4-2001

Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

Published in Journal of Applied Psychology, 2001 April,
Volume 86, Issue 2, Pages 255-264

Assessor Training Strategies and Their Effects on Accuracy, Interrater Reliability, and Discriminant Validity

Filip Lievens
Ghent University

This study compares the effects of data-driven assessor training with schema-driven assessor training and control training. The sample consisted of 229 industrial and organizational psychology students and 161 managers who were randomly assigned to 1 of these training strategies. Participants observed and rated candidates in an assessment center exercise. The data-driven and schema-driven assessor training approaches outperformed the control training on all 3 dependent variables. The schema-driven assessor training resulted in the largest values of interrater reliability, dimension differentiation, and accuracy. Managers provided significantly more accurate ratings than students but distinguished less between the dimensions. Practical implications regarding the design of assessor trainings and the composition of assessor teams are proposed.

In the field of performance appraisal, rater training has emerged as a useful approach to promote more accurate ratings. Evidence of rating accuracy should ultimately mean that ratings are in line with the norms and values espoused by a specific organization. This feature of defining accuracy as deviations from particular organizational norms instead of from some nonexistent gold standard is reflected in the logic behind frame-of-reference training (Bernardin & Buckley, 1981; Woehr, 1994) because this training approach aims to impose a performance theory on raters (Schleicher & Day, 1998). A performance theory represents common (in a specific organization) conceptualizations of what constitute effective and ineffective behaviors on dimensions. For example, behaviors indicative of effective "decision making" may differ across organizations (e.g., automobile industry vs. insurance business). Raters are expected to use the performance theory imposed as a mental schema (instead of their preexisting schemata) when observing and evaluating ratees. Accordingly, frame-of-reference training may serve as a means to let the organization's values and norms influence the way in which employees are evaluated.

Although frame-of-reference training has shown promise in the performance appraisal literature to promote more accurate ratings (Woehr & Huffcutt, 1994), the generalizability of these findings across persons and domains is uncertain (Arvey & Murphy, 1998, p. 159). First, frame-of-reference training has been almost exclusively studied in student populations. Therefore, it is not known whether the positive results also extend to managers. Second, actual applications of frame-of-reference training to other rating settings than performance appraisal are virtually nonexistent. Hence, it remains uncertain whether the frame-of-reference logic may be effective in other human resource management domains, in which raters-assessors play a central role.

Along these lines, a potentially interesting domain is personnel selection and particularly selection through work sample tests and simulation exercises (as is the case in assessment centers). Traditionally, the training of raters-assessors in this area has followed a behavior-oriented strategy (Thornton, 1992). An example of such a training program was originally proposed by Byham (1977). Since then, this approach has been disseminated in many textbooks on assessment center practice (e.g., Ballantyne & Povah, 1995, p. 94). Still, frame-of-reference training may also be relevant in this rating setting because of its distinct advantage of imposing a performance theory on raters, which may legitimately influence that people are assessed in line with the organizational norms and values.

This advantage and the success of frame-of-reference training in the performance appraisal field (Woehr & Huffcutt, 1994) prompted the idea to train assessors of assessment center exercises according to the frame-of-reference logic. Therefore, this study compares the schema-driven frame-of-reference training with the more traditional behavior-oriented assessor training (and with control training). This study extends prior rater training research by (a) comparing these specific rater training approaches with each other, (b) examining their effects in the domain of rating candidates in an assessment center exercise, and (c) using both students and managers as assessors. The effects are determined through a broad set of dependent variables, namely rating accuracy, interrater reliability, and discriminant validity.

Comparative research on assessor training is both of conceptual and practical importance. From a conceptual point of view, it is important to understand which model of human judgment serves as the best foundation of assessor training. From a practitioner's point of view, it is pivotal to know which training strategy leads to more

---

This research was based on my doctoral dissertation. Parts of this study were presented at the Ninth European Congress of Work and Organizational Psychology, Helsinki, Finland, May 1999.

I would like to acknowledge Michael Harris and Wilfried De Corte for their helpful suggestions on an earlier version of this article. I am also especially grateful to Luc Drieghe and Herlinde Pieters for their assistance in conducting the assessment center sessions.

Correspondence concerning this article should be addressed to Filip Lievens, who is a postdoctoral fellow of the Fund for Scientific Research–Flanders, Belgium (F.W.O.), Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. Electronic mail may be sent to filip.lievens@rug.ac.be.

accurate, reliable, and valid ratings. This could also justify the considerable costs and time involved in assessor training.

Theoretical Background and Hypotheses

Conceptually, rater training approaches might be distinguished according to two cognitive orientations: behavioral oriented and schema oriented (Pulakos, 1986). Rater training approaches then differ in the way raters are taught to process information. Behavior-driven theories (Abelson, 1981; Borman, 1978; Rumelhart & Ortony, 1977), also known as data-driven or bottom-up theories, assume people are able to attend to detailed behavior, to classify these many pieces of factual information into distinct categories, and to form relatively objective and accurate judgements. Consistent with these theories, behavior-driven rater training (e.g., Byham, 1977) divides the rating process into different phases (i.e., behavioral observation, classification, and evaluation). Raters are taught to strictly distinguish these phases (especially observation and evaluation) from each other and to proceed to another phase only when the previous one is finished.

Conversely, frame-of-reference training builds on schema-driven theories of human judgment (Cantor & Mischel, 1977; Fiske & Taylor, 1991; Srull & Wyer, 1989). Raters are then instructed to process information in a top-down manner. There is no strict distinction between observation and evaluation because raters are taught to use the performance theory imposed as a mental scheme to "scan" the behavioral stream for relevant incidents and to form on-line evaluations (Cardy & Keefe, 1994; Day & Sulsky, 1995; Sulsky & Day, 1992, 1994).

The distinction between the two training types might also be cast in terms of different processing objectives (Lichtenstein & Srull, 1987).[1] When raters are not given a specific objective to process information, spontaneous impression formation seems to be the default operation. In data-driven rater training, the formation of such a spontaneous impression is prevented by instilling an observational goal on raters. For example, raters are required to carefully observe behavior, to record behavioral observations, and to classify their observations into dimensions. Schema-driven rater training instills an evaluative goal; namely, alternative schemata imposed on raters instead of their preexisting schemata are expected to guide the impression formation process (Cardy & Keefe, 1994).

On the basis of these conceptual models of human judgment and prior performance appraisal research, I formulated several hypotheses regarding the effects of the alternative training strategies on student and managerial assessors of assessment center exercises. A first series of hypotheses is related to the effects on rating accuracy. I hypothesize that trained assessors will make more accurate ratings than will relatively untrained assessors (Hypothesis 1A) and that assessors receiving a schema-driven training such as frame-of-reference training will be more accurate than will assessors receiving a data-driven training (Hypothesis 1B). Two arguments ground this hypothesis. First, frame-of-reference training is likely to outperform behavior observation training in terms of rating accuracy because the former imposes a performance theory on raters, which should ensure that candidates are rated in accordance with the norms and values of a specific organization. Second, frame-of-reference training has proved its effectiveness in the performance appraisal field. The most recent meta-analysis (Woehr & Huffcutt, 1994) found an average effect size of .83 for studies comparing this training approach with control-no training in terms of rating accuracy. For studies comparing behavior observation training with control-no training, the mean effect size on the dependent variables of rating accuracy was .77. Although actual applications of frame-of-reference training in the assessment center field are lacking, there are methodological reasons to expect the positive effects of frame-of-reference training on rating accuracy to generalize to assessment centers. In fact, the laboratory setting, in which frame-of-reference training has typically been examined (e.g., no ratee-rater acquaintance and short delay period between observation and rating), may more closely mirror an assessment center process than a performance appraisal process may (Pulakos, 1986; Woehr, 1994).

Next, I hypothesize that raters with more experience with the performance domain (e.g., as a result of their affiliation or due to experience with performance appraisal) will rate more accurately (Hypothesis 2). Cardy, Bernardin, Abbott, Senderak, and Taylor (1987) found support for this contention in the performance appraisal domain. Personnel administrators were more accurate than were MBA students, who, in turn, were more accurate than undergraduates were. They also found that the schemata, which developed through experience, explained to some extent the relationship between experience and rating accuracy. Other performance appraisal research (Kozlowski, Kirsch, & Chao, 1986; Kozlowski & Mongillo, 1992) also underscores the role of experience and expertise in promoting accurate ratings.

I also hypothesize that the training strategies will yield different effects depending on the type of assessor (Hypothesis 3). In particular, I expect that it will be easier to train industrial and organizational (I/O) psychology students because they do not possess well-established schemata of managerial situations (Heneman, 1988). Consequently, students may more readily adopt the schemata imposed by training (Schleicher & Day, 1998). Instead, managers may have more difficulty using these schemata because of negative transfer due to preexisting schemata (Fiske & Dyer, 1985). As a result of these well-established schemata, managers may also experience more difficulty learning to withhold their judgments as taught in data-driven assessor training. In summary, the hypotheses regarding rating accuracy are the following:

> *Hypothesis 1A:* Ratings of trained assessors (two training conditions) will have higher rating accuracy than ratings of relatively untrained assessors (control training).
>
> *Hypothesis 1B:* Ratings of schema-driven trained assessors will have higher rating accuracy than ratings of assessors receiving a data-driven assessor training.
>
> *Hypothesis 2:* Irrespective of training, ratings of managerial assessors will have higher rating accuracy than ratings of student assessors.
>
> *Hypothesis 3:* The assessor training strategies will yield different effects on rating accuracy depending on the type of assessor: Training will lead to higher rating accuracy among student assessors than among managerial assessors.

The fourth hypothesis deals with the effects of the training programs on interrater reliability. Generally, when assessors rate assessees on completion of a simulation exercise, interrater reliability has been moderate (Thornton, 1992, p. 129). When asses-

[1] I am indebted to an anonymous reviewer for suggesting this idea.

sors base their ratings on more behaviors and discuss their observations beforehand, interrater reliability has been found to increase (Jones, 1981). Similarly, training may be used to put assessor ratings in line. For instance, the behavior-driven assessor training aims to increase reliability by teaching assessors to process the incoming information in a rigorous data-driven way and by providing assessors with an understanding of the dimension definitions. Frame-of-reference training should ensure that assessors use the same performance theory to observe and rate assessees, enhancing interrater reliability. In short, the following is hypothesized.

Hypothesis 4: Interrater reliability will be higher among trained assessors (two training conditions) than among relatively untrained assessors (control training).

A last series of hypotheses is related to the effects on discriminant validity. Prior research has demonstrated that ratings of different dimensions within a single simulation exercise correlate highly, resulting in weak evidence of discriminant validity (Klimoski & Brickner, 1987). One of the explanations for these blurred dimension distinctions is that they are due to the use of exercise-specific performance schemata on the part of assessors (Lievens 1998b; Zedeck, 1986). In particular, Zedeck suggests that a management performance schema guides the observation and evaluation processes in a simulation exercise. For instance, when assessors encounter behavior in a role-play exercise, this behavior is matched against a management performance schema for that exercise (i.e., expectations regarding managerial behaviors when dealing with subordinate problems). Because a management performance schema is exercise specific, and behaviors associated with such a schema may be categorized in more than one dimension, relatively high correlations between different dimension ratings within an exercise arise. The findings of schemata filling in missing information (Taylor & Crocker, 1981) and directing rater attention to consistent information (Fiske & Taylor, 1991) may further contribute to the dimension overlap.

There is some evidence that these biases due to schema-based processing (e.g., exercise specificity, confirmation bias, and selectivity) may be particularly apparent in managerial ratings. For example, Sagie and Magnezy (1997) concluded that managers, in contrast with psychologists, did not provide distinct ratings on assessment center dimensions. In another study (Lievens, 1998a) managers even scored slightly inferior than I/O psychology students in terms of distinguishing among dimensions. Therefore, I hypothesize that managers will discriminate less among the dimensions than I/O psychology students (Hypothesis 5).

Further, I hypothesize that relatively untrained assessors are more susceptible to the aforementioned negative biases of schema-based processing on discriminant validity than trained assessors (Hypothesis 6A). In addition, I expect differences between the two training approaches in terms of their effects on discriminant validity. As noted by Byham (1977) and Thornton (1992), the traditional data-driven assessor training aims to provide assessors with a "shield" against the biases invoked by schema-based processing. Examples include the strict distinction between behavioral observation and evaluation or the provision of definitions of the dimensions. Conversely, frame-of-reference training attempts to counteract the possible negative effects of schema-based processing by imposing more appropriate schemata on assessors. More concretely, the performance theory provides assessors with a mental framework regarding both the assignment of behaviors by dimension and the correct effectiveness level of each behavior (in line with the organization's norms and values). Accordingly, assessors are expected to place relevant incidents—as they occur—in the appropriate mental category. Because of these advantages, I hypothesize that frame-of-reference training will be a better strategy with which to combat the possible drawbacks of schema-based processing on discriminant validity than will data-driven training (see Hypothesis 6B). In short, the following hypotheses are related to discriminant validity:

Hypothesis 5: Managerial assessors (in the control training condition) will differentiate less among the dimensions than student assessors (in the control training).

Hypothesis 6A: Trained assessors (two training conditions) will differentiate more among the dimensions than relatively untrained assessors (control training condition).

Hypothesis 6B: Assessors trained according to a schema-driven strategy will differentiate more among the dimensions than assessors trained according to a data-driven strategy.

Method

Design

Both student and managerial groups were crossed with the two training conditions and the control condition. This yielded a 2 (assessor type) × 3 (training condition) design.

Sample

The total sample consisted of 390 participants. Two hundred and twenty-nine were I/O psychology students and 161 were managers. The I/O psychology students (132 women and 97 men) participated in the study to receive credit for a human resource management course. Ages ranged from 20 to 30 years with a mean age of 22 years and 11 months (SD = 2.4 years). Fifty-one percent of the students had been in college for more than 4 years and had worked as interns in a psychological consulting firm or in a company's personnel department. However, none of them had previously served as an assessor. The students were randomly assigned to the conditions. Eighty-six students were placed in the control condition, 74 were placed in the data-driven training condition, and 69 were placed in the schema-driven training condition. Sample sizes differ because groups of assessors (about 25 to 30 students) instead of the total group were randomly assigned to the conditions.

The managers (126 men, 35 women) were enrolled in an executive MBA program. Their average age was 33 years and 11 months (SD = 4.3 years, range = 25 to 47 years). The managers had an average of 11.2 years full-time working experience (SD = 5.4 years, range = 1 to 26 years), came from a variety of organizations, and had different functional backgrounds (engineering, sales, etc.). They were also randomly assigned to the conditions. Forty-five managers were placed in the control condition, 62 were placed in the data-driven training condition, and 54 were placed in the schema-driven training condition.

In the context of research on the effects of different training strategies, these samples are relevant for two reasons. First, I/O psychology students and managers are usually asked to attend an assessor training because they have neither an in-depth knowledge nor experience with assessment centers. Second, substantial training effects may be expected rather from I/O psychology students and managers than from I/O professional psychologists.

This sample size resulted in a statistical power of .99 for detecting the assessor main effect, .98 for detecting the training main effect, and .98 for

detecting the interaction effect, assuming a medium effect size at an alpha level of .01 (Cohen, 1988).

Experimental Conditions and Procedure

The students and managers were told to assume the role of assessors and to evaluate four videotaped candidates applying for the job of district sales manager. This provided them with an opportunity to further acquire practical experience in observing and rating candidates. Assessors knew that at the end of the experiment they had to explain their observations and ratings to one another. This common assessment center practice served as an incentive to take the assessor task seriously.

Assessor training. The training program was composed of three main parts: (a) an introduction about the basics of assessment centers; (b) a portrayal of the job and the organization; (c) a workshop on the rating process, which included a lecture, practice, and feedback. Across the conditions, only the workshop differed.

In the introductory lecture, assessment centers were defined and framed in the context of human resource management and personnel selection. The trainer also discussed the components, the purposes, the history, and the current usage of assessment centers.

Second, the assessors received details about the main tasks and qualifications required for successful district sales managers, the working context (e.g., place in organizational tree and number of subordinates), and the organization (e.g., the type of business, the size, pictures of products made, and the organizational culture). This information originated from real materials (i.e., actual job posting and annual report of an organization). Next, the three dimensions, which a job analysis had identified to be crucial for the target job, were presented: problem analysis and solving, interpersonal sensitivity, and planning and organization. Assessors were told that given the target job, they would observe and rate videotaped candidates in a sales presentation exercise. After reviewing this exercise, the rating scales were presented to the assessors.

Third, assessors were randomly assigned to one of the three conditions. The *data-driven assessor training* primarily covered the processes of observing, recording, classifying, and evaluating assessee behavior (see Byham, 1977). In fact, assessors were told that accuracy can be fostered by carefully proceeding through these phases and particularly by strictly distinguishing observation from evaluation.

First, the trainer instructed assessors to make behavioral descriptions of assessee behavior instead of nonbehavioral interpretations. This principle was practiced by a Behavior Example Exercise (Byham, 1977, p. 104). Next, the principle of classifying behavior by dimensions was taught to assessors. To this end, assessors received the dimension definitions. For example, planning and organizing was defined as "the ability to systematically structure own and others' activities to achieve maximum work performance." This principle was practiced by a Behavior Classification Exercise (Byham, 1977, p. 107). The last concept conveyed in the training included the rating of dimensions according to the behavior observed. In fact, in line with Byham (p. 108), assessors were taught that a rating of 5 (*excellent*) meant that a great deal of the dimension was shown, a rating of 3 (*average*) that a moderate amount of the dimension was shown, and a rating of 1 (*poor*) that very little of the dimension was shown.

Assessors were provided practice in recording, classifying, and rating real performances. Specifically, they viewed and evaluated a videotaped candidate in a role-play exercise. Afterward, the trainer elicited a discussion of which behaviors were used to decide an assigned rating, clarifying any discrepancies among ratings. Finally, the trainer provided assessors with feedback pertaining to their ratings.

The *schema-driven assessor training* primarily focused on imposing a frame of reference on the assessors (see Stamoulis & Hauenstein, 1993; Woehr, 1994). In fact, assessors were told that accuracy can be fostered by the assessor knowing what the effective, average, and ineffective examples of behavior are within each dimension in this specific organization and using this information as a mental framework to "scan" the stream of behaviors. Assessors were not taught to distinguish observation from evaluation. Instead, they were told to place candidate behavior—as it was observed—into the performance categories (effective, average, and ineffective).

First, the trainer presented the definitions of the dimensions and discussed examples of normative behaviors representing different levels of performance on each of the three dimensions. For instance, normative behaviors representing a 5 on the interpersonal sensitivity dimension were differentiated from normative behaviors representing a 3 or 1 on the same dimension. Assessors were then presented with a written exercise, which listed 20 incidents (see Woehr, 1994, p. 529, for a similar exercise). Assessors had to assign each incident to one of the three dimensions and to one of the three performance categories. Afterward, the trainer discussed the answers and provided feedback.

Assessors were provided practice in rating real assessee performances. Specifically, they viewed and evaluated a videotaped candidate in a role-play exercise. Afterward, the trainer elicited a discussion of how the assessors decided an assigned rating, clarifying any discrepancies among ratings. Finally, the trainer provided assessors with feedback pertaining to their ratings.

In the control (i.e., minimal) training condition, no specific training concepts were conveyed. Assessors were told that they were expected to watch videotaped assessees, take notes if necessary, and provide dimensional ratings. Participants rated the practice tape (see other conditions), but their ratings were not discussed and no feedback was provided.

Procedure. After the random assignment to one of the training conditions, assessors observed the videotaped sales presentation performance of each of the four candidates. Participants used an observation form to record behavior. After each performance, assessors independently rated the candidate on three dimensions (i.e., problem analysis and solving, interpersonal sensitivity, and planning and organization) using a 5-point scale, ranging from 1 (*poor*) to 5 (*excellent*). To control for order effects, four versions of the integral film were developed. Groups of assessors were randomly assigned to a particular version. After all candidates were evaluated, assessors met in teams to share observations, discuss ratings, and write assessee reports.

The entire procedure, including several breaks, lasted about 1 day (i.e., 6 hr). Because of this procedure's longer time span and because it provided assessors with content (e.g., job posting) and context (e.g., organization) information, most of the recommendations to ensure realism in laboratory experiments were followed (see Ilgen, Barnes-Farrell, & McKellin, 1993).

Videotaped Assessee Performances

Videotaped performances of four candidates in a sales presentation exercise were developed. These candidates applied for the position of district sales manager in the organization "Plafox." As part of the selection procedure, each candidate was expected to deliver a sales presentation and argue which of three software systems was most appropriate. This presentation was given to a panel of decision makers who regularly challenged the candidate.

The performances were designed to vary along three dimensions: problem analysis and solving, interpersonal sensitivity, and planning and organizing. To this end, 20 assessors (15 men, 5 women; mean age = 36 years) were asked to provide behaviors that would cause them to judge an assessee as being higher or lower on a specific dimension in a sales presentation exercise. These assessors qualified as experts due to (a) their practical experience as assessors (mean assessor experience = 6 years), (b) their theoretical knowledge of assessment centers, and (c) their familiarity with assessment center research. After eliminating redundancies, this resulted in a list of 45 behaviors. Next, scripts of four performances were written. For reasons of realism, two experienced assessors (2 women; mean age = 33.5 years; mean assessor experience = 4 years) helped writing the

scripts. Next, semiprofessional actors were filmed delivering their scripted sales presentations. On average, each videotaped performance ran about 7 min.

Target scores were estimated on the basis of procedures by Sulsky and Balzer (1988). Five experienced assessors (3 men, 2 women; mean age = 30 years; mean assessor experience = 4 years) were provided with the job posting, the organizational information, and the organizational values. For instance, this organization particularly valued behaviors indicative of "working smarter," "customer orientation," "selling solutions not products," and "people-centered leadership." Using these materials, the expert assessors reached consensus about what constituted effective and ineffective performances on the dimensions for this organization ("performance theory"). Next, they viewed each videotaped performance under optimal conditions. This meant that they could view the performances repeatedly and rewind them. All experts independently rated each performance on a 5-point scale, 1 (*poor*) to 5 (*excellent*). Interrater agreement among the expert ratings equaled .9 (intraclass correlation 2.1, Shrout & Fleiss, 1979). On the basis of the ANOVA approach outlined in Borman (1978), convergent and discriminant validities of the expert ratings were computed and equaled .56 and .58, respectively. These validities were satisfactory as they were similar to the values obtained in Borman's study. Target scores were obtained by averaging the expert ratings. The experts also rated the realism of the videotaped performances on a 9-point scale: 1 (*not at all realistic*) to 9 (*very realistic*). The mean realism ratings were 8.00 (Candidate 1), 7.20 (Candidate 3), 7.00 (Candidate 4), and 5.80 (Candidate 2).

Manipulation Checks

The number[2] of normative and behavioral descriptions recorded by assessors on the observation forms served as manipulation checks. A normative description is a statement that reflects the performance norms espoused by the organization (Schleicher & Day, 1998). Examples of normative descriptions were given during the schema-driven training. A behavioral description is "a behavioral statement that specifically describes what an assessee says or does" (Gaugler & Thornton, 1989, p. 613).

In a preliminary phase, two I/O psychology students coded notes of 15 assessors randomly selected from the assessor pool of this study. These students had been in college for more than 4 years and were unaware of the study's purpose. They were familiarized with the performance theory, including the normative behaviors (see schema-driven training) and with the difference between behavioral and nonbehavioral descriptions (see data-driven training). Next, they independently coded each note into one of the categories (i.e., "normative description," "behavioral description," or "does not apply"). Cohen's (1960) kappa, a coefficient of chance-corrected interrater agreement for nominal scales, was computed and equaled .84. Discrepancies between coders were discussed and resolved. Because the level of interrater agreement among these coders was relatively high and because the observation forms of the remaining 375 assessors yielded a total of 19,551 written descriptions, it was decided to divide the forms in two parts. Each student was randomly assigned one part of observation forms and coded the descriptions of these forms.

Because of the insignificant correlation ($r = .02$) between the two manipulation check variables, two one-way ANOVAs were conducted with training condition as the between-subject factor. A significant training effect was found for the number of behavioral descriptions, $F(2, 383) = 41.75$, $p < .001$. Consistent with expectations, planned comparison tests revealed that assessors receiving data-driven training ($M = 50$, $SD = 20$) noted significantly ($p < .001$) more behavioral descriptions than did assessors receiving either schema-driven training ($M = 38$, $SD = 15$) or control training ($M = 32$, $SD = 13$).

A significant training effect was also found for the number of normative descriptions, $F(2, 383) = 99.52$, $p < .001$. As expected, planned comparison tests showed significant differences ($p < .001$) in terms of the number of normative descriptions between schema-driven trained assessors ($M = 14$, $SD = 9$) and assessors receiving either data-driven training ($M = 5$, $SD = 3$) or control training ($M = 5$, $SD = 3$). In short, these results confirm that the assessors approached their task according to the principles taught in the training.

Analyses

To examine the hypotheses regarding rating accuracy, a 2 × 3 (Assessor Sample × Training Condition) MANOVA was conducted with two differential accuracy indices as dependent variables. In particular, Cronbach's (1955) deviational measure of differential accuracy (DA) was computed (see Sulsky and Balzer, 1988, for the exact formula). DA is an index of the assessor's ability to differentiate among assessees within dimensions. Lower scores on DA indicate better accuracy. Cronbach's other rating accuracy measures were not used because these measures aggregate across dimensions, ratees, or both, and therefore do not directly assess the degree to which performance is accurately rated per ratee on each dimension (Schleicher & Day, 1998).

Besides Cronbach's DA, Borman's (1977) differential accuracy (BDA) was used to assess correlational accuracy or the correlation between ratings on each dimension and the corresponding target scores across assessees. Higher scores on BDA indicate better accuracy. Because BDA provides only correlational information, it is closer to rating validity and is not equivalent to Cronbach's DA. However, BDA tends to correlate significantly with distance accuracy components (Sulsky & Balzer, 1988). Because in this study DA and BDA also significantly ($-.62$, $p < .001$) correlated, MANOVA was preferred to two ANOVAs.

To examine the hypotheses regarding interrater reliability and discriminant validity, I used generalizability analysis (Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). As opposed to classical test theory, generalizability theory regards measurement error as multifaceted. Accordingly, it permits the simultaneous estimation of many sources of variance (e.g., assessors and dimensions) (Kane, 1982; Kraiger & Teachout, 1990; Marcoulides, 1989). In each training condition and assessor sample, the design underlying the generalizability analysis was a two-facet design: assessors (A) and dimensions (D). Assessors and dimensions were completely crossed with each other. Candidates (C) were treated as the object of measurement. In the generalizability analyses, evidence of interrater reliability was derived from the variance component of the Assessors × Candidates interaction (Kane, 1982; Kraiger & Teachout, 1990). This interaction indicates whether assessors differ in their rank ordering of candidates.

Evidence of discriminant validity was derived from the variance component of the Candidates × Dimensions interaction (Kane, 1982; Kraiger & Teachout, 1990). A high value for this interaction indicates substantial differences between dimensions in evaluations of candidates.

Results

Rating Accuracy

According to Hypothesis 1A and 1B assessor training would impact on rating accuracy. Table 1 presents the means and standard deviations by training condition. Consistent with these hy-

[2] A problem with using the number of behavioral descriptions instead of the ratio of the number of behavioral descriptions to the number of descriptions may be that assessors who write down more have a greater chance of receiving higher scores on the number of behavioral descriptions (Gaugler & Thornton, 1989). The same is true for the number of normative descriptions. Therefore, I also computed the ratios. Manipulation check results were similar.

potheses, the 2 × 3 (Assessor Sample × Training Condition) MANOVA showed a significant multivariate main effect for training condition, $F(4, 766) = 13.18$, $p < .001$, Wilks's $\lambda = .88$. A follow-up discriminant analysis yielded one significant eigenvalue, with training condition accounting for 94.4% of the variance in the linear accuracy composite. The structure coefficients from this discriminant analysis indicated that both DA and BDA were driving the discrimination between the different training conditions (−.93 and .84, respectively). This is also reflected in the respective effect sizes (DA $\eta^2 = .11$ and BDA $\eta^2 = .09$).

To further examine Hypothesis 1A, I used a planned comparison test to evaluate the prediction that ratings of untrained assessors would differ from ratings of assessors trained according to either a data-driven or a schema-driven strategy. The planned comparison test was significant ($p < .001$) for both DA and BDA, with trained assessors showing higher differential accuracy than assessors in the control training. Consistent with Hypothesis 1B, a planned comparison test was significant for both accuracy indices (BDA, $p < .01$; DA, $p < .05$), with schema-driven trained assessors being more accurate than assessors receiving a data-driven training.

Hypothesis 2 predicted that managers were more accurate assessors than students. Table 2 presents the means and standard deviations by assessor type. As hypothesized, the MANOVA showed a significant multivariate main effect for assessor type, $F(2, 383) = 5.59$, $p < .01$, Wilks's $\lambda = .97$. An examination of the structure coefficients from a follow-up discriminant analysis revealed that BDA moderately contributed to the discrimination between the managerial and student assessors (.60), whereas DA contributed little (.29). Inspection of the size of the assessor type effect on differential accuracy confirmed this picture (for BDA $\eta^2 = .02$, for DA $\eta^2 = .003$). In addition, a t test showed only a significant comparison for BDA, $t(388) = -2.35$, $p < .05$, with managers being more accurate (see Table 2). These results indicate that, irrespective of training, ratings of managers showed significantly higher differential accuracy (correlational component) than did ratings of students.

Finally, Hypothesis 3 stated that the training approaches would exert different effects on accuracy depending on the type of assessor. Contrary to this hypothesis, the MANOVA showed no significant multivariate Assessor Type × Training Condition interaction effect ($F < 1$).

Table 1
Means and Standard Deviations of Accuracy Indices by Training Strategy

Accuracy index	Control training (n = 131)		Data-driven training (n = 136)		Schema-driven training (n = 123)	
	M	*SD*	*M*	*SD*	*M*	*SD*
BDA	0.89_a	.56	1.13_b	.70	1.41_c	.71
DA	1.18_a	.23	1.04_b	.24	0.97_c	.24

Note. BDA = Borman's differential accuracy; DA = Cronbach's differential accuracy. Higher scores for BDA indicate greater accuracy, whereas lower scores for DA indicate greater accuracy. Means with different subscripts are significantly different at $p < .01$, with the exception of the difference between schema-driven training and data-driven training on DA, which was significant only at $p < .05$.

Table 2
Means and Standard Deviations of Accuracy Indices by Assessor Type

Accuracy index	Psychology students (n = 229)		Managers (n = 161)	
	M	*SD*	*M*	*SD*
BDA	1.07_a	0.66	1.24_b	0.72
DA	1.06_a	0.24	1.07_a	0.28

Note. BDA = Borman's differential accuracy; DA = Cronbach's differential accuracy. Higher scores for BDA indicate greater accuracy, whereas lower scores for DA indicate greater accuracy. Means with different subscripts are significantly different at $p < .05$.

Interrater Reliability

The results of the generalizability analyses of both samples in the various conditions are presented in Table 3. Included are the degrees of freedom, the estimated variance components for each effect (together with their 90% confidence intervals). The variance components represent the variances of the mean candidate ratings attributable to the candidates, to the assessors, to the dimensions, and to the respective interactions among them. Confidence intervals were computed by procedures outlined in Brennan (1992). The percentage contributions of each effect are also given. This percentage contribution refers to the percentage of the sum of the variance components (i.e., the total variance) that each variance component accounts for.

Hypothesis 4 stated that interrater reliability would be higher in the training conditions than in the control training condition. As noted above, evidence of interrater reliability is derived from lower values of the variance component of the Assessors × Candidates (AC) interaction. Table 3 shows that this variance component varied substantially[3] across conditions. In the control training condition, the Assessors × Candidates (AC) interaction accounted for 16.1% in the student sample and 20.5% in the managerial sample. This value decreased to 11.8% for student assessors and 8.1% for managerial assessors in the data-driven training condition. In the schema-driven training condition, the contribution of the Assessors × Candidates (AC) interaction further decreased to 11.0% in the student sample and 4.4% in the managerial sample.

Generalizability coefficients were also computed (see Table 4). These coefficients are intraclass correlations similar in form to the classical reliability coefficients. Values equal or above .80 are considered to be acceptable (Marcoulides, 1989). In the control training, five out of six generalizability coefficients were not acceptable. Alternatively, all generalizability coefficients of dimensions rated by schema-driven trained assessors were above the

[3] A drawback of generalizability analysis is that "no guidelines have been offered for gauging what is to be considered a small, moderate, or large variance component" (Kraiger & Teachout, 1990, p. 32). To limit subjectivity in interpretation of variance components, I followed Kraiger and Teachout and made a priori predictions about the relative size of the variance components. In addition, the percent contribution served as a heuristic to interpret the magnitude of the components (Shavelson & Webb, 1991).

Table 3
Results of Generalizability Analysis of Managers and Students in Various Training Conditions

Effect	Students				Managers			
	df	*VC*	90% confidence intervals	Explained variance (%)	*df*	*VC*	90% confidence intervals	Explained variance (%)
Control assessor training								
A	85	.00	.00 < VC < .00	0.1	44	0[a]		
C	3	.01	.01 < VC < .03	1.2	3	.31	.16 < VC < .91	19.1
D	2	.01	.01 < VC < .03	0.6	2	0[a]		
AC	255	.19	.14 < VC < .28	16.1	132	.33	.24 < VC < .49	20.5
AD	170	.02	.01 < VC < .06	1.5	88	.01	.01 < VC < .03	0.8
CD	6	.36	.19 < VC < 1.05	31.0	6	.35	.18 < VC < 1.02	21.5
ACD	510	.58	.52 < VC < .65	49.5	264	.62	.54 < VC < .71	38.1
Data-driven assessor training								
A	72	.01	.01 < VC < .03	0.4	60	.09	.05 < VC < .23	5.0
C	3	0[a]			3	.20	.10 < VC < .59	11.1
D	2	0[a]			2	0[a]		
AC	216	.20	.15 < VC < .29	11.8	180	.15	.10 < VC < .25	8.1
AD	144	.04	.02 < VC < .12	2.3	120	.02	.01 < VC < .06	1.1
CD	6	.82	.42 < VC < 2.40	47.3	6	.68	.35 < VC < 1.99	37.3
ACD	432	.66	.60 < VC < .73	38.2	360	.68	.60 < VC < .77	37.4
Schema-driven assessor training								
A	65	.02	.01 < VC < .06	1.2	51	.07	.04 < VC < .17	3.3
C	3	0[a]			3	.17	.09 < VC < .50	8.5
D	2	0[a]			2	0[a]		
AC	195	.20	.15 < VC < .29	11.0	153	.09	.05 < VC < .23	4.4
AD	130	.09	.06 < VC < .17	4.8	102	.03	.02 < VC < .09	1.4
CD	6	.92	.47 < VC < 2.69	51.2	6	.92	.47 < VC < 2.69	45.8
ACD	390	.57	.51 < VC < .64	31.8	306	.73	.64 < VC < .84	36.7

Note. The results of six separate generalizability analyses are displayed. VC = estimated variance components; A = assessors; C = candidates; D = dimensions.
[a] Consistent with recommendations of Shavelson and Webb (1991), small negative estimates of variance components were reported as zero.

acceptable level of .80. In the data-driven assessor training, two coefficients were below this level. All these results support Hypothesis 4. In addition, although not hypothesized, schema-driven training slightly outperformed data-driven training in terms of interrater reliability.

Discriminant Validity

Hypothesis 5 stated that in the control training condition, managerial assessors would differentiate less among the dimensions than would student assessors. In line with Hypothesis 5, the Candidates × Dimensions (CD) interaction was smaller for managerial assessors (21.5%) than it was for student assessors (31.0%).

Hypothesis 6A expected trained assessors to differentiate more among dimensions than untrained assessors. According to Hypothesis 6B, assessors receiving schema-driven training would differentiate more among dimensions than would data-driven trained assessors. Consistent with these hypotheses, the value of the Candidates × Dimension (CD) interaction was highest in the schema-driven assessor training condition (51.2% in student sample and 45.8% in managerial sample; see Table 3). These values were substantially higher than were the values found in the control training condition (31.0% for students and 20.1% for managers; see Table 3) and were somewhat higher than the values found in the data-driven assessor training condition (47.3% in student sample and 37.3% in managerial sample; see Table 3).

Because previous studies examined discriminant validity using the multitrait-multimethod matrix (Campbell & Fiske, 1959), I also computed the mean heterotrait-monomethod correlation[4] to test Hypotheses 5, 6A, and 6B. Lower values for the mean heterotrait-monomethod correlation indicate higher discriminant validity. These multitrait-multimethod results corresponded to the generalizability analyses. For example, for managerial assessors the mean heterotrait-monomethod correlation was .17 in the schema-driven training, .24 in the data-driven training, and .39 in the control training. This correlation was also always lower for student assessors than for managerial assessors.

[4] Multitrait–multimethod correlations are less appropriate here because they are based on ratings of only four candidates. In addition, because each assessor rated all four candidates, I was forced to compute these correlations by only using the ratings of a single candidate randomly selected per assessor.

Table 4
Generalizability Coefficients per Dimension for Various Samples and Training Groups

Dimension	Control training	Data-driven training	Schema-driven training
Problem analysis and solving			
Students	.62	.72	.80
Managers	.73	.76	.80
Interpersonal sensitivity			
Students	.79	.85	.88
Managers	.85	.86	.91
Planning and organization			
Students	.64	.77	.83
Managers	.70	.87	.85

Note. These generalizability coefficients were obtained by separate generalizability studies within each dimension. Assessors were the facets. Candidates were the object of measurement.

Discussion

Assessor Training Strategy

This study contrasted the effectiveness of schema-driven (frame-of-reference) training to data-driven training (and control training) for assessors of assessment center exercises. As a first conclusion, this study shows that both the data-driven and schema-driven assessor training strategies clearly outperform the control training. In fact, for assessors in the control training, the analyses reveal substantially lower values indicative of rating accuracy, interrater reliability, and dimension differentiation. Additionally, assessors in the control training condition record significantly less behavioral descriptions than do trained assessors. This control condition consisted of a minimal assessor training, which included information on assessment centers, the job, the organization, the dimensions, the rating scales, and a practice videotape. Apparently, such minimal training is not sufficient to establish adequate levels of accuracy, reliability, and discriminant validity. Therefore, this study confirms the necessity of thoroughly training people prior to serving as assessors (Task Force on Assessment Center Guidelines, 1989).

Second, the present study shows that an assessor training based on the principles behind frame-of-reference training is a better strategy than is behavior observation training to have a performance theory legitimately influence performance evaluations. This conclusion is documented by the higher interrater reliability for assessors receiving frame-of-reference training ratings as compared with data-driven trained assessors. More importantly, rating accuracy was also higher in the frame-of-reference training condition than in the data-driven training condition. This means that the performance theory imposed on assessors ensures that they rate candidates in accordance with the norms and values of a specific organization.

This result contributes to the extant rater training literature because it extends the effectiveness of frame-of-reference training (Woehr & Huffcutt, 1994) to the domain of rating assessment center exercises. The finding of no Training Condition × Assessor Type interaction effect for rating accuracy is also good news for proponents of frame-of-reference training because this means that this training is equally effective for managerial and student assessors.

Third, the schema-driven frame-of-reference training also improves the quality of construct measurement, as evidenced by the higher discriminant validity found. This means that assessors trained according to frame-of-reference training are better able to use the various dimensions differentially. This result contributes to prior research showing that the quality of the output of assessment centers is linked to major assessment center design parameters (see Lievens, 1998b, for a review).

The practical implication of these findings is that an assessor training should include the logic underlying frame-of-reference training. To this end, the following procedure might be used. First, the norms, values, and personal qualities that an organization considers to be crucial to sustain its competitive advantage are made explicit and translated into a performance theory. Each organization may use its own nomenclature to state the normative performance indicators and effectiveness levels (Schleicher & Day, 1998). Next, in a workshop this theory of performance is imposed on raters/assessors. When actually rating candidates of assessment center exercises (and employees), raters are then expected to use this mental framework in a top-down manner to scan the stream of behaviors for relevant incidents and to provide on-line evaluations in light of the organizational norms. In this way, imposing a performance theory on assessors through frame-of-reference training serves as a means to let the organizational values legitimately influence the way candidates are rated. Future research is needed to examine how organizations can formulate such a performance theory, use it in various human resource management domains, and adjust if necessary.

As already noted, at least two elements distinguish frame-of-reference training from other rater training approaches. A first element is the performance theory imposed on raters. A second element is the assumption that assessors use this performance theory as a mental schema to process the incoming information (i.e., scan the behavioral stream and form on-line evaluations) (Cardy & Keefe, 1994; Day & Sulsky, 1995; Sulsky & Day, 1992, 1994). In this way, the schema-driven frame-of-reference approach aims to circumvent the possible pitfalls of schema-based processing by imposing new and more appropriate schemata on assessors. In the social cognitive literature this approach is labeled as "making alternative schemas more salient" (Fiske & Taylor, 1991). Alternatively, the more traditional data-driven strategy aims to counteract possible drawbacks of schema use by teaching assessors to observe behavior and to withhold early interpretations. In the social cognitive literature this practice is referred to as "providing debiasing instructions," namely providing participants with a specific technique to successfully undercut their erroneous impressions (C. G. Lord, Lepper, & Preston, 1984). Because higher levels of accuracy, interrater reliability, and discriminant validity were found for the schema-based frame-of-reference approach than for the data-driven approach, a theoretical contribution of this study is that schema-driven approaches may be preferable to data-driven approaches in order to improve rating accuracy and quality.

Finally, two cautionary remarks should be made on the address of schema-driven training. First, the manipulation checks of this study show that the schema-driven training yields significantly less behavioral descriptions than the traditional data-driven training.

This could mean that a schema-driven training strategy may be superior to other training approaches in terms of rating accuracy but not in terms of behavioral accuracy because schema-driven training does not explicitly teach assessors to separate observation and interpretation. A limitation of this study was that behavioral accuracy was not a dependent variable. In fact, the number of behavioral observations recorded, as used in the manipulation checks, can at best be interpreted as an index of the quality of assessor notes. Because behavioral accuracy is especially important for the feedback in developmental assessment centers, future research should use signal detection theory (R. G. Lord, 1985) to examine this issue more closely.

A second cautionary remark on the address of schema-driven training is related to another limitation of this study, namely that assessors rated candidates in only one assessment center exercise (i.e., sales presentation). This was done to keep the rating task manageable. Therefore, this study should be regarded as a first demonstration of the effectiveness of the schema-driven frame-of-reference approach for promoting more accurate, reliable, and valid ratings of assessment center candidates. Future research is needed to extend these findings to other assessment center exercises, to other dimensions, and to working assessment centers. Analogously, the stability of these findings over longer time periods should be examined.

Assessor Type

On the one hand, this study reveals that managers had somewhat more difficulty in using dimensions differentially, as evidenced by the lower percentage of variance due to the Candidates × Dimension interaction in the managerial sample than in the student sample. This corroborates prior research, which found lower criterion-related (Gaugler et al., 1987) and discriminant validity (Lievens, 1998a; Sagie & Magnezy, 1997) for managerial assessors as compared with psychologist assessors. On the other hand, a more positive result for managerial assessors is that they rate candidates with higher differential accuracy (i.e., correlational component) than psychology students rate. This implies that managers' ratings converge more closely to the target ratings, which reflect the norms and values of the organization.

A plausible explanation for these results is that managers are better able to derive relevant performance indicators from the portrayal of the organization presented to them. Accordingly, this finding confirms the belief that managerial assessors take organizational norms and values into account when evaluating assessment center candidates (Klimoski & Brickner, 1987). Another explanation is that in the control condition managers rely on their preexisting schemata to rate candidates. The use of these prior expectations and knowledge structures, in turn, results in higher accuracy. This parallels findings in the performance appraisal field documenting the positive relationship between experience, use of performance schemata, and rating accuracy (Cardy et al., 1987).

On the basis of these results with respect to the use of managers as assessors in assessment center exercises, the following applied implications can be stated. When assessment center exercises are used *for hiring-selection purposes,* managerial assessors should be included in assessor teams. As shown by this study, inclusion of managers leads to more accurate (i.e., more in line with the organizational norms and values) candidate ratings, ensuring a better fit between the person hired and the organization. When assessment center exercises are used *for developmental purposes,* practitioners should realize that managerial assessors have more difficulty providing distinct ratings on the dimensions, which is a prerequisite for detailed attribute-based feedback to the candidates.

References

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist, 36,* 715–729.

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

Ballantyne, I., & Povah, N. (1995). *Assessment and development centres.* Aldershot, England: Gower.

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

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

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63,* 135–144.

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Byham, W. C. (1977). Assessor selection and training. In J. L. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 89–125). New York: Pergamon Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cantor, N., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. *Journal of Personality and Social Psychology, 35,* 38–48.

Cardy, R. L., Bernardin, H. J., Abbott, J. G., Senderak, M. P., & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational and Organisational Psychology, 60,* 197–205.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Day, D. V., & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80,* 158–167.

Fiske, S. T., & Dyer, L. M. (1985). Structure and development of social schemata: Evidence from positive and negative transfer effects. *Journal of Personality and Social Psychology, 48,* 839–852.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition.* Reading, MA: Addison-Wesley.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

Gaugler, B. B., & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74,* 611–618.

Heneman, R. L. (1988). Traits, behaviors, and rater training: Some unexpected results. *Human Performance, 1,* 85–98.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54,* 321–368.

Jones, A. (1981). Inter-rater reliability in the assessment of group exercises at the UK assessment centre. *Journal of Occupational and Organisational Psychology, 54,* 79–86.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6,* 125–160.

Klimoski, R. J., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40,* 243–260.

Kozlowski, S. W. J., Kirsch, M. P., & Chao, G. T. (1986). Job knowledge, ratee familiarity, conceptual similarity and halo error: An exploration. *Journal of Applied Psychology, 71,* 45–49.

Kozlowski, S. W. J., & Mongillo, M. (1992). The nature of conceptual similarity schemata: Examination of some basic assumptions. *Personality and Social Psychology Bulletin, 18,* 88–95.

Kraiger, K., & Teachout, M. S. (1990). Generalizability theory as construct-related evidence of the validity of job performance ratings. *Human Performance, 3,* 19–35.

Lichtenstein, M., & Srull, T. K. (1987). Processing objectives as a determinant of the relationship between recall and judgement. *Journal of Experimental Social Psychology, 23,* 93–118.

Lievens, F. (1998a, August). *Assessment centers and what they measure: Disentangling assesee, assessor, exercise, and dimension effects.* Paper presented at the Annual Meeting of the Academy of Management, San Diego, CA.

Lievens, F. (1998b). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6,* 141–152.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47,* 1231–1243.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70,* 66–71.

Marcoulides, G. A. (1989). The application of generalizability analysis to observational studies. *Quality and Quantity, 23,* 115–127.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes, 38,* 76–91.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–136). Hillsdale, NJ: Erlbaum.

Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70,* 103–108.

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

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review, 96,* 58–83.

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.

Sulsky, L. M., & Balzer, W. K. (1988). The meaning and measurement of performance rating accuracy: Some methodological 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). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology, 79,* 535–543.

Task Force on Assessment Center Guidelines. (1989). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management, 18,* 457–470.

Taylor, S. E., & Crocker, J. (1981). Schematic bases of social information processing. In E. T. Higgins, S. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario symposium* (pp. 89–134). Hillsdale, NJ: Erlbaum.

Thornton, G. C., III. (1992). *Assessment centers in Human Resource Management.* Reading, MA: Addison-Wesley.

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

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

Zedeck, S. (1986). A process analysis of the assessment center method. *Research in Organizational Behavior, 8,* 259–296.

Received August 16, 1999
Revision received April 4, 2000
Accepted April 4, 2000 ■