A New Frame for Frame-of-Reference Training: Enhancing the Construct Validity of Assessment Centers

Deidra J. Schleicher
University of Tulsa

Bronston T. Mayes
California State University, Fullerton

David V. Day
The Pennsylvania State University

Ronald E. Riggio
Claremont McKenna College

The authors undertook a comprehensive examination of the construct validity of an assessment center in this study by (a) gathering many different types of evidence to evaluate the strength of the inference between predictor measures and constructs (e.g., reliability, accuracy, convergent and discriminant relationships), (b) introducing a theoretically relevant intervention (frame-of-reference [FOR] training) aimed at improving construct validity, and (c) examining the effect of this intervention on criterion-related validity (something heretofore unexamined in the assessment center literature). Results from 58 assesses and 122 assessors suggested that FOR training was effective at improving the reliability, accuracy, convergent and discriminant validity, and criterion-related validity of assessment center ratings. Findings are discussed in terms of implications and future directions for both FOR training and assessment center practice.

From its initial use for military officer selection in the 1940s, the assessment center (AC) has become a popular technique for assessing managerial potential. Debates over the validity of this method are also popular. After a brief review of the evidence for AC validity, the present study examines the efficacy of a specific intervention for improving the validity of ACs.

AC Validity

The most impressive and consistent evidence for the validity of ACs comes in the form of criterion-related validity. From Bray and Grant’s (1966) early validation of AT&T’s Management Progress Study, to more recent meta-analyses (Gaugler, Rosenthal, Thornton, & Benston, 1987; Hunter & Hunter, 1984; Schmitt, Gooding, Noe, & Kirsch, 1984), there have been consistent findings of moderate to high validity coefficients for ACs in predicting managerial success. The evidence for the construct validity of ACs, however, has been less clear. Although researchers have often found evidence of convergent validity (in the form of a general trait or person factor), it has been accompanied by a corresponding lack of discriminant validity (Archambeau, 1979; Neidig, Martin, & Yates, 1979; Sackett & Dreher, 1982; Shore, Thornton, & McFarlane Shore, 1990; Turnage & Muchinsky, 1982). This robust finding of method variance (referred to as the exercise effect) has led many researchers to question the construct validity of assessment ratings. Interventions such as reducing the number of dimensions to be rated (Bycio, Alvares, & Hahn, 1987), using behavioral checklists (Reilly, Henry, & Smither, 1990), and varying the rating instructions (Silverman, Dalessio, Woods, & Johnson, 1986) have been proposed in an attempt to improve the construct validity of ACs.

Although such interventions have shown some success, they are lacking in a couple regards. First, these investigations generally have solely looked at the effect of interventions on convergent and discriminant relationships within AC ratings (i.e., mono- and hetero-trait and method correlations). However, recent conceptualizations of validity have invoked a broader approach to establishing validity, and such convergent and discriminant relationships represent only one type of evidence for construct validity. Second, previous investigations have ignored theoretically relevant interventions from other literatures. Specifically, the role of assessor training in improving the construct validity of ratings has been ignored, despite the theoretical and empirically documented importance of training for improving the quality of judgments. Finally, as noted by Klimgoski and Brickner (1987), previous studies examining variables thought to improve the construct validity of assessment ratings have not examined the effect on criterion-related validities, even though criterion-related validity is an important aspect of construct validity (Landy, 1986). Others have noted the possibility that criterion-related validities would improve if abilities were better measured within exercises (e.g., Bycio et al.,...
1987; Sackett, 1987; Silverman et al., 1986) and have encouraged such examinations. The fact that the impact of an improvement in construct validity on the criterion-related validity of ACs has not been examined seems to be a particularly large oversight given that it is the predictive validity by which the effectiveness of ACs is most likely to be judged. In the present study, we attempted a more comprehensive examination of the construct validity of an AC by (a) widening the lens for viewing evidence of construct validity, (b) introducing a theoretically relevant training intervention aimed at improving construct validity, and (c) examining the impact of this intervention on criterion-related validity as well as construct validity.

Assessor Training

As set forth by the guidelines on ACs (Task Force on Assessment Center Guidelines, 1989; Task Force on Assessment Center Standards, 1980), assessor training is an integral component of the AC program. The potential for increasing the reliability and the construct validity of ratings through more careful attention to assessor training has been emphasized by many authors (e.g., Bray & Grant, 1966; Norton, 1981; Silverman et al., 1986; Spychalski, Quinones, Gaugler, & Pohley, 1997; Thomson, 1970; Turnage & Muchinsky, 1982). Training has also been specifically singled out as an important determinant of the strength of exercise factors (Schmitt, Schneider, & Cohen, 1990; Turnage & Muchinsky, 1982). Despite all the suppositions, however, there is no empirical research on the effectiveness of various assessor training strategies for improving AC validity. Even the guidelines on ACs (Task Force on Assessment Center Guidelines, 1989; Task Force on Assessment Center Standards, 1980) are vague on this point, simply noting that “whatever the approach” (Task Force on Assessment Center Guidelines, 1989, p. 465) to assessor training, the objective is obtaining accurate judgments. This emphasis on accuracy, especially considered in conjunction with Harris, Becker, and Smith’s (1993) comment that future AC research may benefit by considering work in the performance appraisal area, suggests a specific assessor training strategy that should be effective for improving the construct validity of AC ratings. Frame-of-reference (FOR) training is an intervention designed to improve the validity of trait judgments in the performance appraisal context. It is theoretically relevant to an investigation of the construct validity of AC ratings, as described below.

Overview of FOR Training

Initially proposed by Bernardin and Buckley (1981), the primary goal of FOR training was to eliminate idiosyncratic standards held by raters and replace them with a common frame of reference for rating. The overall intent of the training was to bring individual perceptions into closer congruence with those held more widely in the organization. The steps involved in such training include (a) teaching important dimensions comprising the job and behaviors indicative of each dimension, (b) discussing behaviors indicative of various effectiveness levels within each dimension, (c) providing practice evaluations with the new frame of reference, and (d) giving feedback on the accuracy of the ratings (Pulakos, 1984).

There is abundant and consistent evidence that FOR training increases the accuracy of performance appraisal ratings (e.g., Athey & McIntyre, 1987; Cardy & Keefe, 1994; Day & Sulsky, 1995; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986; Schleicher & Day, 1998; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992, 1994; Woehr, 1994). A meta-analysis evaluating the relationship between performance appraisal training and rating outcomes (e.g., halo, leniency, observational and rating accuracy) concluded that FOR training results in the largest overall increase in rating accuracy (Woehr & Huffcutt, 1994). The theoretical relevance of FOR training for improving the construct validity of AC ratings is highlighted in the following section by comparing explanations for the exercise effect with known outcomes and advantages of FOR training.

FOR Training for Assessors

Most explanations found in the literature for the exercise effect invoke the limited information-processing capacity of assessors. To make their ratings, assessors must observe and recall behaviors for each assesse in each exercise and then categorize these behaviors into relevant dimensions. This is a demanding task, and it may be unrealistic to expect assessors to do all of this accurately and reliably, given the basic limitations in human information processing (Bycio et al., 1987; Reilly et al., 1990). These limitations are often exacerbated by the lack of clarity in the dimension definitions. That is, a great deal of the exercise effect may be caused by scoring the same behavior for several abilities (i.e., dimensional dependency; Brannick, Michaels, & Baker, 1989) or interpreting the dimension differently from one exercise to another (Robertson, Gratton, & Sharpley, 1987). Reilly et al. observed that definitions of dimensions are typically written in general terms and are not clearly related to the operational definitions (i.e., behaviors elicited by particular exercises). It is then left to the already cognitively overburdened assessors to decide which behaviors are part of which dimensions.

Another explanation for the exercise effect involves the use of categorization schemas. Shore et al. (1990) noted that schemas greatly simplify the challenging task required of assessors. That is, the ability to organize performance information around relevant categories should make the rating task both easier and more accurate. However, the AC may provide such a strong organizational framework in terms of exercises that assessors are forced to code the behavior by exercise rather than by dimension (Silverman et al., 1986), resulting in the exercise effect. Strong manipulations may be necessary to modify these schemas and therefore allow the assessors to successfully reorganize the information they have obtained by dimension.

The previously discussed explanations for the exercise effect seem to converge in suggesting an intervention that would (a) reduce information-processing demands placed on assessors, (b) provide greater clarity to the dimension definitions, and (c) change categorization schemas to be more dimension oriented. These are all established outcomes of FOR training. First, the central role of dimensions in FOR training is readily apparent. Most relevant to the present study is the fact that a large part of FOR training involves a discussion of the dimensions and the behaviors indicative of each dimension. Obviously, such an emphasis should help assessors distinguish between dimensions and thereby help to reduce the exercise effect. Moreover, assessors also learn how to apply consistent standards across situations (i.e., exercises). This strategy is believed to reduce the cognitive demands placed on assessors. Furthermore, previous research on FOR training (Sulsky
& Day, 1992) has shown that such training teaches raters to form more accurate on-line impressions, reducing the need to recall behaviors for each assessees in each exercise and then to categorize these behaviors into relevant dimensions. With the cognitive demands reduced, FOR-trained assessors should provide more accurate and reliable dimension ratings. A final point relevant to the exercise effect is that previous research has shown that FOR training changes categorization schemas to more trait-based representational forms (Schleicher & Day, 1998). Therefore, FOR training should lead to assessors organizing candidate information around dimension categories, as opposed to exercise categories, thereby improving the construct validity of the assessment ratings.

FOR Training and the Construct Validation of ACs

As Landy (1986) noted, all validation efforts should be construed as part of the broader process of scientific theory development (i.e., traditional hypothesis testing)—a process of accumulating various types of judgmental and empirical evidence to support an inferential network of psychological constructs and their operational measures (Binning & Barrett, 1989). Accordingly, any information gathered in the process of developing or using a test is relevant to its validity in that it contributes to users' understanding of the underlying construct (Anastasi, 1986). In that spirit, the present study examines many different sources of evidence for the construct validity of ACs.

In the context of theory building for the construct validity of ACs, FOR training of assessors can be thought of as an intervention that impacts the strength of two inferences: (a) the link between the predictor measure and the predictor construct (i.e., what is traditionally thought of as construct validity) and (b) the link between the predictor measure and the criterion measure (i.e., what is traditionally thought of as criterion-related validity). It is believed that FOR training will improve the overall quality of the predictor measure (i.e., assessment ratings) and therefore the strength of both of these inferences. In other words, training serves as a moderator of both construct validity (e.g., convergent–discriminant) relationships and criterion-related relationships. The specific components involved in testing these two inferences are described next.

Inference 1: Predictor Measure → Predictor Construct

This inference pertains to the level of confidence a researcher has that a predictor measures a specific construct of interest. Such evidence usually takes the form of empirically based relationships and judgments that are both convergent and discriminant in nature (Campbell & Fiske, 1959; Cook & Campbell, 1979; Cronbach & Meehl, 1955); however, examining the reliability of a predictor measure (e.g., AC ratings) can also provide construct-oriented evidence regarding how well components (e.g., different assessors’ ratings) model the underlying predictor construct (Anastasi, 1986; Binning & Barrett, 1989). Therefore, we examined the impact of FOR training on improving the strength of this inference with regard to both reliability and convergent and discriminant relationships. We hypothesized that, because the goal of FOR training is to teach all assessors to view performance with a common frame of reference, the interrater reliability among FOR-trained assessors will be higher than the interrater reliability among control-trained assessors (Hypothesis 1).

As noted above, the strength of Inference 1 is also shown by patterns of convergent and discriminant correlations between the predictor measure and other measure of similar and different constructs. Although FOR training is generally expected to improve both the convergence and the discrimination of assessment ratings, this does not necessarily translate into straightforward hypotheses with regard to the pattern of mono- or hetero-trait and of method correlations. That is, FOR training is not expected to increase the monotrait–heteromethod correlations as an index of convergent validity. As Sackett (1987) noted, “Ratings of dimensions across exercises aren’t intended as merely repeated measures that should correlate perfectly well were it not for measurement error; rather, they are intended to tap only partially overlapping subsets of the behavioral domain that the dimension label represents” (p. 19). Therefore, FOR training, by improving the quality of assessment ratings, would not necessarily increase these monotrait–heteromethod correlations. Moreover, this type of convergent validity has often been found in previous research, even with untrained assessors (e.g., Neidig et al., 1979; Sackett & Dreher, 1982; Turnage & Muchinsky, 1982). In fact, it is possible that these correlations might be higher for control raters, given that they would be more likely to use broader categories (i.e., more general trait or person factors) for rating, which would contribute to the illusion of convergent validity. Therefore, nothing is hypothesized with regard to the size of monotrait–heteromethod correlations. What FOR training is expected to do, however, is to improve discriminant validity by clarifying the differences between assessment dimensions. We therefore expected to see lower correlations within exercises across different dimensions (heterotrait–monomethod correlations; Hypothesis 2a) and lower correlations between different dimensions in different exercises (heterotrait–heteromethod correlations; Hypothesis 2b) for FOR assessors than for control assessors.

What has been described above represents only one way of assessing the convergent and discriminant validity of AC ratings, however. Several researchers have noted that examining relationships across and within exercises may be lacking as the sole strategy for judging claims of the construct validity of ACs. Perhaps a more appropriate standard would involve other methods of assessing similar constructs (e.g., Fleenor, 1996; Riggio, Aguirre, Mayes, Belloli, & Kubiak, 1997; Shore et al., 1990). Reilly et al. (1990) noted that the convergent and discriminant validity of assessment ratings should be judged on a multitrait–multimethod matrix that includes total scores on the AC dimensions correlated with other measures of those dimensions obtained outside of the AC exercises (e.g., personality and cognitive ability measures). These prescriptions suggest more of a nomological network (Cronbach & Meehl, 1955) approach to establishing the construct validity of assessment ratings. The present study adopts this more inclusive view of convergent and discriminant validity and examines relationships between dimension ratings and other personality and skills measures that assess both similar and different constructs.

The personality and skills measures that we examined are those that were collected at the same time the assessees participated in the AC and include the need for dominance (Steers & Braunstein, 1976), self-monitoring (Snyder & Gangestad, 1986), and social skills (Riggio, 1989). We expected these measures to converge with ratings of the three dimensions assessed in the present AC (leadership, communication skills, and decision-making skills).
Specifically, self-monitoring, social skills, and the need for dominance should converge with the leadership dimension. High self-monitors tend to have generally outgoing and extraverted social styles that lead them to initiate conversations in interpersonal situations (Ickes & Barnes, 1977), as would those individuals high in social skills. This style would also lend itself to proportionally more contributions to group discussions, which have been shown to be predictive of leader emergence in groups (Stein & Heller, 1979). The need for dominance assesses individuals’ attempts to control their environments and to influence others and the extent to which they enjoy the role of leader (Steers & Braunstein, 1976); such factors should obviously converge with AC leadership ratings. Self-monitoring and social skills should also converge with the communication skills dimension, given the importance of such skills in effective interpersonal communication. Therefore, because of these expected convergent relationships between personality and skills measures and the assessment dimensions, we expected that the dimension ratings of FOR-trained assessors, compared with the control-trained assessors, would demonstrate higher correlations with similar personality measures (Hypotheses 3a, 3b, 3c, 4a, and 4b).

An additional way to examine the convergent validity of the assessment ratings is to compare them with ratings on the same dimensions made by others at a different time in a different context. In the present study, this second source of dimension ratings is provided by the assesses’ supervisors, and we hypothesized that the ratings of FOR-trained assessors, as compared with those of control-trained assessors, would converge better with the dimensional ratings provided by the supervisors (Hypothesis 5). Accuracy is the final criterion used to evaluate the effectiveness of FOR training at strengthening Inference 1. We hypothesized, on the basis of extensive research on FOR training, that FOR-trained assessors would provide more accurate ratings than control-trained assessors (Hypothesis 6).

**Inference 2: Predictor Measure → Criterion Measure**

FOR training is believed also to play a moderating role in the predictor measure–criterion measure link, thereby affecting the criterion-related validity of AC ratings. In general, we expected that the ratings produced by FOR-trained raters—because of their better reliability, accuracy, and convergent–discriminant validity—will show stronger relationships with job performance than will those produced by control-trained raters (Hypothesis 7).

**Method**

**Participants**

**Assessors.** Approximately 500 former business students who participated in an undergraduate AC were recruited. All assesses completed personality and skills measures in the AC, which was conducted approximately 4 years prior to the collection of the criterion data. All AC performance was videotaped.

Assesses were contacted via mailed surveys to collect the criterion information. If assesses failed to return the surveys, a follow-up phone call was made. Seventy-five respondents returned a survey (20% of total valid surveys sent; 105 surveys were returned from the U.S. Postal Service as undeliverable); there was complete data for 58 of the respondents. The breakdown of the sample was 54% male, 46% female; 50% White, 25% Asian, 13% Hispanic, 9% Filipino, 2% American Indian, and 2% Middle Eastern; mean age was 29.0 years (SD = 7.7); mean tenure was 32.0 months (SD = 26.3); 41% worked in the business sector (e.g., accounting, banking, investments, real estate), 25% worked in the service industry, 9% were in the manufacturing industry, and another 9% were in the transportation, communication, or public utilities sectors.

Although the number of participants is consistent with typical validation sample sizes in industry (Schmidt, Hunter, & Ury, 1976), given the low but not atypical survey return rate (Edwards, Thomas, Rosnfeld, & Booth-Kewley, 1997), potential response biases were explored. Differences were examined between respondents and nonrespondents on all feasible demographic and personality variables. There were no differences in terms of age, gender, ethnic group membership, or any of the personality variables (all ps > .05). Taken together, these comparisons suggest little reason to be concerned that response bias was a significant factor in the present study.

**Assessors.** Students at a different state university than the assesses (n = 122) served as assessors and viewed the videotaped AC performance of the 75 assesses who had returned the survey. Sixty-seven (55%) of these participants were male, and 98% of them were psychology students (with 3 master of business administration students). The mean age of the assessor participants was 22.4 years (SD = 0.5). Twenty-nine of the participants had managerial experience, 37 had conducted performance appraisals as part of their jobs, and 6 had participated in an AC. These assessors participated in the present study in groups of 4 to 6. Each group (22 total) was randomly assigned to a training condition (FOR or control; n = 11 groups each) and observed and rated the performance of 5 to 6 assesses.

**AC Exercises**

Three commonly used AC exercises were included in the present study: (a) leaderless group discussion (LGD), (b) presentation exercise, and (c) mock hiring interview. Each exercise was designed to provide an opportunity to rate three dimensions: (a) decision-making skills (including problem solving, analytical and decision-making ability, and creativity), (b) communication skills (including clarity of communication, expressiveness, and language and listening skills), and (c) leadership (including persuasiveness, influence on others, and human relations abilities). For control purposes and to keep the rating task manageable, only the first 5 min of both the LGD and the interview were observed and rated by the assessors.

**Development of Rating Scales and Comparison Scores**

Seven-point behaviorally anchored rating scales (BARS) were used to rate performance in each exercise on each of the three dimensions. Similar to the approach taken by Reilly et al. (1990), the first step in the development of these scales was to identify behaviors relevant to the three dimensions that are elicited by specific exercises. A group of five undergraduate research assistants trained on the meaning of the three dimensions viewed the videotapes of all 500 assesses and recorded critical incident behaviors that were indicative of exceptionally good or poor performance. Procedures outlined by Smith and Kendall (1963) were then followed for selecting behavioral anchors for the BARS. Specifically, the critical incidents were retranslated into the three dimensions by the five research assistants and the primary investigator, Deidra I. Schleicher. A criterion of 83% agreement (five out of six) was used to select behaviors for each dimension. The research team then assigned effectiveness ratings to these behaviors. On the basis of the means and the standard deviations of these effectiveness ratings, behavioral anchors for the BARS were selected. Each behavioral anchor had three components representing a behavioral example from each of the three exercises (such a rating format is considered a hybrid “task-BARS” approach; Harvey, 1991).

Comparison scores were developed for each of the assesses from the ratings of the research team. These comparison scores were used in computing the accuracy components. The expert assessors received extensive FOR training prior to independently rating each asseree (as recommended by Sulsky & Balzer, 1988); initial consensus was estimated at .86
(intraclass correlation). Experts then discussed their ratings to arrive at consensus comparison scores for each assessees on each dimension in every exercise.

### Procedure

Assessors were randomly assigned in groups of 4 to 6 to receive control or FOR training. They were given basic knowledge of the design and uses of an AC and then received either control or FOR training (90 min) before rating the target assessees (90 min).

**FOR training.** Both the FOR and control training conditions included those practices common to assessor training programs in most ACs (Spychalski et al., 1997), with the FOR training including additional components. FOR assessors received a description of the three exercises in the AC, including a videotaped demonstration. Copies of the dimension definitions and BARS were distributed, and the trainer read aloud the dimension definitions and scale anchors, including the different ways in which the dimension could be manifested across exercises. Behaviors pertaining to different effectiveness levels for each dimension were also discussed by the trainer. The overall goal of this training was to create a common performance theory (i.e., frame of reference) among assessors so that they would agree on the standards used to evaluate assesses’ behaviors.

Assessors then practiced rating 4 assesses with their new frame of reference. The assessors were instructed to evaluate each assesse after each exercise on each performance dimension. After each exercise, assessors were asked to share their ratings with the rest of the group; the trainer led a discussion about the assigned ratings, clarifying discrepancies among raters and providing feedback on the appropriate effectiveness level on each dimension portrayed by each assesse. After training, assessors viewed the videotape of the 5 or 6 target assessees and, after each exercise, independently provided ratings on each dimension by using the BARS.

**Control training.** The introduction to the design and uses of ACs was lengthier in the control condition than in the FOR condition to equate overall training session time between the two conditions. The assessors in the control training condition also received a complete description of the three exercises in the AC, saw the videotaped demonstration of the exercises, and were given the dimension definitions and BARS to examine. Control assessors also viewed and rated the practice assesses, but their ratings were not discussed and no feedback was provided. Instead, the demonstration of the exercises and the practice ratings were used to emphasize the importance of accurately observing and recalling behavior. After viewing the practice tapes, the control assessors watched the target videotape of the 5 or 6 assessees and provided ratings for each assesse on each dimension after each exercise. We should note that each assesse’s performance was rated by both a FOR-trained group and a control group.

### Criterion Variables

For each assesse, there were within-exercise dimension ratings, across-exercise dimension ratings, and a mechanically derived overall assessment rating (OAR; a unit-weighted composite of the across-exercise dimension ratings). In addition to examining relationships within and between exercises and dimensions (i.e., convergent and discriminant relationships based on within-exercise dimension ratings), other relationships included those between the ratings and the comparison scores (i.e., accuracy relationships based on across-exercise dimension ratings), those between the ratings and other personality and ability measures (based on across-exercise dimension ratings), and those with job performance criterion variables in a predictive design (based on the OAR).

#### Accuracy

Five indexes of accuracy were computed. Cronbach’s (1955) four component indexes include (a) elevation (the differential grand mean), (b) differential elevation (the differential main effect of rates), (c) stereotype accuracy (the differential main effect of dimensions), and (d) differential accuracy (the differential Ratee × Dimension Interaction). Borman’s (1977) differential accuracy is the average of the z-transformed correlations between a rater’s ratings for each dimension and the corresponding true scores across rates. Sulsky and Balzer (1988) provide a more complete description of these accuracy indexes.

#### Reliability

The interrater reliability for each assesse on each dimension was computed for all groups of assessors. In accordance with Schmitt (1977) and Schneider and Schmitt (1992), assessors in each group were treated as items (and assessees were treated as cases), enabling the calculation of internal consistency (i.e., coefficient α) estimates for each dimension. Confidence intervals were constructed around these reliability estimates (McGraw & Wong, 1996), allowing for comparison between FOR and control groups’ interrater reliability. Because longer scales (i.e., more assessors) lead to greater internal consistency when all other things are equal, some assessors in each group were randomly excluded from the reliability analyses, resulting in equal numbers of FOR and control-trained assessors for each videotape.

#### Personality and skills.

The Self-Monitoring Scale (Snyder & Gangestad, 1986) is an 18-item true–false instrument designed to assess the extent to which people “monitor” (observe, regulate, and control) the public appearances of self they display in social situations and interpersonal relationships” (Snyder, 1987, pp. 4–5). Although the Self-Monitoring Scale has its critics on issues such as dimensionality (e.g., Briggs & Cheek, 1988), a recent study by Gangestad and Snyder (2000) showed that self-monitoring is a unitary construct with regard to its relations with such external criteria as expressive control, nonverbal decoding skills, and behavioral variability. In conjunction with research demonstrating that high self-monitors are more likely to emerge as leaders (Day, Schleicher, Unckless, & Hiller, 2002) and have stronger interpersonal and communication skills (e.g., Sypher & Sypher, 1983), these relationships illustrate the relevance for examining expected convergent relationships. Coefficient alpha for the Self-Monitoring Scale was .71 in the present study.

The Social Skills Inventory (Riggio, 1989) consists of 90 items rated on a 5-point Likert-type scale. A high score on this scale indicates a high degree of competence in social situations. The Social Skills Inventory demonstrated convergent and discriminant validity in a series of studies reported by Riggio (1986). Specifically (and relevant to its use in the present study), the Social Skills Inventory correlated in predicted directions with the Affective Communication Test (Friedman, Prince, Riggio, & DiMatteo, 1980) and the Profile of Nonverbal Sensitivity (Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979). Reliability (Cronbach’s α) for this scale was .88.

Finally, the need for dominance—referring to an individual’s attempts to control the environment, influence others, and assume leadership roles—was assessed via the Manifest Needs Questionnaire (Steers & Braunstein, 1976). This scale consists of 20 items (5 items per subscale; only the Need for Dominance subscale was included in the present hypotheses) with a 7-point Likert-type response format. The reliability estimate for the dominance subscale was .71 (coefficient α).

#### Supervisor ratings.

The survey mailed to assesses included several measures that were to be given to the assesses’ immediate supervisors or managers to be filled out and mailed directly back to the researcher Deidra J. Schleicher. Three of these measures were the same 7-point BARS used by assessors, with the behavioral anchors modified so that they were tied less to the AC setting. In addition, the supervisors also completed two other measures of performance: an overall evaluation of the employee and an assessment of the employee’s future potential (both scaled from 1 to 7). These five ratings were combined to create an overall index of job performance (α = .89).

### Results and Discussion

Intercorrelations and descriptive statistics for the assesssee variables are reported in Table 1. Table 2 reports intercorrelations and descriptive statistics for the assessor variables. These same intercorrelations are reported separately for FOR and control assessors in Table 3.
Interrater Reliability

The first six hypotheses examined the effect of FOR training on the quality of assessors’ ratings. Hypothesis 1 stated that the interrater reliability would be higher for the FOR-trained assessors than for the control-trained assessors. Internal consistency (i.e., coefficient α) estimates for each dimension were as follows: communication skills, .83 (FOR) and .76 (control); decision-making skills, .89 (FOR) and .69 (control); and leadership skills, .75 (FOR) and .65 (control). Across the three dimensions, the average reliability for the FOR ratings was .82 (coefficient α), compared with .70 for the control ratings. These results reveal that the AC ratings overall had quite high levels of interrater reliability (consistent with previous research; e.g., Klimoski & Brickner, 1987) and that the FOR ratings were consistently more reliable than the control ratings. Moreover, for all four comparisons, the confidence intervals around the reliability estimates were nonoverlapping (McGraw & Wong, 1996), indicating that the FOR ratings were significantly more reliable than the control ratings. Thus, Hypothesis 1 was supported. Given that the goal of FOR training is to teach all of the assessors to view performance with a common frame of reference, such a finding is not surprising, although it is an important one. However, as always, reliability information is most appropriately viewed as a necessary, but insufficient, condition for claims of construct validity.

Mono- and Hetero- Trait and Method Correlations

Hypotheses 2a and 2b predicted that the heterotrait–monomethod correlations and the heterotrait–heteromethod correlations would be lower for FOR assessors than for control assessors, respectively. Correlations between ratings on the same and different exercises and dimensions are reported in Table 4, for both the FOR and control assessors. To test the hypotheses, z transformations were performed on these individual correlations, and the resulting coefficients were averaged to obtain measures of convergent and discriminant validity. Although not part of the hypotheses, the average convergent validity coefficient (i.e., monotrait–heteromethod correlation) was .34 for FOR assessors and .48 for control assessors (both ps ≤ .05). A t test for significant differences between nonindependent correlations (Steiger, 1980) revealed that this was a significant difference, t(16) = −2.14, p ≤ .05 (two-tailed; R² = .246); the monotrait–heteromethod correlations were higher for control assessors than for FOR assessors. The

Table 1
Intercorrelations, Means, and Standard Deviations for the Assessee Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Gender</td>
<td>1.5</td>
<td>0.5</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Age</td>
<td>29.0</td>
<td>7.7</td>
<td>.29*</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Tenure</td>
<td>32.0</td>
<td>26.3</td>
<td>.05</td>
<td>.69**</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Performance</td>
<td>5.7</td>
<td>0.9</td>
<td>.08</td>
<td>—</td>
<td>.11</td>
<td>.05</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Self-monitoring</td>
<td>9.1</td>
<td>3.4</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Social skills</td>
<td>278.0</td>
<td>28.1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Need for dominance</td>
<td>21.6</td>
<td>4.4</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. FOR rating</td>
<td>4.4</td>
<td>1.0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Control rating</td>
<td>4.0</td>
<td>1.3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Ns range from 40 to 75. *p ≤ .05. **p ≤ .01.

Table 2
Intercorrelations, Means, and Standard Deviations for the Assessor Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Gender</td>
<td>1.5</td>
<td>0.5</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Age</td>
<td>22.4</td>
<td>4.5</td>
<td>.05</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Tenure</td>
<td>2.0</td>
<td>0.2</td>
<td>.04</td>
<td>—</td>
<td>—</td>
<td>.19*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Managerial experience</td>
<td>1.8</td>
<td>0.4</td>
<td>.21*</td>
<td>.37**</td>
<td>16</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Performance appraisal</td>
<td>1.7</td>
<td>0.5</td>
<td>.22*</td>
<td>.31**</td>
<td>13</td>
<td>.43**</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Assessment center</td>
<td>2.0</td>
<td>0.2</td>
<td>.06</td>
<td>—</td>
<td>.19*</td>
<td>.21*</td>
<td>.23**</td>
<td>10</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Elevation</td>
<td>0.7</td>
<td>0.9</td>
<td>—</td>
<td>−.13</td>
<td>—</td>
<td>−.11</td>
<td>−.08</td>
<td>.05</td>
<td>.03</td>
<td>.00</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Differential elevation</td>
<td>0.6</td>
<td>0.2</td>
<td>−.13</td>
<td>.05</td>
<td>−.01</td>
<td>−.07</td>
<td>−.15</td>
<td>.16</td>
<td>.02</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Differential accuracy</td>
<td>0.4</td>
<td>0.1</td>
<td>.05</td>
<td>.02</td>
<td>−.15</td>
<td>.02</td>
<td>.01</td>
<td>.01</td>
<td>.18*</td>
<td>−.04</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Stereotype accuracy</td>
<td>0.3</td>
<td>0.2</td>
<td>−.03</td>
<td>.03</td>
<td>.02</td>
<td>−.09</td>
<td>−.12</td>
<td>−.18*</td>
<td>10</td>
<td>−.03</td>
<td>.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Borman’s differential</td>
<td>1.0</td>
<td>0.5</td>
<td>.13</td>
<td>−.09</td>
<td>.22*</td>
<td>.09</td>
<td>.02</td>
<td>−.10</td>
<td>−.09</td>
<td>−.28**</td>
<td>−.37**</td>
<td>−.18*</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. N = 122. For managerial, performance appraisal, and assessment center experience, 1 = yes and 2 = no.
*p ≤ .05. **p ≤ .01.
average discriminant validity coefficient (i.e., heterotrait–monomethod correlations) was .66 for FOR assessors and .74 for control assessors. As we predicted (Hypothesis 2a), the heterotrait–monomethod correlations were significantly smaller for FOR assessors than for control assessors. This was a significant difference, t(16) = −2.54, p ≤ .05 (one-tailed; $R^2 = .338$). The heterotrait–heteromethod correlations were .28 and .48 for FOR and control assessors, respectively. This was a significant difference, t(34) = −5.62, p ≤ .01 (one-tailed; $R^2 = .497$), and thus, Hypothesis 2b was also supported.

The support for Hypotheses 2a and 2b, in conjunction with the magnitude of the monotrait–heteromethod correlations, suggests an intriguing conclusion. For all three types of correlations, control raters had larger correlations than FOR raters. This finding suggests that control assessors did not discriminate between dimensions as effectively as FOR assessors. This lack of discrimination is dramatically illustrated by the sizable magnitude of the heterotrait–heteromethod correlation for control assessors (.48), which represents the average correlation between ratings of different traits in different exercises. Although it may not be realistic to expect FOR training to increase the monotrait–heteromethod correlations, it does appear that FOR training can improve assessors’ discrimination between dimensions, thereby enhancing the discriminant validity component of construct validity. Furthermore, it is this evidence of discriminant validity that has proven most elusive to past AC researchers.

**Convergence With Personality Measures and Supervisors’ Dimension Ratings**

We hypothesized (Hypothesis 3) that, compared with control assessors, FOR-trained assessors would show stronger correlations between their leadership ratings and (a) self-monitoring, (b) social skills, and (c) need for dominance. These correlations are shown in

---

### Table 4

**Mono- and Hetero-Trait and Method Correlations for Frame-of-Reference (FOR) and Control Assessors—Hypothesis 2**

<table>
<thead>
<tr>
<th>Dimensions within exercises</th>
<th>Leaderless group discussion</th>
<th>Presentation</th>
<th>Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>D</td>
<td>L</td>
<td>C</td>
</tr>
<tr>
<td><strong>Leaderless group discussion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Presentation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interview</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** FOR raters are below the diagonal, and control raters are above the diagonal. Individual correlations are based on an N = 122. C = Communication skills; HM = heterotrait–monomethod correlations; MH = monotrait–heteromethod correlations; HH = heterotrait–heteromethod correlations; D = Decision-making skills; L = Leadership skills.
Table 5. Results from t tests for significant differences between two nonindependent correlations support Hypothesis 3b (social skills) and Hypothesis 3c (need for dominance): the test for Hypothesis 3b (social skills) was not statistically significant (r = .28 for FOR ratings and r = .14 for control ratings), t(55) = 2.38, p < .05 (one-tailed; R² = .097). Results for Hypothesis 3c (need for dominance) were in the hypothesized direction, but these were not statistically significant (p > .05, one-tailed; R²s = .007 and .009, respectively).

We also hypothesized (Hypothesis 4) that, compared with control assessors, there would be stronger correlations between the communication skills ratings provided by FOR raters and (a) self-monitoring and (b) social skills personality variables. These results are presented in Table 5. Results for Hypothesis 4a (self-monitoring) and Hypothesis 4b (social skills) were in the hypothesized direction but were not statistically significant (self-monitoring: r = .32 for FOR ratings and r = .27 for control ratings), p > .05, one-tailed, R² = .013 (social skills: r = .22 for FOR ratings and r = .13 for control ratings), t(55) = 1.34, p = .08 (one-tailed; R² = .033). In sum, support for the personality hypotheses was somewhat weak. In all five of the hypothesized relationships for the personality variables, the pattern of the FOR and control correlations was in the predicted direction, but only one of these differences was significant.

Perhaps the most relevant convergent measures available in the present study are the supervisors’ ratings of leadership, communication skills, and decision-making skills. These measures are essentially an assessment of the same dimensions as were evaluated in the AC, by different people, at a different time, and in a different setting. As such, these measures are also seen as convergent measures in a construct validity framework. An examination of these relationships did indeed reveal convergence. Specifically, Hypothesis 5 predicted that the dimension ratings for FOR-trained assessors, as compared with those for control-trained assessors, would be more highly correlated with the corresponding dimension ratings as provided by the supervisor. As Table 5 shows, this hypothesis was supported for two of the three dimensions. The correlations between assessment dimension and corresponding supervisor dimension ratings were .12 (FOR) versus .00 (control) for communication skills, .26 (FOR) versus .28 (control) for decision-making skills, and .42 (FOR) versus .27 (control) for leadership skills. For communication and leadership skills, the FOR correlation was significantly (p < .05) higher than the control correlation (see Table 5 for statistical test results). Given the pronounced differences across these two measurement episodes, these findings indicate relatively high levels of convergence (with the exception of the communication skills dimension) and suggest that the failure to obtain significant results with the personality measures may be more a function of the personality variables available than it is evidence for the lack of convergent validity of the AC ratings (particularly the FOR ratings).

### Table 5

**Convergent Correlations Between Assessment Center (AC) Ratings and Personality Variables and Supervisor Dimension Ratings—Hypotheses 3–5**

<table>
<thead>
<tr>
<th>Leadership AC ratings</th>
<th>FOR</th>
<th>Control</th>
<th>t*</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-monitoring</td>
<td>.36*</td>
<td>.32*</td>
<td>0.62</td>
<td>.007</td>
</tr>
<tr>
<td>Social skills</td>
<td>.28*</td>
<td>.14</td>
<td>2.38*</td>
<td>.097</td>
</tr>
<tr>
<td>Need for dominance</td>
<td>.44*</td>
<td>.40**</td>
<td>0.70</td>
<td>.009</td>
</tr>
<tr>
<td>Supervisor leadership</td>
<td>.42**</td>
<td>.27</td>
<td>2.15*</td>
<td>.117</td>
</tr>
<tr>
<td>Communication skills AC ratings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-monitoring</td>
<td>.32*</td>
<td>.27</td>
<td>0.84</td>
<td>.013</td>
</tr>
<tr>
<td>Social skills</td>
<td>.22</td>
<td>.13</td>
<td>1.34</td>
<td>.033</td>
</tr>
<tr>
<td>Supervisor</td>
<td>.12</td>
<td>.00</td>
<td>1.80*</td>
<td>.085</td>
</tr>
</tbody>
</table>

Note. N = 40–58. FOR = frame of reference.

* p ≤ .05. ** p ≤ .01.

| Decision-making skills AC ratings | Supervisor decision-making skills ratings | .26  | .28 | −0.28  | .002 |

Accuracy

The final hypothesis bearing on the strength of Inference 1 predicted that FOR-trained assessors would make more accurate ratings than would control assessors (Hypothesis 6). Given the conceptual overlap of the five accuracy indexes and their statistically significant intercorrelations (see Table 2), this hypothesis was tested via multivariate analysis of variance, with training (FOR vs. control) as the independent variable and the five accuracy indexes as the multiple dependent variables. As we predicted, ratings provided by FOR-trained assessors were significantly more accurate than those made by control assessors, F(5, 94) = 8.10, p < .001, Wilks’s Λ = .70; partial η² = .242. Follow-up discriminant function analyses revealed that this effect was driven primarily by elevation, differential elevation, and differential accuracy, with standardized discriminant function coefficients of −.73, −.75, and −.45, respectively. As a means of estimating the effect size associated with each accuracy dependent variable, we also computed univariate analyses of variance (ANOVAs). Table 6 reports the results of these ANOVAs along with the accuracy means for the FOR and control groups. Not surprisingly, results from these individual ANOVAs parallel those for the follow-up discriminant function analyses. The current results regarding accuracy are in line with robust findings showing that FOR training is the best approach for improving rating accuracy (see Woehr & Huffcutt, 1994) and extend the effectiveness of FOR training from the performance appraisal setting to the AC setting.

Unlike typical performance appraisal situations in which political motivations often color-the ratings (Longenecker, Sims, & Gioia, 1987), there are unlikely to be political motivations involved in ACs, given that assessors typically neither know the assesses (Spychalski et al., 1997) nor will be accountable to them for their ratings. Therefore, one must assume that any exercise effects originate from an inability to distinguish between the dimensions rather than any motivational reasons on the part of the assessors, making accuracy an important criterion in ACs.

### Criterial Relationships

No previous studies that have examined variables thought to improve the convergent and discriminant validity of assessment ratings have also assessed to what extent criterion-related validities are similarly affected (Klimoski & Brickner, 1987). This is likely due to the difficulty of having the degree of control required to manipulate factors believed to affect construct validity and yet having real-world outcomes available as criteria. The present study
Table 6  
Analysis of Variance Results for the Five Accuracy Components—Hypothesis 6

<table>
<thead>
<tr>
<th></th>
<th>FOR</th>
<th>Control</th>
<th>F</th>
<th>p</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td>0.36</td>
<td>1.02</td>
<td>17.31</td>
<td>&lt;.001</td>
<td>.14</td>
</tr>
<tr>
<td>Differential elevation</td>
<td>0.53</td>
<td>0.67</td>
<td>11.55</td>
<td>&lt;.001</td>
<td>.09</td>
</tr>
<tr>
<td>Differential accuracy</td>
<td>0.42</td>
<td>0.45</td>
<td>3.81</td>
<td>&lt;.05</td>
<td>.03</td>
</tr>
<tr>
<td>Stereotype accuracy</td>
<td>0.27</td>
<td>0.30</td>
<td>0.40</td>
<td>&gt;.05</td>
<td>.01</td>
</tr>
<tr>
<td>Borman’s differential accuracy</td>
<td>1.03</td>
<td>0.91</td>
<td>0.01</td>
<td>&gt;.05</td>
<td>.00</td>
</tr>
</tbody>
</table>

Note. For elevation, differential elevation, differential accuracy, and stereotype accuracy, smaller numbers represent greater accuracy. For Borman’s differential accuracy, larger numbers represent greater accuracy. FOR = frame of reference.

provided a unique opportunity for examining both of these issues. Hypothesis 7 predicted that there would be greater criterion-related validity with the FOR ratings than with the control ratings. The criterion of interest was supervisory job performance ratings. The criterion-related validity coefficient for the FOR ratings was .32 (p < .05), compared with .21 (p > .05) for the control ratings. These two correlations were significantly different, t(40) = 1.95, p < .05 (one-tailed; R² = .091), indicating that the FOR ratings were significantly more predictive of performance than the control ratings, thus supporting Hypothesis 7. Supplemental analyses were undertaken to examine whether the AC ratings provided incremental validity beyond college grade point average and American Assembly of Collegiate Schools of Business scores (a standardized exam assessing knowledge of core business areas). The results indicated that whereas FOR ratings accounted for a significant (p < .05) amount of unique variance in performance over these other predictors, the control ratings did not (complete results for these analyses are reported in Table 7).

We noted above that the improved criterion-related validity of FOR-generated ratings was likely due to their better construct validity (e.g., their superior reliability, accuracy, and convergent–discriminant validity). Unfortunately, there is no way to directly test the hypothesized mediating role of these variables in the present study. However, indirect evidence for the importance of accuracy comes from examining the criterion-related validity associated with the comparison score ratings. The comparison scores were slightly more predictive of the job performance ratings than were the FOR ratings (.36 vs. .32, respectively), implying that it may be their similarity to comparison scores (i.e., accuracy) that was responsible for the improved criterion-related validity of the FOR ratings. Interestingly, two separate aspects of the results suggest that the effectiveness of FOR training may lie in more finely tuned calibration of the assessors on the dimensions rather than in merely teaching the correct rank order of assessees on dimensions. First, Borman’s (1977) differential accuracy index, which provides only correlational information, was not improved as a result of FOR training but the Cronbach accuracy components, which provide distance information, were. Second was the high correlation (r = .93) we found between the FOR and control ratings. Although this finding may initially lead one to question the significance of a difference between FOR and control ratings, it should be noted that this correlation represents only rank-order association. The practical significance of both of these findings is that assessors may reach similar conclusions on rank order of candidates regardless of training. However, the other differences between FOR and control ratings (e.g., mean differences in distance accuracy and convergent and discriminant relations) become extremely important when ACs are used for training and development purposes.

The criterion-related validity coefficients obtained in the present study, particularly the FOR-based coefficients, approach the magnitude of the meta-analytic coefficients found in the literature (.37, corrected [Gaugler et al., 1987]; .43 [Hunter & Hunter, 1984]; and .41 [Schmitt et al., 1984]). As such, the present study provides further support for the predictive validity of ACs. These findings are particularly good news for this specific AC, given that there were many factors in the present study (e.g., data were collected across multiple organizations, no direct or indirect criterion contamination; Klimoski & Brickner, 1987) that should, if anything, lead to a conservative estimate of the criterion-related validity.

Conclusion and Future Directions

In summary, the present study examined many different types of evidence to demonstrate the effectiveness of FOR training in improving the construct and criterion-related validity of ACs. Results showed that FOR training, as compared with control training, improved the reliability and the accuracy of AC ratings. There was also improved discriminant validity associated with the FOR assessment ratings in the form of smaller heterotrait–monomethod and heterotrait–heteromethod correlations and somewhat improved convergent validity in the form of larger correlations with external measures of the same and similar constructs. Finally, FOR training significantly improved the criterion-related validity of the current AC for predicting supervisors’ ratings of job performance. Implications for both FOR training and AC practice are discussed below, after a review of some of the possible limitations associated with the current study.

Limitations of the Present Study

The two primary methodological goals in the present study were (a) to ensure the ability to make strong internal validity inferences regarding the effect of FOR training and (b) to eliminate most, if not all, potential contaminants (i.e., confounds) that typically exist in AC validity studies (e.g., direct and indirect criterion contamination; Klimoski & Brickner, 1987). In light of these goals, we address specific potential generalizability concerns. First, because of the low response rate from the survey, potential response biases are of concern. Although there were no differences between re-
spondents and nonrespondents in terms of age, gender, ethnicity, or any of the personality variables, this is a limited set of variables on which possible differences were assessed. Accordingly, there may be representativeness issues with our sample; this is a possible limitation that should be addressed in future studies of this type. Second, two of the AC exercises were truncated to 5 min. Although the impression formation process literature suggests that assessors’ ratings of performance are unlikely to change after the first 5 min (e.g., Ambady & Rosenthal, 1992), this decision presents possible generalizability concerns. It also suggests that it may be interesting for future research to examine whether FOR training mitigates this tendency to form impressions so early on and to not revise them in light of additional behavior. Finally, there may be concerns arising from the use of students as both assessees and assessors. With regard to the former, for a variety of reasons (e.g., mean age, business student status, previous work experience, industries entered), the current assessees may be precisely the candidates likely to be involved in most ACs used in industry (with the exception of those designed for executive-level assessment). The issue of student assessors is admittedly more challenging. However, it is not clear who should have been chosen as “assessor” subjects (i.e., there are very few who possess the job title “assessor”). Instead, special effort was made to recruit subjects as assessors who had managerial experience so as to improve this aspect of generalizability.

Implications and Future Directions for FOR Training

Previous research on FOR training offers several possible explanations for its effectiveness in the AC context in the present study, including a reduction in cognitive load. For example, Day and Sulsky (1995) have suggested that FOR training helps raters form more accurate on-line impressions, reducing the need to recall behaviors for each assessee in each exercise and then categorize these behaviors into relevant dimensions. This FOR effect would likely be even larger if a greater number of performance dimensions were used than those in the present study. That is, control raters have a better chance of keeping 3 dimensions straight rather than 8 or 10. As the number of assessment dimensions increases, so might the advantages associated with FOR training.

It has also been noted that the AC may provide such a strong, exercise-based organizational framework that assessors are forced to code the behavior by exercise rather than by dimension. Strong manipulations may be necessary to modify these schemata to allow the assessors to successfully reorganize information by dimension (Harris et al., 1993; Silverman et al., 1986). FOR training represents one such manipulation. Research has shown that FOR training changes the impression formation process of people to more trait-based (as opposed to behavior-based) representations (Schleicher & Day, 1998). We assumed that this FOR training feature would reduce exercise effects in the present study. Results suggest that FOR training did have some impact in this regard. Moreover, supplemental exploratory analyses on the present data revealed that impressions formed by FOR-trained assessors (coded from a free-recall task) were significantly more target referent and significantly more abstract than those formed by control-trained assessors (Carlston, 1994). In other words, FOR assessors’ impressions primarily used categories and traits (i.e., assessment dimensions)—as opposed to behaviors—as the main representational forms. Future research should further pursue this proposed mediating role of the impression formation process of assessors.

Implications and Future Directions for ACs

The present results regarding the effectiveness of FOR training also have implications for AC practice. They suggest that FOR training should be useful in AC practice given that it is no more lengthy or expensive than control training, enhances the legal defensibility of the AC (Werner & Bolino, 1997), and appears to enhance AC validity. That is, previous research has established FOR training as effective in improving rating accuracy (e.g., Woehr & Huffcutt, 1994), its relationship to components of procedural justice has been noted (see Schleicher & Day, 1998), and the preliminary evidence presented here suggests that it improves both construct and criterion-related validity of AC ratings.

FOR training represents a fundamentally different approach to training assessors than what has typically been used. For example, many AC training programs focus on the accuracy of the observation and recall of behavior (see Byham, 1977; Spychalski et al., 1997). FOR training, in contrast, focuses on the accuracy of impressions formed on the various performance dimensions. In fact, FOR training often leads to less accurate recall of behavior while it simultaneously enhances the accuracy of impressions (Sulsky & Day, 1992). The point of this distinction is to assure the AC practitioner that the present study does more than simply show that training assessors can enhance both the construct and criterion-related validity of assessment ratings; rather, it also suggests that the best approach to training may be one that is fundamentally different from the typical approach.

An important, albeit implicit assumption that guided the present study was that there is no true exercise effect (Neidig & Neidig, 1984); that is, dimensions, not exercises, underlie AC ratings. This point is reminiscent of the sign versus sample distinction (Wernimont & Campbell, 1968) with its corresponding questions of how to view the AC and thus validate it. The present position is that the AC is a sign, not merely a work sample, in that it offers an opportunity to observe where assessees fall on dimensions that may be important for future performance. In fact, the criterion-related validity coefficients are strongly supportive of such a position. That is, dimension ratings in an educational AC predicted job performance 4 years later across several different jobs and organizations. It is not reasonable to assume that these exercises met the requirements for being considered work samples of the jobs now held by these assesses. Therefore, the basis behind the criterion-related validity of the present AC must be an accurate evaluation of assessees on dimensions believed to be important for job performance (i.e., a sign) and not merely a work-sample test.

The above discussion begs the question of whether it matters if we are measuring performance on dimensions (a sign) or performance on exercises (a sample), as long as the AC ratings predict performance on the job. There are several reasons to believe that this is an important distinction and that paying attention to the dimensions included in ACs matters. First, as the present study suggests, using a dimension-focused approach to training can have a positive impact on the criterion-related validity of the AC. Second, ACs are increasingly being used for training and development purposes (Cohen, 1980; Spychalski et al., 1997), making the accurate assessment of personnel on construct-valid dimensions increasingly more important. Finally, the emphasis on di-
mensions is crucial to the extent that the AC is designed to assess potential, and early identification of management potential is still one of the most common purposes for ACs (Gaugler et al., 1987).

As Klimoski and Brickner (1987) noted, “In terms of priorities, given the real need of organizations to assess potential (apart from competencies), it would seem most important to establish if, or under what conditions, ACs can be made to produce valid measures of constructs” (p. 255). The present study preliminarily suggests that FOR training provides at least one of these conditions.

References


Received July 24, 2000
Revision received November 20, 2001
Accepted November 20, 2001