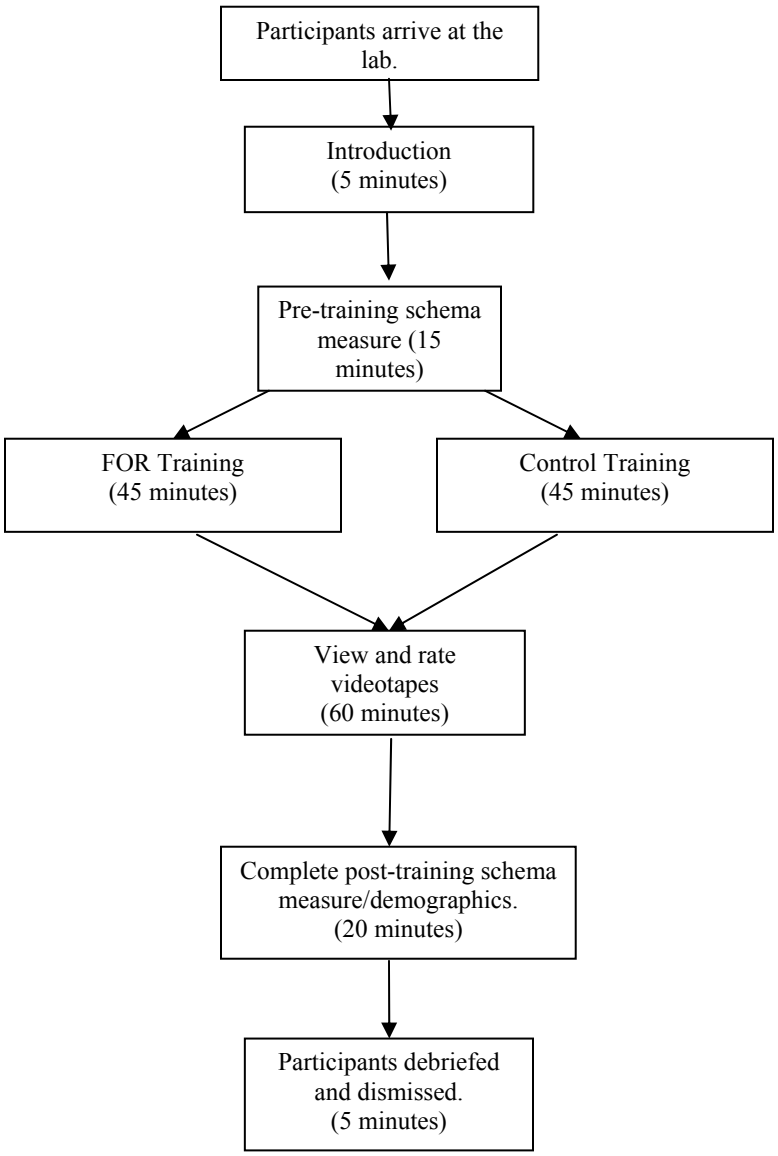

Figure 1. Timeline of Study Methods

curriculum. To control for the possibility of confounding effects due to candidate performance level and sex, each participant viewed 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).

To test for the possibility of confounding differences in the performance episodes due to candidate attractiveness, a pilot session was conducted in which six undergraduate participants viewed videotaped images of the candidates. While viewing the images, the participants responded to six items pertaining to the attractiveness of the candidates using a Likert-type scale ranging from 1 (*Disagree very much*) to 7 (*Agree very much*). (See Appendix D). The results of this pilot session revealed no significant differences between the candidates in terms of their attractiveness ratings, $F(3, 20) = 1.27$, *ns*. See Table 1 on page 16 for the means and standards deviations of the attractiveness ratings by candidate.

Rating Form and Comparison Scores

The rating form consisted of a blank sheet of paper with spaces to record the ratings for each dimension. (See Appendix E). Participants recorded candidate behaviors on their rating 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 each videotape, participants recorded their rating for each dimension in the spaces provided. Each dimension was rated using an 11-point Likert-type rating scale adapted from Tennessee Assessment Center (2002) (1.0 = *extremely weak* to 5.0 = *exceptional*). (See

Table 1

Means and Standard Deviations of Attractiveness Ratings by Candidate

Candidate	*M*	*SD*
Male/Below Average Ratings	4.28	.46
Male/Above Average Ratings	5.00	1.07
Female/Below Average Ratings	3.92	1.06
Female/Above Average Ratings	4.75	1.40

Appendix F). An overall evaluation scale was also included (an 11-point Likert-type scale ranging from 1.0 = *extremely weak* to 5.0 = *exceptional*).

In order for rating accuracy to be measured, a set of comparison scores was needed. Thus, using procedures recommended by Sulsky and Balzer (1988), three upper level graduate students in industrial and organizational psychology serving as subject matter experts (SMEs) independently observed and rated the videotaped episodes. Each of the SMEs was a trained assessment center assessor and thus, intimately familiar with the role play exercise and the dimensions being rated. After independently rating the performances, the SMEs met to discuss rating differences and, through consensus, generated a set of comparison scores. See Table 2 on page 17 for the consensus ratings for each dimension of each episode.

In addition to providing expert ratings, the SMEs also completed the performance schema instrument (described below). Following the recommendation of Day et al., (2001), the experts' schema ratings were averaged to generate a referent schema that served as the comparison for evaluating performance schema accuracy.

Table 2

Means and Standard Deviations of Performance Ratings by Dimension

Dimension	Consensus Expert Rating	FOR (n = 73)		Control (n = 71)	
		M	*SD*	*M*	*SD*
Analysis					
Male/Below Average	2.7	2.34	.53	2.80	1.02
Male/Above Average	4.0	3.79	.64	3.92	.80
Female/Below Average	3.5	3.72	.76	3.78	.96
Female/Above Average	3.7	3.49	.61	3.26	.99
Decisiveness					
Male/Below Average	2.7	2.40	.81	3.02	1.20
Male/Above Average	3.5	3.60	.41	3.44	.90
Female/Below Average	2.7	2.63	.82	3.03	1.12
Female/Above Average	3.7	3.90	.65	3.28	1.08
Leadership					
Male/Below Average	2.7	2.21	.75	2.72	.98
Male/Above Average	3.7	3.91	.67	3.74	.88
Female/Below Average	2.7	2.86	.69	3.00	1.11
Female/Above Average	4.0	3.82	.72	3.39	1.02
Confrontation					
Male/Below Average	2.7	2.15	.67	2.28	.87
Male/Above Average	4.0	3.54	.57	3.51	1.02
Female/Below Average	2.5	3.35	.62	3.46	1.16
Female/Above Average	4.0	3.76	.72	3.38	1.12
Sensitivity					
Male/Below Average	3.5	2.80	.93	3.69	1.06
Male/Above Average	3.7	3.89	.64	3.78	.88
Female/Below Average	3.0	2.59	.96	2.92	1.17
Female/Above Average	3.7	3.83	.76	3.45	.99

Note. FOR = frame of reference.

Rater Training

Participants were randomly assigned to either FOR or control training sessions. All training sessions were conducted by the author using a standard written set of procedures.

FOR training. See Appendix G for the script that was used by the experimenter for the FOR training condition. The FOR training proceeded according to the following set of procedures outlined by Pulakos (1984, 1986): (a) Participants are told that they will evaluate the performance of ratees on separate performance dimensions.

(b) Participants are given rating scales and instructed to read them as the trainer reads the dimension definitions and scale anchors aloud.

(c) The trainer discusses ratee behaviors that illustrate different performance levels for each scale.

(d) Participants are shown a videotape of a practice vignette and are asked to evaluate the ratee using the scales provided.

(e) Ratings are written on a blackboard and discussed by the group of participants.

(f) The trainer provides feedback to participants explaining why the ratee should receive a particular rating (target score) on a given dimension.

Accordingly, participants in the FOR training condition were informed that they would be evaluating job performance on the five performance dimensions. The participants were given a copy of the rating form and the trainer read the definition of each dimension and the scale anchors aloud. Next, participants were read a partial list of example

behaviors and asked to indicate which dimension each behavior reflects. See Appendix H for the full list of example behaviors. The trainer then presented and discussed examples of behaviors that represent different levels of performance (i.e., good performance versus poor performance) on each dimension. To illustrate, behaviors representing a 2.0 on a particular dimension were differentiated from behaviors that represent a 4.0 on the same dimension. To further practice matching behaviors and dimensions, participants were given a list of sample behaviors (similar to those seen in the videotapes) and asked to indicate which dimension each behavior reflects (adapted from Tennessee Assessment Center, 2002). (See Appendix I). The trainer then discussed these behaviors and provided feedback as to the dimension and level of performance (weak or effective) represented by each behavior. Participants then observed and rated a practice videotape (also a role play exercise using another assessment center candidate) similar to the ones used as the rating stimuli. To ensure that participants had exposure to examples of both weak and effective performance, the practice videotape consisted of a mixed performance episode where the candidate displayed both positive and negative behaviors across the five dimensions. Next, the trainer collected the ratings, wrote them on the board, and discussed the ratings with the group. Finally, the trainer provided feedback to the participants, explaining why the candidate should have received a particular rating on each dimension according to the ratings of the SMEs. The entire training session lasted about 45 minutes.

Control training. See Appendix J for the script that was used by the experimenter for the control training condition. Participants in the control training were instructed that they would be evaluating job 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 performance appraisal (adapted from Business & Legal Reports, Inc.) was shown. This particular training video was amenable to the control training condition because it used non-technical language and was intended for a broad audience. See Appendix K for the written consent from Business & Legal Reports, Inc., to use their performance appraisal lecture slides for the purposes of this study. The control training session also lasted approximately 45 minutes.

Dependent Variables

Rating accuracy. Using the formulas provided by Sulsky and Balzer (1988), rating accuracy was assessed via Cronbach's (1955) four indexes of rating accuracy: (a) elevation (E), (b) differential elevation (DE), (c) differential accuracy (DA), and (d) stereotype accuracy (SA). Each index reflects a different portion of the distance between participants' ratings and the target scores derived from the SMEs. Developed using an analysis of variance (ANOVA) framework, elevation represents the differential grand mean, differential elevation represents the differential main effect of ratees, stereotype accuracy refers to the differential main effect of dimensions, and differential accuracy refers to the differential Ratee x Dimension interaction (Sulsky & Balzer, 1988). Lower scores on these measures represent higher accuracy, whereas higher scores indicate lower levels of accuracy.

Borman's (1977) differential accuracy (BDA) was also assessed. Borman's differential accuracy is the average of the z-transformed correlation between a rater's ratings for each dimension and the corresponding true scores across ratees. Higher scores

on the index reflect better rating accuracy. It has been argued that Borman's differential accuracy is an index of rating validity as it provides correlational information and is thus insensitive to distances between ratings and true scores (Sulsky & Day, 1994).

Rather than utilizing a single overall accuracy index, multiple rating accuracy indexes were assessed because an overall accuracy index collapses across potentially important information that may be meaningful for understanding the effects of FOR training (Sulsky & Balzer, 1988). Moreover, some individual accuracy components may be more important in certain rating situations that others (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982). However, FOR training should lead to improved accuracy with respect to all of the Cronbach component indexes (Sulsky & Day, 1994), and previous researchers have found increases in all of Cronbach's indexes as a result of FOR training (e.g., Pulakos, 1986). In terms of the present study, multiple accuracy indexes will be needed to determine the relative influence of each component with respect to performance schema accuracy.

Performance schema accuracy. Performance schema accuracy (PSA) refers to the degree to which individuals have schemas of performance that are similar to an expert schema of performance. Each participant's performance schema was measured using a paired comparison computer program in which participants rated the degree of similarity of randomly paired job behavior statements. To select the behaviors to be included in the measure, four trained assessors were asked to rank order the behaviors within each dimension that were most relevant to the role play exercise used in the present study. The three behaviors from each dimension with the highest average rankings were retained for inclusion in the measure, for a total of 15 behavior statements. This resulted in a

measure consisting of 105 randomly paired comparisons (see Appendix L for the instructions for the measure (adapted from Organizational Research Group, 1998) and Appendix M for a list of the behavior statements used in the measure (adapted from Tennessee Assessment Center, 2002)).

To generate the referent, or expert, schema, the three SME similarity data matrices were first analyzed using multidimensional scaling. The number of dimensions was determined by constraining the number of dimensions to be between 2 and 5 because (a) there were not enough behaviors per dimension to warrant asking for more than 5 dimensions (Kruskal & Wish, 1978), and (b) the number of underlying dimensions should not exceed the number of theoretical dimensions. The 5-dimensional solution provided the best fit with a substantial R^2 of .99. Then, consistent with previous research using expert similarity data matrices (e.g., Day et al., 2001), the similarity ratings of the three SMEs were averaged to create the expert data matrix.

PSA was assessed using multidimensional scaling (MDS). MDS is a geometric modeling technique that has been found to be useful for representing the organization of knowledge (e.g., Forgas, 1981; Rentsch et al., 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 goodness-of-fit measure, and values of R^2 range from 0 to 1 with higher values reflecting better fit. To measure PSA, individual differences Euclidian distance (INDSCAL) MDS analyses were conducted on the SME similarity data matrix and each participant's similarity data matrix. The resulting R^2 value for each participant was operationalized as PSA in subsequent analyses.

Declarative knowledge. A behavioral classification measure was used to assess participants' declarative knowledge. This measure required participants to match 15 managerial behaviors to their respective dimensions. (See Appendix H). The number of correctly classified behaviors was operationalized as declarative knowledge in subsequent analyses.

III. Results

Pilot Study

To address potential concerns regarding participant fatigue due to the time length of the sessions in the present study, a pilot study was conducted in which 15 undergraduate students participated in a FOR training session, 7 of whom rated only 2 performance episodes and 8 of whom rated all 4 performance episodes. After rating the videotapes, each participant responded to a set of items designed to measure his/her level of fatigue. See Table 3 on page 25 for the list of items used and a summary of the results. Overall, the results of this pilot study revealed no significant increase in the fatigue levels of participants who rated all 4 episodes. Moreover, the two episodes shown in the 2-episode condition were shown last in the 4-episode condition, allowing for a test of fatigue-driven rating differences between the two conditions. A comparison of the elevation component of rating accuracy revealed no significant differences in elevation between the 2-episode condition (M = .62, SD = .17) and the 4-episode condition (M = .59, SD = .17), t (13) = .34, *ns*. Hence, the primary study was conducted as proposed using the original 4 episodes.

Primary Study

Intercorrelations and descriptive statistics for the study variables are reported in Table 4 on page 26. These same intercorrelations are reported separately for FOR- and control-trained participants in Table 5 on page 27. Means and standard deviations are reported separately for each condition in Table 6 on page 28.

Table 3

Means and Standard Deviations for Pilot Study Items by Training Condition

Item	2 videos (n = 7)	4 videos (n = 8)
1. I felt tired after rating the 2(4) videotapes.	6.57_a (.79)	5.00_b (1.69)
2. I don't think the quality of my ratings was affected by fatigue.	2.86_a (1.68)	4.88_b (1.73)
3. 2 (4) videotapes was enough practice for me.	6.14 (1.57)	5.25 (1.39)
4. By the end of the 2nd (4th) videotape, I was too tired to concentrate.	3.71 (2.06)	2.75 (1.16)
5. I would have been willing to rate more than 2 (4) videotapes in this study.	1.71 (1.25)	3.00 (2.33)
6. The amount of time I spent rating the videotapes was reasonable.	3.43 (1.81)	5.00 (1.07)

Note. Participants responded to each item using a 7-item Likert-type rating scale (1 = *disagree very much* to 7 = *agree very much*). Values in parentheses are standard deviations. Means with different subscripts are significantly different at $p < .05$.

Table 4

Means, Standard Deviations, and Intercorrelations for Study Variables

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. Gender[a]	1.56	.50	-										
2. Age	21.44	3.73	.10	-									
3. GPA	3.16	.41	.05	-.08	-								
4. Rating experience	.76	1.72	.11	.06	-.04	-							
5. Knowledge score	10.92	2.45	.00	.04	.06	.14	-						
6. E[b]	.73	.22	.03	-.07	-.08	.00	-.39**	-					
7. DE	.40	.20	.05	.03	-.06	.08	-.23**	.64**	-				
8. SA	.28	.11	-.09	-.06	-.13	-.10	-.30**	.62**	.25**	-			
9. DA	.37	.14	.07	-.15	-.16	-.06	-.36**	.66**	.09	.44**	-		
10. BDA	.76	1.72	-.10	-.02	.11	.02	.40**	-.64**	-.44**	-.44**	-.46**	-	
11. PSA	.89	.07	-.09	-.02	-.02	.03	.24**	-.26**	-.17*	-.16*	-.18*	.31**	-

Note. $N = 144$. GPA = grade point average. Rating experience = total number of times having rated the job performance of another person. E = elevation. DE = differential elevation. SA = stereotype accuracy. DA = differential accuracy. BDA = Borman's differential accuracy. PSA = performance schema accuracy.

[a] 1 = female, 2 = male.

[b] Correlations with E, DE, SA, and DA are negative because smaller values on these indexes represent greater accuracy.

*$p < .05$. **$p < .01$.

Table 5

Intercorrelations for Study Variables by Training Condition

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Gender[a]	-	.01	.16	.26*	.29*	-.12	-.07	-.25*	.01	. 03	-.03
2. Age	.18	-	.07	.12	.12	-.17	.06	-.24*	-.27**	.17	-.06
3. GPA	-.09	-.17	-	.01	.22	-.24	-.07	-.29*	-.32**	.21	-.12
4. Rating experience	-.05	.01	-.09	-	.18	.02	.05	-.02	-.09	.10	.08
5. Knowledge score	-.21	-.05	-.02	.11	-	-.18	-.08	-.22	-.27*	.30*	.01
6. E[b]	.02	.03	-.05	-.04	-.27*	-	.54**	.62**	.63**	-.57**	-.02
7. DE	.07	.04	-.15	.17	-.08	.59**	-	.22	-.09	-.26*	-.11
8. SA	-.02	.13	-.01	-.28*	-.11	.36**	-.06	-	.44	-.47*	.03
9. DA	-.01	-.07	-.05	-.01	-.12	.35**	-.04	.10	-	-.47*	.06
10. BDA	-.07	.22	.17	-.09	.12	-.29**	-.37**	-.05	.07	-	.08
11. PSA	-.05	-.01	.08	.01	.21*	-.21*	.01	-.14	-.19	.25*	-

Note. Frame-of-reference participants (n = 73) are below and 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. E = elevation. DE = differential elevation. SA = stereotype accuracy. DA = differential accuracy. BDA = Borman's differential accuracy. PSA = performance schema accuracy.

[a] 1 = female, 2 = male.

[b] Correlations with E, DE, SA, and DA are negative because smaller values on these indexes represent greater accuracy.

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

Table 6

Means and Standard Deviations of Study Variables by Training Condition

Variable	FOR (n = 73)		Control (n = 71)	
	M	*SD*	*M*	*SD*
1. Gender[a]	1.48	.50	1.63	.49
2. Age	21.52	4.33	21.37	3.02
3. GPA	3.12	.41	3.20	.41
4. Rating experience	.77	1.56	.76	1.88
5. Knowledge score	11.93	1.89	9.89	2.53
6. Elevation	.61	.12	.86	.23
7. Differential Elevation	.32	.14	.48	.22
8. Stereotype Accuracy	.24	.08	.32	.13
9. Differential Accuracy	.31	.08	.42	.16
10. Borman's Differential Accuracy	1.07	.44	.44	.55
11. Performance Schema Accuracy	.91	.06	.87	.06

Note. FOR = frame of reference. GPA = grade point average. Rating experience = total number of times having rated the job performance of another person. For elevation, differential elevation, stereotype accuracy, and differential accuracy, small numbers represent greater accuracy. For Borman's differential accuracy, larger numbers represent greater accuracy.
[a] 1 = female, 2 = male.

Analysis of Demographic Variables

Prior to conducting any analyses concerning the study hypotheses, appropriate two-sample tests were conducted on all relevant demographic variables for the FOR-trained group and the control-trained group. Results of these analyses revealed no significant differences in the two training groups for age, $t(142) = .25$, gender, $\chi^2(1) = 3.47$, race, $\chi^2(3) = .78$, GPA, $t(142) = 1.12$, or rating experience, $t(142) = .02$.

Performance Schema Accuracy

Hypothesis 1a predicted that PSA would be significantly greater after FOR training than before FOR training. For each individual PSA analysis, the number of dimensions was constrained to be 5 because this was the number of dimensions derived in the expert solution. Hypothesis 2 was tested by conducting a paired-samples *t*-test on the means of the Fisher-*z* transformed square roots of the R^2 values for FOR-trained participants pre- and post-training. Results revealed that the mean R^2 for the FOR-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 $d = .90$. In contrast, 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*. Hence, Hypothesis 1a was fully supported.

Hypothesis 1b predicted that PSA would be significantly greater for participants in the FOR training condition than for participants in the control training condition. To test this hypothesis, an independent-samples *t*-test was conducted on the means of the Fisher-*z* transformed square roots of the R^2 values for participants in the FOR and control training conditions. Analysis of these data revealed that the mean R^2 for the FOR-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, therefore, fully supported.

Rating Accuracy

Hypothesis 2 predicted that FOR-trained participants would provide more accurate ratings than control-trained participants. As Schleicher et al., (2002) pointed out, because of the conceptual overlap of the five accuracy indexes and their statistically significant intercorrelations (see Table 4), a multivariate framework is more appropriate for testing this hypothesis. Thus, Hypothesis 2 was tested via multivariate analysis of variance, with training (FOR vs. control) as the independent variable and the five rating accuracy indexes as the multiple dependent variables. Hypothesis 4 was fully supported, as results revealed that ratings provided by FOR-trained participants were significantly more accurate than those made by control-trained participants, $F(5, 138) = 16.66$, $p < .001$; Wilks's $\Lambda = .62$; partial $\eta^2 = .38$. A summary of the accuracy means for each training group is provided in Table 6 on page 28.

A follow-up 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 both elevation and BDA were driving the discrimination between the different training conditions (.50 and -.51, respectively).

Follow-up univariate analyses of variance (ANOVAs) were calculated to estimate the effect size associated with each accuracy dependent variable. A summary of these results is provided in Table 7 on page 31. Overall, the results of this analysis are

Table 7

Analysis of Variance Results for the Five Accuracy Components

Accuracy	FOR	Control	*F*	*p*	R^2
Elevation	.61	.86	66.23	< .001	.32
Differential elevation	.32	.48	24.83	< .001	.15
Stereotype accuracy	.24	.32	17.76	< .001	.11
Differential accuracy	.31	.42	28.89	< .001	.17
Borman's differential accuracy	1.07	.44	57.83	< .001	.29

Note. $N = 144$. For elevation, differential elevation, stereotype accuracy, and differential accuracy, small numbers represent greater accuracy. For Borman's differential accuracy, larger numbers represent greater accuracy. FOR = frame of reference.

consistent with previous FOR research using assessment center simulations as stimuli (e.g., Schleicher et al., 2002), and they support the ubiquitous research finding that FOR training is an effective approach for improving rating accuracy.

Performance Dimensions

Further corroborating the efficacy of FOR training was the additional finding that FOR-trained participants ($M = 4.44$, $SD = .42$) used a significantly larger number of performance dimensions to code candidate behaviors on their rating sheets than did control-trained participants ($M = 3.89$, $SD = .69$), $t(142) = 5.83$, $p < .001$ (one-tailed); Cohen's $d = .98$. This result was obtained by averaging the number of coded performance dimensions across all four candidates for both training groups.

Performance Schema Accuracy - Rating Accuracy Relationships

Hypothesis 3 predicted that PSA would be positively related to the five rating

accuracy indexes. As evidenced in Table 3, this hypothesis was fully supported as PSA correlated positively and significantly with each of the five rating accuracy indexes.[1] A closer inspection of Table 4 reveals that these correlations were generally larger in the FOR condition as compared to the control condition. However, none of these differences were statistically significant.

Declarative Knowledge

Hypothesis 4 predicted that FOR-trained participants would score higher on a measure of declarative knowledge than control-trained participants. This hypothesis was tested by conducting an independent-samples *t*-test on the mean knowledge scores for the two training conditions. Results indicated that FOR-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, thus, fully supported.

Incremental Validity of Performance Schema Accuracy

Finally, Hypothesis 5 predicted that PSA would account for a unique amount of variance in all five measures of rating accuracy over and above that of a measure of declarative knowledge. This hypothesis was tested by conducting 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. As evidenced in Table 8 on page 33, PSA accounted for a significant

[1] Correlations with elevation, differential elevation, stereotype accuracy, and differential accuracy are negative because smaller values on these indexes represent greater accuracy.

Table 8

Regression Results for the Incremental Validity of PSA

Accuracy Index	R	R^2	ΔR^2
Elevation			
Step 1			
Declarative Knowledge	.39	.15	
Step 2			
PSA	.42	.18	.03*
Differential Elevation			
Step 1			
Declarative Knowledge	.23	.05	
Step 2			
PSA	.26	.07	.01
Stereotype Accuracy			
Step 1			
Declarative Knowledge	.30	.09	
Step 2			
PSA	.31	.10	.01
Differential Accuracy			
Step 1			
Declarative Knowledge	.36	.13	
Step 2			
PSA	.37	.14	.01
Borman's Differential Accuracy			
Step 1			
Declarative Knowledge	.40	.16	
Step 2			
PSA	.46	.21	.05**

Note. $N = 144$. PSA = performance schema accuracy.
*p < .05. **p < .01.

amount of unique variance in elevation and BDA over and above that of declarative knowledge. Thus, partial support was found for Hypothesis 5.

IV. Discussion

Summary of Present Study Results

The purpose of the present study was to add to the literature on the cognitive effects of FOR training by examining the influence of FOR training on raters' schemas of performance. Results of the present study indicated that PSA was greater for FOR-trained raters than control-trained raters after training, and PSA improved significantly from pre-FOR training to post-FOR training compared to no pre-post improvement in control-trained raters. Moreover, FOR-trained raters provided more accurate ratings than control-trained raters, and PSA was positively associated with multiple indexes of rating accuracy. Finally, FOR-trained raters scored higher on a measure of declarative knowledge than did control-trained raters, and PSA added incremental variance to the prediction of two indexes of rating accuracy over and above that of a declarative knowledge measure.

Contributions of the Present Study

The results of the present study offer three important contributions to the FOR training literature. First, the present study complements previous work that has examined the extent to which FOR training influences raters' schemas of performance knowledge. Previous researchers (e.g., Woehr, 1994) have studied rater schemas by analyzing the organization of recalled ratee behaviors. In the present study, a standardized paired comparison technique was utilized because it allowed for an evaluation of raters' performance schema accuracy relative to an expert model.

Second, the present study is the first to examine the accuracy of FOR-trained rater schemas. Previous studies of the cognitive effects of FOR training are limited in that

they failed to assess the degree to which the cognitive variables that were measured compared to those of experts. This is surprising given that expert ratings have long been used in rater training research as a means of establishing rating accuracy. The present study addressed the issue of expert rater cognition, and perhaps this will prompt rater training researchers to consider experts not only as a source for developing "true" scores, but also as potential resources for evaluating the cognitive effects of training.

Third, the present study is also the first to provide a direct test of the cognitive changes that are hypothesized to occur as a result of FOR training. Previous studies have examined only the post-training cognitive effects of FOR training, inferring the existence of a change based on training-control differences. The results of the present study provided direct evidence that rater schemas of performance become more accurate as a result of FOR training, whereas control-trained raters showed no increase in schema accuracy. A possible avenue for future research would be to examine changes in rater schemas over time. For example, Sulsky and Day (1994) found that FOR-trained raters provided significantly more accurate ratings than control-trained raters even after a 48-hour delay, and Roch and O'Sullivan (2003) found no significant decay in rating accuracy after a two-week delay between FOR training and the rating task. Based on the results of the present study, a likely explanation for these findings might be that FOR training fosters the development of relatively stable schemas of performance, which in turn should account for the stability of rating accuracy over time. Further studies in this domain should consider the temporal stability of schema accuracy in addition to rating accuracy.

Results of the present study also further corroborate the vast number of studies documenting the efficacy of FOR training for improving rating accuracy. Specifically, the FOR training effect was strongest for elevation and BDA. There is some debate in the literature as to which components of rating accuracy are most relevant to FOR training. Schleicher and Day (2001), for example, argued that differential accuracy should be the primary component of interest because it directly assesses the degree to which performance is accurately rated per ratee on each dimension. Other studies, however, have found mixed results as to which components were influenced the most by FOR training. Based on results from the previously reviewed FOR training literature, the conflicting results appear to be due, at least in part, to choice of analysis (univariate vs. multivariate), and if multivariate, whether BDA was included in the analysis. In their discussion of these mixed findings, Sulsky and Day (1994) concluded that the specific components that are influenced most are likely to vary across studies. Perhaps future research could help shed some light on this issue by determining under what training conditions each component is most likely to be affected.

One interesting finding that emerged from the present study was the pattern of relatively large correlations between Cronbach's rating accuracy components. This finding is in contrast to previous research that has indicated these components are empirically independent of one another (e.g., Roach & Gupta, 1992). One explanation for this finding may be found in the design of the present study. To be specific, the rating stimuli used in the present study were chosen to control for possible differences due to performance level. Thus, ratings across each dimension for the above average performance level candidates tended to be very similar, as did the ratings across each

dimension for the below average performance level candidates (see Table 2). This may have led to the large correlations between the elevation component and the other accuracy components because raters were distinguishing between overall levels of ratee performance but were not making fine-grained distinctions between individual ratees and dimensions. It is not surprising, then, that elevation and Borman's differential accuracy showed the largest rating accuracy differences across the two training conditions. It should be noted, however, that significant differences were found for all rating accuracy indexes between the FOR and control training conditions.

A second interesting finding from the present study was that rater age and gender was significantly correlated with some of the components of rating accuracy in the control training condition, but not in the FOR training condition. It should be noted that a similar pattern of correlations was observed in the Schleicher et al. (2002) study, although the authors did not attempt to interpret these results. One explanation may be that FOR training's emphasis on creating a standard with which to judge performance has the added benefit of reducing certain rating biases. In other words, left to their own devices, raters may be more likely to use their own standards for evaluating performance, which then allows for the possibility that extraneous variables will systematically influence their ratings. Although beyond the scope of the present study, this is an interesting research question that would be better answered by incorporating these demographic variables into the design of an experimental procedure.

Limitations and Strengths of the Present Study

As with any study, one must be cautious in generalizing the results of the present study. The present study utilized student raters who were previously unfamiliar with the

rating format and the rating situation. In addition, raters in actual organizational or assessment center (AC) rating situations would likely be expected to be more invested in the outcome of the training and, perhaps, the training itself. Moreover, the AC candidates that appeared in the stimulus episodes for the present study were relatively homogeneous with respect to some demographic characteristics (e.g., age and race), which may not be representative of the population of ratees who are assessed in some organizations and assessment centers. The inclusion of demographically diverse ratees as rating stimuli is a potentially valuable topic for further research.

Nonetheless, there are many methodological aspects of the present study that can be viewed as improvements upon previous FOR studies. The rating conditions associated with the present laboratory study are more associated with those of assessment centers than other studies. For example, FOR research has historically utilized standardized tapes of teaching performance in which confederate graduate students deliver a lecture that can be classified as generally favorable, unfavorable, or mixed with regards to teaching performance. The present study offers an alternative to this approach by utilizing tapes of actual managers engaged in an actual AC exercise. In contrast, Schleicher et al. (2002) used tapes of business students participating in an undergraduate AC. Moreover, Schleicher and associates used only the first 5 minutes of their tapes "for control purposes and to keep the rating task manageable" (p. 738). Given that the first 5 minutes of a meeting is likely to revolve around superficial conversation, this can lead to a tremendous reduction in observable behaviors that may be important for making dimensional ratings. Finally, the present research answers, in part, Lievens (2001) call

for FOR training studies to employ additional AC exercises beyond presentation exercises.

Practical Implications

Two clear practical implications for FOR training emerged from the results of the present study. First, the finding that FOR-trained raters have greater levels of PSA than control-trained raters and that PSA predicted two indexes of rating accuracy (elevation and Borman's differential accuracy) over and above declarative knowledge suggests that PSA may be considered a meaningful outcome variable of FOR training. This implication is consistent with previous studies that have found that schema measures convey unique information about training that is not available in traditional measures of learning (e.g., Davis et al., 2003; Dorsey et al., 1999; Stout et al., 1997). Results of the present study indicate that PSA conveys meaningful information about the impact of FOR training and the development of rating accuracy. FOR training researchers might consider incorporating performance schema measures as training criteria in addition to traditional indexes of rating accuracy. Such information may be useful for determining which aspects of FOR training contribute most to the development of PSA. Incorporating performance schema measures may also lead to further refinements in the measurement of performance schemas, such as determining the ideal number of dimensions to include and which dimensions result in greater levels of PSA.

Second, PSA may be a potential tool for identifying idiosyncratic raters. Bernardin and Buckley (1981) originally proposed FOR training as a method for identifying raters with idiosyncratic frames of reference, a suggestion that has largely been ignored by FOR training researchers (Hauenstein & Foti, 1989). One reason for this

apparent oversight may be the lack of a standardized method for identifying idiosyncratic raters. PSA may provide useful information as to which raters have idiosyncratic schemas (frames of reference). This information may be useful for determining which raters may require further training. Moreover, Hauenstein and Foti (1989) recognized that training raters who already possess an appropriate frame of reference may be a waste of training resources. PSA could be assessed pre-training to identify those who already possess an appropriate frame of reference and thus may not benefit from the training. Additional research in this area could be directed toward the development of a model of schema idiosyncrasy, including making an empirical connection between rating idiosyncrasy and schema idiosyncrasy.

Conclusion

Previous research has found consistently positive effects of FOR training for improving rating accuracy. Many researchers have recognized the need for a better understanding of the cognitive mechanisms involved in FOR training, and consequently, numerous FOR 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 failed to account for the positive effects of FOR training in AC and other rating situations in which memory and recall are not as important. In all fairness, most of the research on FOR training has been conducted with the intention of generalizing the results to performance evaluations in organizations, in which memory and recall can become very salient factors with respect to rating accuracy. Only recently has FOR training been applied to AC rating situations, but this shift has signaled the need for more sophisticated cognitive measurement techniques that extend beyond memory

and recall. 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.

List of References

List of References

American Psychological Association (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060-1073.

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology*, *49*, 141-168.

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.

Bannister, B. D. (1986). Performance outcome feedback and attributional feedback: Interactive effects on recipient responses. *Journal of Applied Psychology*, *71*, 203-210.

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

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.

Bittner, R. H. (1948). Developing an industrial merit rating procedure. *Personnel Psychology*, *1*, 403-432.

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.

Business & Legal Reports, Inc. (n.d.). *Performance appraisals: How to conduct effectively.* Retrieved October 3, 2006, from http://hr.blr.com

Cannon-Bowers, J. A., & Salas, E., & Converse, S. A. (1993). Shared mental models in expert team decision making. In N. J. Castellan (Ed.), *Individual and group decision making* (pp. 221-246). Hillsdale, NJ: Lawrence Erlbaum.

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* (p. 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.

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*, 001-009.

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 posttraining 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.

Fulk, J., Brief, A. O., & Barr, S. H. (1985). Trust-in-supervisor and perceived fairness and accuracy of performance evaluations. *Journal of Business Research*, *13*, 301-313.

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.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology*, *73*, 68-73.

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.

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.

Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management*, *20*, 403-437.

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 Cannon-Bowers, J. A., & Salas, E. (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 115-153). Washington, D.C.: 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: Sage.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72-107.

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.

Loftus, E. F. (2003). Make believe memories. *American Psychologist, 58*, 864-873.

Loftus, E. F. (2004). Memories of things unseen. *Current Directions in Psychological Science, 13(4)*, 145-147.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, *85*, 273-283.

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.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationships between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology, 67*, 320-325.

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.

Organizational Research Group (1998). *Organizational Measurement Series: Vol. 1. Team Assessment and Evaluation Measures*. Knoxville, TN.

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., & 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 and 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.

Roach, D. W., & Gupta, N. (1992). A realistic simulation for assessing the relationships among components of rating accuracy. *Journal of Applied Psychology, 77*, 196-200.

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.

Rouse, W. B., Cannon-Bowers, J. A., & Salas, E. (1992). The role of mental models in team performance in complex systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*, 1296-1308.

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*, *38*, 78-91.

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.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review*, *11*, 22-40.

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.

Spool, M. D. (1978). Training programs for observers of behavior: A review. *Personnel Psychology*, *31*, 853-888.

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.

Tennessee Assessment Center (2002). *Assessment Center Rating Materials*. Knoxville, TN.

Werner, J. M., & Bolino, M. C. (1997). Explaining U. S. Courts of Appeals decisions involving performance appraisal: Accuracy, fairness, and validation. *Personnel Psychology*, *50*, 1-24.

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., & Feldman, J. (1993). Processing objective and question order effects on the causal relation between memory and judgment in performance appraisal: The tip of the iceberg. *Journal of Applied Psychology*, *78*, 232-241.

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.

Appendices

Appendix A

Demographic Questionnaire[1]

Participant I.D.: _____

The following information will be used ONLY for statistical purposes. All responses will be kept strictly confidential.

Demographic Information:

Age: _____

Major: _____

Gender
(Circle one): F M

Grade Point Average (GPA):_____

Race
(Circle one): African American
Asian/Pacific Islander
Native American
Caucasian
Other: _____

Class Rank
(Circle one): Freshman
Sophomore
Junior
Senior

Work Experience:

Do you Currently hold a job? Y N

If yes,

1. How long have you been at your current job? _____ Months
2. How many hours per week do you work? _____ Hours per week
3. Is your current job to be a career-oriented position or a job of convenience? (circle one)

Rating Experience:

1. How many total times have you rated the job performance of another person? __Times

2. How many different people have you rated? _____People

[1] Demographic questionnaire extracted and adapted from Organizational Research Group (1998)

Appendix B

Character Sketch for Role Player[1]

You will be rating scenarios depicting a meeting between two employees of a medical supply company. You will be rating the performance of the regional manager, who is meeting with a role player playing the part of a district manager named Christine Hawkins.

Here are some things you should know about the meeting. Christine is one of the company's best district managers. In fact, she received an award from the company for her impressive sales numbers. Christine is known to be moody, tyrannical, and obsessive, but she is also a great counselor and trainer. Her employees either love her or hate her.

Christine called this meeting because she wants to fire John Taylor, a poorly performing employee. He has had miserable sales for the last 18 months, and although she worked with him, he has not improved at all. Since only the regional manager has the authority to terminate him, Christine will try to convince the regional manager that this is a necessary step.

You should also know that Christine's customer satisfaction numbers have dropped significantly in recent months, although they are now improving. Christine will try to explain this away by suggesting that she pushed her sales people to focus on new sales, thus somewhat ignoring new customers.

[1] Character sketch extracted and adapted from Tennessee Assessment Center (2002)

Appendix C

Performance Dimensions, Definitions, and Behavioral Examples[1]

Analysis – The ability to identify problems, secure relevant information by effectively asking questions, relate data from different sources, and identify the possible causes of problems.

Behavioral Examples:

- Correctly identifies basic issues, including: data, facts, names/titles of people
- Correctly identifies relationships among: data, people, and problems
- Integrates information across sources
- Recognizes priorities among issues, materials, and data
- Secures relevant information by asking probing questions
- Identifies possible solutions

Decisiveness – Readiness to make decisions, render judgments, take action, or commit oneself; firmly expressing one's opinions and ideas.

Behavioral Examples:

- Makes specific recommendations
- Uses emphatic speech
- Commits to a course of action
- Delineates clear action plans
- Strongly expresses beliefs
- Recognizes the need for immediate action

Leadership – Utilizing appropriate interpersonal styles and methods in guiding individuals (subordinates/peers/superiors) or group toward task accomplishment.

Behavioral Examples:

- States goals and purposes for a meeting
- Maintains control of a meeting
- Provides direction/redirects discussion
- Solicits input from employees
- Establishes multiple agendas
- Articulates smooth transitions between topics
- Clarifies roles
- Resists the manipulations of other employees

- Attempts to motivate others

Confrontation – The ability and willingness to disagree or express opposing viewpoints in a tactful style; the willingness to stand up for thoughts and beliefs even when challenged.

Behavioral Examples:

- Confronts others about ideas or proposals
- Defends own positions when challenged
- Corrects others
- Voices dissenting opinions
- Challenges the ideas of others
- Asserts and uncommon/unpopular position

Sensitivity – The extent to which an individual shows consideration for the feelings and needs of others, asks for the opinions of others, and gives encouragement.

Behavioral Examples:

- Attentive behaviors (eye contact, nodding, “um”)
- Establishes rapport (small talk)
- Uses humor
- Exchanges social pleasantries
- Acknowledges contributions of others
- Does not interrupt others

Appendix D

Manager and Role Player Attractiveness Scale

Participant ID:_____ Episode:_____

Please respond to the following statements about the manager and the role player in the video you just watched. Using the scale below, please indicate your level of agreement with each statement.

1 = Disagree very much
2 = Disagree moderately
3 = Disagree slightly
4 = Neither agree nor disagree
5 = Agree slightly
6 = Agree moderately
7 = Agree very much

Manager:

Compared to most working adults that I know, the manager…

____ 1. appeared to be an attractive person.
____ 2. seemed to be a likable person.
____ 3. was not a friendly person.
____ 4. seemed like a pleasant person.
____ 5. was not very appealing.
____ 6. had a professional appearance.

Role Player:

Compared to most working adults that I know, the role player…

____ 1. appeared to be an attractive person.
____ 2. seemed to be a likable person.
____ 3. was not a friendly person.
____ 4. seemed like a pleasant person.
____ 5. was not very appealing.
____ 6. had a professional appearance.

Appendix E

Simulation Rating Form

Participant ID:_____ Episode:_____

Ratings
Analysis
Decisiveness
Leadership
Confrontation
Sensitivity
Overall

Appendix F

Rating Scale for Performance Dimensions[1]

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

[1] Rating scale adapted from Tennessee Assessment Center (2002)

Appendix G

Script for FOR Training Condition

INTRODUCTION

[*As participants arrive at the lab, ask them to sign in and have them take a seat at any available computer terminal.*]

Welcome, my name is Allen Gorman and I will be running today's session. Before we begin, please turn off your cell phones and please note that no food or drink is permitted in the lab.

First of all, thank you very much for volunteering to participate in this project. Your honest efforts during this session are greatly appreciated.

To begin, please take a look at the consent form that is placed at your workstation. This form gives you some information about what you will be doing today. I will read a few sections to you, then you can read the rest of the sheet carefully to yourself. Feel free to ask any questions you may have.

[*The following is a copy of the consent form:*]

CONSENT TO PARTICIPATE
DEPARTMENT OF MANAGEMENT
THE UNIVERSITY OF TENNESSEE
KNOXVILLE, TENNESSEE 37996

Title	Evaluating Work Performance
Purpose	The researchers listed below are conducting a study on how people evaluate work performance. The primary task of this study will require you to review videotapes of managerial performance and document managerial behaviors.
Activities	As a part of this study, I will learn about evaluating job performance, I will view and rate videotapes of managerial performance, and I will respond to some surveys about my experiences during the study. During and after the study is complete, all data and related information will be kept in a locked laboratory indefinitely.
Compensation	I will receive extra course credit in exchange for my participation in this study.
Confidentiality	I understand that my identity will remain anonymous and that I will not be identified in any report or publication.

Risks	There are no known risks.
Freedom to withdraw	I realize that research participation is completely voluntary. I understand that I am free to refuse to participate in this study or withdraw at any time. There is no penalty of any kind for either non-participation or withdrawal.
Availability of results	A summary of these results will be available from the researcher 5/15/08. The summary will include only aggregated (i.e., combined) data for the entire sample. No individual results will be available.
Investigator availability	The research investigators are listed below and if you have concerns or questions about the research, they can be reached at the listed telephone numbers or at The University of Tennessee's Department of Management (974-3161).

C. Allen Gorman	974-1681	**Joan Rentsch**	974-1671
Principal Investigator		Co-investigator & Faculty Advisor	

Consent My signature below indicates that I consent to participate in this research investigation.

_____ _____

Signed Date

Name (Please Print Neatly)

[*Read over consent form with participants:*]

During this study you will participate in a training program designed to train managers how to effectively conduct performance appraisals in organizations. You will learn how to recognize and rate dimensions of managerial job performance, and then you will have the opportunity to practice rating videotapes of managerial performance. Any information obtained about you during the study will be kept strictly confidential and will be stored in a locked cabinet in a locked room at a University of Tennessee location. There are no known risks in this study. The benefits of participating in this study include an opportunity to learn about the evaluation of managerial job performance.

In exchange for your participation, you will receive extra course credit. Your participation in this study is voluntary. You may decline to participate without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw before

the data collection is complete, your data will be destroyed. However, be aware that you must complete the session in order to receive extra credit.

Any questions? Okay, go ahead and sign the form.

[*Collect the consent forms.*]

For this project, I want you to be aware that everything is straightforward, meaning I am not trying to trick you, so please ask any questions you might have.

Okay, so before we talk about evaluating job performance in organizations, I would like you to complete a survey about your perceptions of job performance. Please select the 'Start' button on your desktop and then select 'Run'. In the dialog box, please type in the following exactly as it looks on the white board: (E:\PairedComparison\compare.exe E:\PairedComparison\cmd1.txt). Please type in the ID number that is printed on your folder. Please follow along as I read the instructions.

[*Read the instructions for the Performance Schema Measure.*]

Does anyone have any questions? You may begin?

[*Allow time for participants to complete the Performance Schema Measures (approx. 15 minutes).*]

FOR TRAINING

As I said earlier, you are taking part in a training program that will help you learn how to recognize and rate dimensions of job performance. Many of you are likely to be a manager at some point in your career, and managers are often called upon to conduct performance appraisals for their organizations. Thus, the things that you learn in this training may help you when the time comes for you to evaluate other people in your organization.

Before we begin, let me tell you a few things about performance appraisals:

- The purpose of the appraisal process is to inform employees of how they are doing and how they can improve the quality of their performance.

- Properly conducted performance appraisals are motivational and help employees grow and develop.

- Preparing for and conducting performance appraisals are among the most important things you will do as a supervisor.

For an organization to be successful, every employee needs to be able to perform to the best of their abilities. They can only do so if they have adequate feedback and clearly defined goals.

Performance appraisals are an essential tool for accomplishing these tasks. They offer a formal and official way to:

> • Recognize accomplishments. Every company will define recognizable accomplishments differently, but it's important to have a reward system in place.
>
> • Guide employee progress. Effective performance appraisals continue to refine the initial job description of what is expected of employees as they learn new skills and gain experience.
>
> • Improve performance. Whether making good performance better or correcting poor performance, performance appraisals are an important step in identifying the situation and laying out the course for improvement.

Okay, now that you have had an introduction to the idea of performance appraisal, I am going to show you some example work behaviors and a typical rating scale.

[*Hand out dimension definitions.*]

Take a look at the handout entitled "Dimensions, Definitions, and Behavioral Examples." These 5 dimensions are commonly used to categorize the job performance of managers: Analysis, Decisiveness, Leadership, Confrontation, and Sensitivity.

Please read the definition of each dimension and the example behaviors to yourself while I read them aloud.

Analysis – The ability to identify problems, secure relevant information by effectively asking questions, relate data from different sources, and identify the possible causes of problems.

Behavioral Examples:

- Correctly identifies basic issues, including: data, facts, names/titles of people
- Correctly identifies relationships among: data, people, and problems
- Integrates information across sources
- Recognizes priorities among issues, materials, and data
- Secures relevant information by asking probing questions
- Identifies possible solutions

Decisiveness – Readiness to make decisions, render judgments, take action, or commit oneself; firmly expressing one's opinions and ideas.

Behavioral Examples:

- Makes specific recommendations
- Uses emphatic speech
- Commits to a course of action
- Delineates clear action plans
- Strongly expresses beliefs
- Recognizes the need for immediate action

Leadership – Utilizing appropriate interpersonal styles and methods in guiding individuals (subordinates/peers/superiors) or group toward task accomplishment

Behavioral Examples:

- States goals and purposes for a meeting
- Maintains control of a meeting
- Provides direction/redirects discussion
- Solicits input from employees
- Establishes multiple agendas
- Articulates smooth transitions between topics
- Clarifies roles
- Resists the manipulations of other employees
- Attempts to motivate others

Confrontation – The ability and willingness to disagree or express opposing viewpoints in a tactful style; the willingness to stand up for thoughts and beliefs even when challenged

Behavioral Examples:

- Confronts others about ideas or proposals
- Defends own positions when challenged
- Corrects others
- Voices dissenting opinions
- Challenges the ideas of others
- Asserts and uncommon/unpopular position

Sensitivity – The extent to which an individual shows consideration for the feelings and needs of others, asks for the opinions of others, and gives encouragement.

Behavioral Examples:

- Attentive behaviors (eye contact, nodding, "um")
- Establishes rapport (small talk)
- Uses humor
- Exchanges social pleasantries
- Acknowledges contributions of others
- Does not interrupt others

You will be using these dimensions and example behaviors to rate some videotapes of managerial performance. As you watch each videotape, you are going to record specific behaviors for each dimension. Behaviors refer to those things individuals actually do or say. Behaviors are gathered directly from our observation of others; they represent information that has yet to be processed.

For example, the following statements are examples of leadership behaviors: "He solicited information from the subordinate" or "He provided little overall direction to the meeting." This is not a leadership behavior: "He is a deficient leader." This statement does not tell you anything about what the person actually did.

Okay, now let's practice putting behaviors in their correct dimensions.

[*Read examples from classification practice Form A. Allow time to discuss answers with group.*]

Good, now let me tell you a little bit more about the tapes you will be rating.

[*Hand out character sketch.*]

Each of these scenarios depicts a meeting between two employees of a medical supply company. You will be rating the performance of the regional manager, who is meeting with a role player playing the part of a district manager named Christine Hawkins. Prior to the meeting, the regional manager was provided with information that may be useful in the meeting with Christine, including various sales figures and charts.

Here are some things you should know about the meeting. Christine is one of the company's best district managers. In fact, she received an award from the company for her impressive sales numbers. Christine is known to be moody, tyrannical, and obsessive, but she is also a great counselor and trainer. Her employees either love her or hate her.

Christine called this meeting because she wants to fire John Taylor, a poorly performing employee. He has had miserable sales for the last 18 months, and though she worked with him, he has not improved at all. Since only the regional manager has the authority to terminate him, Christine will try to convince the regional manager that this is a necessary step.

You should also know that Christine's customer satisfaction numbers have dropped significantly in recent months, though they are now improving. Christine will try to explain this away by suggesting that she pushed her sales people to focus on new sales, thus somewhat ignoring new customers.

Now let's talk about some example behaviors you might see in the meeting with the role player and how you might rate these behaviors using this rating scale.

[*Hand out and read over rating scale and anchors.*]

[*Interact with participants in the following section: e.g., ask for volunteers to name the dimensions represented by the behaviors.*]

Let's start with the Analysis dimension. If, for example, a manager in this scenario asked Christine many specific and probing questions, integrated information across different sources, and utilized the charts and graphs that are provided during the meeting, he or she might get a rating of 4.0 (highly effective) or higher in the Analysis dimension.

On the other hand, if a manager does not recognize the major issues surrounding the meeting with Christine, is unfamiliar with the provided materials, and does not ask Christine any specific and probing questions, he or she might get a rating of 2.0 (weak) or lower in the Analysis dimension.

If a manager identifies the basic issues in the meeting, shows some familiarity with the materials, and asks minimal probes, he or she might get a rating around a 3.0 (satisfactory).

A rating of 3.0 using this scale is considered average. If you give a rating above a 3.0 or below a 3.0 on any dimension, the behaviors that you recorded in the dimension should reflect the above or below average rating. In other words, if you give a below average rating, you should have some negative behavior(s) documented on the rating form that support your rating. On the other hand, if you give an above average rating, you should have some positive behavior(s) documented on the rating form that support your rating.

Does anyone have any questions at this point?

Moving on to Decisiveness, if a manager in this scenario makes solid decisions, uses a strong tone, and articulates a detailed plan of action, he or she would get an above average rating. However, if a manager in this scenario wavers or is hesitant to make decisions, refuses to make clear decisions, or offers no time frame for when decisions will be made, he or she would get a below average rating.

How far above or below average your rating will depend on the behaviors that you have documented. Don't be afraid to use the entire rating scale. Just be sure that you can support your rating using the documented behaviors. Remember, your ratings should be based on actual behaviors, not just your impressions of the manager.

For Leadership, if a manager in this scenario maintains control of the meeting with Christine, asks for input from Christine, or attempts to motivate Christine, he or she would get an above average rating. If he or she loses control of the meeting or has no impact on the outcome of the meeting, this would person would receive a below average rating.

For the Confrontation dimension, if a manager in this scenario willingly confronts Christine, is tactful, defends his or her perspective on the issues, and follows through when Christine disagrees, this person would receive an above average rating. On the other hand, if a manager overtly avoids conflict or does not confront Christine in a tactful manner, he or she would receive a below average rating.

Finally, for the Sensitivity dimension, a manager in this scenario would receive an above average rating if he or she was polite, attentive, and respectful during the meeting and if he or she attempted to build rapport with Christine and was sensitive to the fate of other employees. However, a manager would receive a below average rating if he or she was insulting or disrespectful, or interrupted Christine during the meeting.

Let's talk some more about recording behaviors on your rating form. One way to document behaviors is to write down quotes. So, as you watch the videotapes, write down things that the manager says on your rating form. Then, write down the behavior that the quote exemplifies using the behavioral examples on your definition sheet.

[*Write down following example on white board.*]

For example, if a manager were to ask Christine, "What is your market share?" this would be an example of securing relevant information by asking probing questions, which is an Analysis behavior. So, on your rating form, you would write down what the manager said, word for word, and beside it you would write something like "asked probing questions" to help you categorize the quote.

Now, when you watch the videotapes, you will see that there is a lot of information to write down. What I recommend is that you write down as many quotes as you can while you are watching the videotape. Then, when the videotape is over, use a different color pen to write down the behavior that the quote exemplifies. You should notice at your seat that I have provided you with red colored pens so you can do this easily.

It is also helpful to note whether the behavior you wrote down is positive, negative, or neutral. To do this, I recommend that you write a +, –, or 0 next to each behavior to indicate whether it is a positive, negative, or neutral behavior.

After you have watched each videotape and recorded the behaviors on your sheet you will then assign a rating to each dimension using the scale provided. After you have rated each dimension, you will record a global rating on your rating form using the same rating scale. The global rating is your overall impression of the manager's performance across all the dimensions.

Any questions so far? Okay, now let's practice putting example behaviors into their correct dimensions.

[*Hand out behavior classification practice Form B. Allow for time to complete, then discuss answers with group.*]

Okay, now let's take a 5 minute break.

[*Allow for 5 minute break before rating practice videotape. Tell participants the locations of bathrooms and water fountains.*]

Okay, now that you are familiar with the dimensions and example behaviors, let's watch a practice videotape. Use your rating form to record quotes and behaviors as you observe them. At the conclusion of the tape, remember to use your red pen to indicate the behavior that each quote reflects, and then assign your ratings in the spaces provided. Don't forget to put a global rating on your rating form after rating all the dimensions. When you are done, we will go over the observed behaviors and the dimension ratings together.

[*Play practice videotape (Episode F). Allow for time at end to write +,-,or 0 and assign ratings. Give feedback regarding behaviors and ratings based on SME target ratings.*]

[*Ask participants to give their ratings and write them on the board.*]

Four subject matter experts rated this videotape independently, and then they agreed upon the following ratings:

Analysis: The experts gave a 3.7 in Analysis, noting the following behaviors:

> Asked probing questions: "Do you have any information [on his] sales? What do you tell them the goals are? What's the history of this region? What would his response be to your comments?
>
> Integrated across materials: integrates customer service index (CSI) w/ John Taylor's performance for his customers; integrates CSI with turnover (TO): "Do you think turnover impacts your customer satisfaction?"

Decisiveness: The experts gave a 3.0 in Decisiveness, noting the following behaviors:

> Refused to sign the termination letter: "I'd like to meet with him first"
>
> BUT
>
> No action plan for his conversation with John Taylor (was unsure what he would discuss with Taylor)

Leadership: The experts gave a 2.7 in Leadership, noting the following behaviors:

> Directed meeting with probes
> Solicited input: "What are your goals"
> Resisted Christine's manipulations: "I'll talk to Mr. Lane"
>
> BUT
>
> No multiple agenda
> No impact except refusal to sign
> No attempt to stop Christine from going to Mr. Lane
> Allowed Christine to walk out of meeting

Confrontation: The experts gave a 3.7 in Confrontation, noting the following behaviors:

> Pointed out weaknesses: "I think your sales are fantastic, but [customer service is important too]"; "So you know you're in the middle and there is a trend downwards?"
> Disagreed: "I'm not sure I agree entirely"
> Defends his position: "Christine, if I was going behind your back, I wouldn't have told you"
> Used tact while disagreeing: "As you know, the decision to terminate is mine, and while I respect your opinion, I think to be fair, I need to meet with Mr. Taylor"

Sensitivity: The experts gave a 3.0 in Sensitivity, noting the following behaviors:

> Polite: attentive; "okay"; "I appreciate your opinion"
> Attempted to develop rapport: "You're more a veteran here than I am"
> Praise: "You won the Lane Award"; "Your numbers have been very good"
> Empathetic: "I understand your concerns"

Now that you have seen a practice videotape, we're going to practice one more time matching behaviors and dimensions.

[*Hand out Behavior Classification Form A.*]

Okay, now that you have learned about rating job performance and had a chance to observe and rate the job performance of another person, I would like you to complete another survey about your perceptions of job performance. You will notice that this survey is very similar, although not identical, to the first survey you took. Please select the 'Start' button on your desktop and then select 'Run'. In the dialog box, please type in the following exactly as it looks on my computer screen:(E:\PairedComparison\compare.exe E:\PairedComparison\cmd1.txt). Please type in the ID number that is printed on your folder. [*If ID# 10001, now 20001*]. Please follow along as I read the instructions.

[*Read the instructions for the Performance Schema Measure.*]

Does anyone have any questions? You may begin?

[*Allow time for participants to complete the Performance Schema Measures (approx. 15 minutes).*]

Okay, now let's take a 5 minute break.

[*Allow for 5 minute break before rating videotapes.*]

Okay, now you are going to rate some more videotapes like the one you just saw. You will be rating 4 different videotapes of 4 different managers interacting with Christine Hawkins. You will write down quotes, record behaviors, and assign ratings just like in the example we just went through. We will repeat this process for all 4 videotapes.

Are there any questions before we begin?

[*Videotapes: B (Male Low); C (Female High); D (Female Low); E (Male High)*]

[*Participants watch and rate first videotape. Ask participants to put training materials and rating forms in folders at their desks.*]

Now I want you to identify whether each behavior was positive, negative, or neutral by placing either a +, –, or 0 next to the behavior.

[*Allow time to indicate sign of behavior.*]

Now, I want you to identify the dimension that each behavior belongs to by writing the name of the dimension next to each behavior.

[*Allow time to record dimensions.*]

Okay, now assign your ratings for each dimension using the spaces provided. Don't forget to put a global rating on your rating form.

[*Allow time to make ratings.*]

Now I would like you to answer a few questions about the manager and the role player in the videotape.

[*Allow time to complete attractiveness measure.*]

[*Steps are repeated until all 4 tapes have been rated.*]

[*Hand out demographic form.*]

Okay, now I would like you to answer a few questions about yourself. Please read these items carefully and enter your answers on the sheet. These items will only be used for statistical purposes.

[*Collect demographic form.*]

[*Hand out debriefing form.*]

Now I will give you a sheet that gives you some additional information about the study. Please don't discuss the details of this study with anyone. We expect that this study may contribute to the development of future training programs, but this will only happen if participants enter the session uninformed. So please keep quiet about your experiences, other than to suggest to your friends that they can participate.

Thank you very much for your participation. You may be dismissed.

Appendix H

Behavior Classification Practice Form A[1]

When conducting performance appraisals, it is important to accurately classify the behaviors that you observe. Your task is to categorize the following behaviors into their respective performance dimensions.

Analysis
Decisiveness
Leadership
Confrontation
Sensitivity

	Dimension
1. Strongly expresses beliefs	
2. Clarifies roles	
3. Identifies possible solutions	
4. Corrects others	
5. Acknowledges contributions of others	
6. Attempts to motivate others	
7. Integrates information across sources	
8. Uses humor	
9. Makes specific recommendations	
10. Voices dissenting opinions	
11. Maintains control of a meeting	
12. Does not interrupt others	
13. Secures relevant information by asking probing questions	
14. Commits to a clear course of action	
15. Solicits input from employees	

[1] Behavior classification practice form extracted and adapted from Tennessee Assessment Center (2002)

Appendix I

Behavior Classification Practice Form B[1]

When conducting performance appraisals, it is important to accurately classify the behaviors that you observe. Your task is to categorize the following behavioral statements into their respective performance dimensions.

Analysis
Decisiveness
Leadership
Confrontation
Sensitivity

	Dimension
1. She shuffled through papers while the role player was speaking.	
2. "I am sorry that your previous order was incorrectly filled. I have verified this order myself, and I have approved a 10% discount for this order."	
3. She solicited the input of the role player on three different occasions.	
4. "I'm sorry, but I don't think your solution is feasible."	
5. He failed to recognize the importance of the lack of training the employees received.	
6. After the role player expressed his disagreement, the manager redirected the meeting, saying, "I appreciate your concern, but we need to move on the next issue."	
7. "It seems clear to me that all of your problems are due to lack of motivation."	

[1] Behavior classification practice form extracted and adapted from Tennessee Assessment Center (2002)

Appendix J

Script for Control Training Condition

INTRODUCTION

[*As participants arrive at the lab, ask them to sign in and have them take a seat at any available computer terminal.*]

Welcome, my name is Allen Gorman and I will be running today's session. Before we begin, please turn off your cell phones and please note that no food or drink is permitted in the lab.

First of all, thank you very much for volunteering to participate in this project. Your honest efforts during this session will be greatly appreciated.

To begin, please take a look at the consent form that is placed at your workstation. This form gives you some information about what you will be doing today. I will read a few sections to you, then you can read the rest of the sheet carefully to yourself. Feel free to ask any questions you may have.

During this study you will learn about the evaluation of job performance in organizations. I will give a brief lecture on performance appraisal in organizations, and then you will have the opportunity to practice evaluating videotapes of managerial performance. Any information obtained about you during the study will be kept strictly confidential and will be stored in a locked cabinet in a locked room at a University of Tennessee location. There are no known risks in this study. The benefits of participating in this study include an opportunity to practice rating someone else's job performance.

In exchange for your participation, you will receive extra course credit. Your participation in this study is voluntary. You may decline to participate without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw before the data collection is complete, your data will be destroyed. However, be aware that you must complete the session in order to receive extra credit.

Any questions? Okay, go ahead and sign the form.

[*Collect the consent forms.*]

For this project, I want you to be aware that everything is straightforward, meaning I am not trying to trick you, so please ask any questions you might have.

Okay, so before we talk about evaluating job performance in organizations, I would like you to complete a survey about your perceptions of job performance. Please select the 'Start' button on your desktop and then select 'Run'. In the dialog box, please type in the following exactly as it looks on the white board: (E:\PairedComparison\compare.exe E:\PairedComparison\cmd1.txt). Please type in the ID number that is printed on your folder. Please follow along as I read the instructions.

[*Read the instructions for the Performance Schema Measure.*]

Does anyone have any questions? You may begin?

[*Allow time for participants to complete the Performance Schema Measures (approx. 15 minutes).*]

[*Allow participants to take a 5 minute break before training starts.*]

CONTROL TRAINING

As I said earlier, you will have the opportunity to practice rating other peoples' job performance during this session. But before we do that, I need to give you some information about how people evaluate job performance in organizations. Many of you are likely to be a manager at some point in your career, and managers are often called upon to conduct performance appraisals for their organizations. Thus, the following lecture is intended to give you an introduction to some important things to consider when the time comes for you to evaluate other people in your organization. You may take notes if you wish using the scratch paper that is provided. However, you will not be tested on this material as a part of this study.

[*Begin PowerPoint presentation*]:

In this training session, we will discuss how to conduct effective performance appraisals.

- The purpose of the appraisal process is to inform employees of how they are doing and how they can improve the quality of their performance.

- Properly conducted performance appraisals are motivational and help employees grow and develop.

- Preparing for and conducting performance appraisals are among the most important things you will do as a supervisor.

We will discuss:

- The importance and benefits of performance appraisals
- How to avoid discrimination during the appraisal process
- How to measure and document performance
- How to set performance goals
- How to prepare for and conduct appraisal interviews
- How to deal with different levels of employee performance

Feel free to ask questions during the presentation if anything is unclear or needs further explanation.

For an organization to be successful, every employee needs to be able to perform to the best of their abilities. They can only do so if they have adequate feedback and clearly defined goals.

Performance appraisals are an essential tool for accomplishing these tasks. They offer a formal and official way to:

- Recognize accomplishments. Every company will define recognizable accomplishments differently, but it's important to have a reward system in place.

- Guide employee progress. Effective performance appraisals continue to refine the initial job description of what is expected of employees as they learn new skills and gain experience.

- Improve performance. Whether making good performance better or correcting poor performance, performance appraisals are an important step in identifying the situation and laying out the course for improvement.

Performance appraisals also provide the opportunity to:

- Review how well employees have met job requirements and goals.
- Set new performance goals, including additional responsibilities.
- Identify areas in which performance needs to be improved.
- Discuss career advancement, including training opportunities and promotion.

Performance appraisals offer many benefits to the company, including:

- Documentation of performance issues, disciplinary actions, written goals, and so on—all signed by the involved parties

- A system for providing employee development opportunities

- A regular outlet for providing performance feedback

- Legal protection should the company be involved in accusations of discrimination or illegal termination

- Morale boost to motivate employees through a recognized and defined reward system

Performance appraisals offer many of the same benefits to employees, including:

- Clear direction in their work regarding what's expected of them and of their role in the company's overall goals

- A regular outlet in which to receive feedback on performance and expectations

- A regular time in which to give input on their job, their department, or the company

- Motivation to perform their best because they know they will be recognized and/or rewarded

Typical legal problems associated with performance appraisals involve charges of discrimination.

- Title VII of the Civil Rights Act of 1964 prohibits employment discrimination, including discrimination in the evaluation of employee performance, because of race, national origin, religion, age, or sex.

- The Americans with Disabilities Act prohibits discrimination against disabled employees—for example, judging their performance more harshly because of their disability.

- Other fair employment laws, such as the Pregnancy Discrimination Act and Equal Pay Act, also prohibit discriminatory practices related to performance appraisals.

Legal problems and discrimination charges may arise from:

- Failure to clearly communicate performance standards
- Failure to give timely feedback when performance does not meet standards
- Failure to allow employees the opportunity to correct inadequate performance
- Inconsistency in measuring performance from employee to employee

- Failure to document performance objectively

We'll now learn about the specifics of the appraisal process, beginning with appraisal forms.

In order to be most effective, performance appraisal forms need to be well constructed and easy to understand. They should include the following items:

- Well-defined performance expectations in areas that include: adaptability, communication skills, cooperation, dependability, initiative, innovation, job knowledge, organization, productivity, and quality of work
- Clearly described measurement tools
- A concrete rating system
- Space to write down specific job examples
- A section for setting timely, measurable performance goals

There are many ways to measure performance, but the main thing to remember is that the more objective the measurement, the better.

- Use specific rating scales—whether numbers or terms—when assessing performance.
- Use a system that is fair and flexible in assessing workplace situations and performance.
- Be consistent in using the same measurement systems for all employees.
- Make sure the measurement system is clear about what is being measured. Also make sure it is understood by all employees.
- Measurements need to be a useful tool that enables you to give a meaningful assessment as well as enabling employees to know exactly how their performance measures up.

One of the most common rating scales is numerical because numbers are perceived to be the most objective. It's a good idea to also attach words describing what each number means, however, to make sure the numbers are used consistently. For example, on a 1 to 5 scale, 1 could mean "well below standard," 3 means "meets standards," and 5 means "well above standard."

• Measurement terms can also be used as long as they are specific, meaningful, and distinct from one another. A typical spread goes from "unsatisfactory" through "satisfactory" to "outstanding."

• Management by Objectives (MBO) is a system of ratings that measure how well an employee reached specific goals or objectives, such as producing x number of pieces per shift or making x number of calls per hour.

• Systems can also measure effort or results with behaviors you can observe and track, such as attendance or initiative.

Once you have measured employee performance, you need to document your findings in a useful way that will help you prepare for appraisal interviews and avoid discrimination charges.

• Make sure all performance documentation is objective, based on performance not personalities.

• Document performance of all employees, not just troublemakers or star performers.

• Be sure that your documentation provides complete and accurate information that will support your conclusions about employees' performance. Include both favorable and unfavorable comments to give a realistic picture of performance. No one is perfect. No one is without some redeeming qualities.

• Document performance on a regular basis not just before a scheduled performance appraisal—for example, at least once a month on each employee.

Since goal-setting is central to an effective performance appraisal, it's crucial to get it right. Performance goals should be set with employees and meet the following criteria.

• Goals must be based on actual job requirements. Use the job description when setting performance goals.

• Goals must be realistic and achievable—otherwise they will frustrate rather than motivate employees. They should account for changing conditions and priorities.
• They must also be measurable, which means that they are specific and practical.

• Goals need to be observable in any number of areas, including time spent or results produced.

• Goals must remain challenging. They need to evolve with time. Once previous standards are met, raise the bar.

• Goals must be clearly prioritized so that employees know which are most important to you and the organization.

For the performance appraisal process to be most effective, you need to get your employees involved from the beginning.

• Employees must be encouraged to take an active role in:
  - Setting their performance goals
  - Designing the action plans to help them achieve their goals
  - Identifying their professional strengths and weaknesses, and giving their input about how to improve these identified areas of their performance

• Employees also need to be very involved in the performance appraisal meeting from preparation through the final report. Employees are much more fully invested in their performance when they play a large part in designing and guiding it.

Preparation for performance appraisals involves both you and your employees. Employees need to prepare for their performance appraisal meeting by:

• Reviewing their performance during the evaluation period as objectively as possible, considering their achievements and how well they have met their goals.

• Thinking about new performance goals for the next evaluation period.

Your preparation includes:

• Objectively reviewing employee performance

• Completing a written appraisal using the company's form

• Thinking about new goals for employees

• Scheduling a time and place for the meeting and giving employees ample notice so that they have time to prepare

Getting the appraisal meeting off to a good start is essential.

• Start by laying out a flexible agenda that includes plenty of time for feedback and discussion. Also set a positive tone with a few comments reminding

employees that the purpose of the performance appraisal is to improve performance and not to find fault or cast blame.

• Talking about money is a sensitive area. Raises are often associated with performance appraisals, and employees may expect to receive a raise immediately after their performance appraisal, especially if they receive a good evaluation. It's very important, therefore, to make sure employees know company policy on raises—preferably before you begin the performance appraisal meeting.

• Encourage input right from the beginning by asking for employees' understanding of every point of discussion.

• Give the good news first about successes, goals exceeded, and goals met.

Conducting effective performance appraisals is difficult partly because of the sensitive nature of being "on review." Positive presentation of the issues during performance appraisal meetings is, therefore, very important.

• Make sure you always keep the conversation focused on professional behavior and performance. Don't get personal.

• Also, stick to objective examples, especially when pointing out an area that needs improvement.

• Continue to invite response from employees throughout the meeting; never let it get to be one-sided.

• Listen actively by looking at employees, nodding, and using affirmative phrases such as "okay" and "I see."

• Create a "we" mentality, which shows that supervisor and employee are working together to help the employee give the best performance and have the best career opportunities that are available.

During the performance appraisal meeting be sure to review performance:

• Specifically as it relates to the goals that were set at the last performance appraisal

• Making note of strengths and accomplishments during the period so that employees know that you noticed

- Noting where performance fell short, but doing so along with encouragement and a listing of any resources, such as training or coaching, that can help employees meet their goals

Set goals for the next period based on company goals and employee performance.

- For example, if the company sets new production standards, create employee goals that help employees help the company meet its goals.

- Or if an employee has not met previous goals, reassert the goals but in a modified form that you and the employee agree is a realistic challenge.

Endings are as important as beginnings when it comes to performance appraisal meetings. An effective meeting must end by clearly setting a path for the future in order to motivate employees to do their best.

- End on a positive note by letting employees know where they're doing a good job and encouraging them to take advantage of professional opportunities to improve their performance even more.

- Lay out a detailed action plan that includes measurable tasks and a timetable for accomplishing them.

- When performance has been inadequate, confirm that employees know what will happen if they don't improve. Be specific—for example, failure to improve production within 30 days will result in discipline up to and including termination.

- Make sure employees understand what's expected of them in the action plan and are in agreement that the plan is realistic and challenging.

Now we'll discuss the need for continuous feedback between formal appraisals.

The key to superior employee performance is continuous feedback.

- That's why it is important to follow the organization's required schedule for conducting formal performance appraisals.

- Informal performance appraisals can also be helpful, especially if a performance problem arises and the annual review is months away. You can also take advantage of an informal appraisal if an employee makes an outstanding accomplishment that needs to be recognized and/or rewarded outside the regularly scheduled review.

- The main point is to keep the feedback flowing. Create an atmosphere of open communication between you and your employees so that performance issues can be discussed as they happen. With a climate of continuous feedback in place, formal performance appraisal meetings can be much more comfortable and productive events.

Positive reinforcement is a proven effective tool for encouraging outstanding performance, and there are many ways to accomplish this.

- The simplest and easiest way is to verbally acknowledge a good job at the time it's accomplished. This can be in private or in the presence of co-workers and is a verbal "pat on the back" that gives anyone a lift.

- Public recognition of accomplishments is another, more formal and more important, reward for accomplishment. This can be done through an announcement at a companywide meeting, an article with photo in a company newsletter, or even a write-up in the local newspaper.

- Tangible rewards include time off, a new piece of equipment or an upgrade, or a move to a bigger office.

- Monetary rewards include more than raises. Gift certificates or bonuses are also valuable and motivating rewards.

Identifying and dealing effectively with poor performance is also critical.

- Act immediately when you see a performance problem—don't wait for the next performance appraisal. Early intervention is key to successfully dealing with sub-par performance.

- Use tact when you approach an employee about performance problems, but be direct. Focus on the particular issue with specific examples of the behavior—and with encouraging comments and a list of resources to help the employee improve.

- Be prepared to deal with employees' reactions to being criticized. Remain calm if employees get upset. Keep the discussion focused on ways to get performance back on track. Ask for the employees' input on what help they need to improve performance.

- There may be occasions when performance does not improve. Whether it's because employees are unable or unwilling to improve, the written performance appraisal is a valuable tool that documents performance problems as well as plans

for improvement. This document forms a basis from which to adapt improvement plans for employees who are unable to meet certain goals. It also serves as formal, objective evidence of performance problems for employees who are unwilling to improve, and it can be the basis for discipline procedures.

The performance appraisal process helps identify and track problems. In most cases, the outcome is improved performance: Problem solved. But not always.

- The continuous feedback system will keep you in tune with employees so that you can recognize when problems continue.

- Talk with employees as soon as possible after you become aware of a continuing problem and encourage them with specific resources that are available to help them meet expectations. At the same time, be clear about company policy regarding performance.

These are the main points you should take away from this training session.

- You must conduct objective appraisals on a scheduled basis.
- Appraisals tell employees how they're doing and how they can improve.
- Appraisals help create a system of motivation and rewards based on performance.

Do you have any questions about anything we've discussed today concerning performance appraisals?

[*End PowerPoint*]

Okay, now that you are familiar with the idea of performance appraisal, I am going to show you an example rating form.

[*Hand out rating scale and dimension definitions.*]

Take a look at the handout entitled "Dimensions, Definitions, and Behavioral Examples." These 5 dimensions are commonly used to categorize the job performance of managers: Analysis, Decisiveness, Leadership, Confrontation, and Sensitivity.

Please read the definition of each dimension and the example behaviors to yourself while I read them aloud.

Analysis – The ability to identify problems, secure relevant information by effectively asking questions, relate data from different sources, and identify the possible causes of problems.

Behavioral Examples:

- Correctly identifies basic issues, including: data, facts, names/titles of people
- Correctly identifies relationships among: data, people, and problems
- Integrates information across sources
- Recognizes priorities among issues, materials, and data
- Secures relevant information by asking probing questions
- Identifies possible solutions

Decisiveness – Readiness to make decisions, render judgments, take action, or commit oneself; firmly expressing one's opinions and ideas.

Behavioral Examples:

- Makes specific recommendations
- Uses emphatic speech
- Commits to a course of action
- Delineates clear action plans
- Strongly expresses beliefs
- Recognizes the need for immediate action

Leadership – Utilizing appropriate interpersonal styles and methods in guiding individuals (subordinates/peers/superiors) or group toward task accomplishment

Behavioral Examples:

- States goals and purposes for a meeting
- Maintains control of a meeting
- Provides direction/redirects discussion
- Solicits input from employees
- Establishes multiple agendas
- Articulates smooth transitions between topics
- Clarifies roles
- Resists the manipulations of other employees
- Attempts to motivate others

Confrontation – The ability and willingness to disagree or express opposing viewpoints in a tactful style; the willingness to stand up for thoughts and beliefs even when challenged

Behavioral Examples:

- Confronts others about ideas or proposals
- Defends own positions when challenged
- Corrects others
- Voices dissenting opinions
- Challenges the ideas of others
- Asserts and uncommon/unpopular position

Sensitivity – The extent to which an individual shows consideration for the feelings and needs of others, asks for the opinions of others, and gives encouragement.

Behavioral Examples:

- Attentive behaviors (eye contact, nodding, "um")
- Establishes rapport (small talk)
- Uses humor
- Exchanges social pleasantries
- Acknowledges contributions of others
- Does not interrupt others

Does anyone have any questions about these dimensions or the example behaviors from each dimension?

Okay, now I will give you some information about the videotapes that we are about to watch.

[*Hand out Christine Hawkins character sketch.*]

Each of these scenarios depicts a meeting between two employees of a medical supply company. You will be rating the performance of the regional manager, who is meeting with a role player playing the part of a district manager named Christine Hawkins. Prior to the meeting, the regional manager was provided with information that may be useful in the meeting with Christine, including various sales figures and charts.

Here are some things you should know about the meeting. Christine is one of the company's best district managers. In fact, she received an award from the company for her impressive sales numbers. Christine is known to be moody, tyrannical, and obsessive, but she is also a great counselor and trainer. Her employees either love her or hate her.

Christine called this meeting because she wants to fire John Taylor, a poorly performing employee. He has had miserable sales for the last 18 months, and though she worked

with him, he has not improved at all. Since only the regional manager has the authority to terminate him, Christine will try to convince the candidate that this is a necessary step.

You should also know that Christine's customer satisfaction numbers have dropped significantly in recent months, though they are now improving. Christine will try to explain this away by suggesting that she pushed her sales people to focus on new sales, thus somewhat ignoring new customers.

Okay, now take a look at the rating form again. As you watch each videotape, you are going to record specific behaviors for each dimension. Behaviors refer to those things individuals actually do or say. Behaviors are gathered directly from our observation of others; they represent information that has yet to be processed.

For example, the following statements are examples of leadership behaviors: "He solicited information from the role player" or "He provided little overall direction to the meeting." This is not a leadership behavior: "He is a deficient leader." This statement does not tell you anything about what the person actually did.

One way to document behaviors is to write down quotes. So, as you watch the videotapes, write down things that the manager says on your rating form. Then, write down the behavior that the quote exemplifies using the behavioral examples on your definition sheet.

[*Write down following example on white board.*]

For example, if a manager were to ask the role player, "What is your market share?" this would be an example of securing relevant information by asking probing questions, which is an Analysis behavior. So, on your rating form in the Analysis section, you would write down what the manager said, word for word, and beside it you would write something like "asked probing questions" to help you categorize the quote.

Now, when you watch the videotapes, you will see that there is a lot of information to write down. What I recommend is that you write down as many quotes as you can while you are watching the videotape. Then, when the videotape is over, use a different color pen to write down the behavior that the quote exemplifies. You should notice at your work station that I have provided you with red colored pens so you can do this easily.

It is also helpful to note whether the behavior you wrote down is positive, negative, or neutral. To do this, I recommend that you write a +, –, or 0 next to each behavior to indicate whether it is a positive, negative, or neutral behavior.

After you have watched each videotape and recorded the behaviors, you will then assign a rating to each dimension using the scale provided.

[*Read over rating scale and anchors.*]

Your ratings should be based on the behaviors that you recorded in each dimension. For example, if you assign a rating of 4.0 in a given dimension, you should be able to support that rating with the behaviors that you recorded in that dimension. After you have rated each dimension, you will write a global rating at the top of the rating form using the same rating scale. The global rating is your overall impression of the manager's performance across all the dimensions.

[*Allow for 5 minute break before rating videotapes.*]

Okay, let's summarize what you will be doing next. You will be rating 4 different videotapes of 4 different managers interacting with the role player, Christine Hawkins. Record the manager's quotes and behaviors as you observe them on the rating form, and when you are done watching the videotape, please use the red pen to describe your quotes using the example behaviors from each dimension. Then you will record your rating for each dimension in the space provided, as well as your overall rating at the top of the rating form. We will repeat this process for all 4 videotapes.

Are there any questions before we begin?

[*Participants watch and rate first videotape. Collect training materials and rating forms.*]

Now I want you to identify whether each behavior was positive, negative, or neutral by placing either a +, –, or 0 next to the behavior. Then, I want you to identify the dimension that each behavior belongs to by writing the name of the dimension next to the behavior.

[*Allow time to record dimensions.*]

Now I would like you to answer a few questions about the manager and the role player in the videotape.

[*Allow time to complete attractiveness measure.*]

[*Steps are repeated until all 4 tapes have been rated.*]

Okay, now that you have had a chance to observe and rate the job performance of other people, I would like you to complete another survey about your perceptions of job performance. You will notice that this survey is very similar, although not identical, to the first survey you took. Please select the 'Start' button on your desktop and then select

'Run'. In the dialog box, please type in the following exactly as it looks on the white board: (E:\PairedComparison\compare.exe E:\PairedComparison\cmd1.txt). Please type in the ID number that is printed on your folder. Please follow along as I read the instructions.

[*Read the instructions for the Performance Schema Measure.*]

Does anyone have any questions? You may begin?

[*Allow time for participants to complete the Performance Schema Measures (approx. 15 minutes).*]

[*Hand out demographic form.*]

Okay, now I would like you to answer a few questions about yourself. Please read these items carefully and enter your answers on the sheet. These items will only be used for statistical purposes.

[*Collect demographic form.*]

[*Hand out debriefing form.*]

Now I will give you a sheet that gives you some additional information about the study. Please don't discuss the details of this study with anyone. We expect that this study may contribute to the development of future training programs, but this will only happen if participants enter the session uninformed. So please keep quiet about your experiences, other than to suggest to your friends that they can participate.

Thank you very much for your participation. You may be dismissed.

Appendix K

Written Consent to Use Copyrighted Performance Appraisal Lecture Slides

RE: Permission to use Power Point slides

From: Chris Kilbourne <ckilbourne@blr.com>
To: cgorman1 <cgorman1@utk.edu>
Date: Wednesday, March 07, 2007 04:07 PM
Subject: RE: Permission to use Power Point slides

You have our permission to use up to 30 slides from the presentation, provided they are attributed to Business & Legal Reports. We'd appreciate it if you could indicate that the material is copyright 2007 Business & Legal Reports, Inc., and is available in its entirety on HR.BLR.com.

Thank you for respecting our copyright.

Chris Kilbourne

Christopher Kilbourne
Director of Editorial Development and Special Initiatives
Business & Legal Reports, Inc.
860-510-0100
Free Newsletters that Make Your Job Easier:
http://www.blr.com/newsletters

Appendix L

Performance Schema Measure Instructions[1]

This is a survey of your perceptions of managerial job performance. You will be presented with statements of job behaviors that may occur on a typical day in the life of a manager. Read each pair of behaviors and rate how similar they are. Please rate how similar the behaviors are in terms of the meaning that they have for you. Ask yourself: "What does it mean to me about job performance that each of these behaviors happens?" and "Do they mean the same thing about job performance?"

Please click on the number on the scale presented with each pair of behaviors that best indicates the degree of similarity of these behaviors. If the occurrence of both behaviors means the same thing to you, click on "+5" to indicate "very similar." If the occurrence of one behavior means something different to you than the occurrence of the other behavior, click on the number that indicates the degree of that dissimilarity.

The scale presented with each pair of behaviors will look like the following:

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
   -5    -4    -3    -2    -1     0   +1    +2    +3    +4    +5
  Very                                                      Very
Dissimilar                                                Similar
```

For example, the two behaviors to be rated may be:

> A manager adopts others' suggestions when executing tasks.
> A manager gives credit where it is due.

Both of these behaviors may mean that the manager works effectively as part of a group. Thus, a rating of "+5" to indicate "very similar" would be appropriate.

Please remember that people judge behaviors in different ways. This means that there are NO RIGHT OR WRONG answers. Two behaviors that are seen as similar by one person may be seen as dissimilar by another. Everyone's responses are important in this project. It is important for me to know how you as an individual see these behaviors.

You will notice as you respond that the same behavior will appear more than once. This is no trick. What we want you to do is to tell us how similar each behavior is to a number of other behaviors. To do this, we need to present each behavior more than once.

[1] Schema measure instructions adapted from Organizational Research Group (1998)

Your individual responses will not be disclosed to anyone. Only group-level responses will be reported. Your confidentiality is completely assured.

[1] Schema measure instructions adapted from Organizational Research Group (1998)

Appendix M

Job Behavior Statements Included in Performance Schema Measure[1]

ANALYSIS

Correctly identifying basic issues.
Correctly identifying relationships among data, people, and problems.
Securing relevant information by asking probing questions.

DECISIVENESS

Making specific recommendations.
Committing to a specific course of action.
Articulating clear action plans.

LEADERSHIP

Maintaining control of a meeting.
Clarifying roles.
Resisting being manipulated by others.

CONFRONTATION

Defending a position when challenged.
Confronting others about ideas or proposals.
Challenging the ideas of others.

SENSITIVITY

Acknowledging contributions of others.
Establishing rapport.
Not interrupting others.

[1] Dimensions and behaviors extracted and adapted from Tennessee Assessment Center (2002)

Vita

C. Allen Gorman originally hails from McDonough, Georgia. After high school, Allen earned his B. A. in Psychology at the University of Georgia in 2000. From there, he went to the University of Nebraska at Omaha where he received his M. A. in I/O Psychology in 2004.

Allen is currently employed as an assistant professor in the I/O Psychology program at Angelo State University in San Angelo, Texas.