Journal of Applied Psychology
2002, Vol. 87, No. 4, 675–686

0021-9010/02/$5.00 DOI: 10.1037//0021-9010.87.4.675

Trying to Understand the Different Pieces of the Construct Validity Puzzle of Assessment Centers: An Examination of Assessor and Assessee Effects

Filip Lievens
Ghent University

This study examined the effects of assessor-related factors (i.e., type of assessor) and assessee-related factors (i.e., type of assessee profile) on the construct validity of assessment center ratings. In particular, 3 types of assessors (26 industrial/organizational [I/O] psychologists, 20 managers, and 27 students), rated assessee performances that varied according to cross-exercise consistency (i.e., relatively inconsistent vs. relatively consistent) and dimension differentiation (relatively undifferentiated vs. relatively differentiated). Construct validity evidence was established for only one assessee profile and only in the I/O psychologist and managerial samples. More generally, these results indicate that 3 factors (poor design, assessor unreliability, and especially cross-situational inconsistent assessee performances) may explain why construct validity evidence is often not established in operational assessment centers.

Assessment centers serve a variety of human resource functions such as selection and development. An important advantage of the developmental assessment center approach is that participants receive detailed feedback concerning their strengths and weaknesses on the managerial dimensions measured. Accordingly, it is critical that the dimensions measured are valid indicators of managerial abilities (Lievens & Klimoski, 2001). Since the early 1980s, however, a recurring theme in the literature is that in assessment centers the quality of construct measurement is relatively poor (e.g., Chan, 1996; Fleenor, 1996; Sackett & Dreher, 1982). The general conclusion is that, within exercises, the distinctions between dimensions are blurred, as scores on one dimension correlate highly with scores on other dimensions (i.e., low discriminant validity). When people are rated on the same dimension in more than one exercise, correlations among the obtained ratings are low (i.e., low convergent validity).

The most frequently researched explanation for these construct validity findings is that they are due to assessors' biases and inaccuracy, which may result from poor assessment center design. This research attention seems warranted because improvements in construct validity were generally found when the design of the assessment center was adjusted to facilitate assessor rating processes (see Arthur, Woehr, & Maldegen, 2000; Lievens & Conway, 2001, for recent reviews). Examples of specific design considerations included limiting the number of dimensions to be rated and providing assessors with behavioral checklists.

This research was supported by a grant as Postdoctoral Fellow of the Fund for Scientific Research, Flanders, Belgium. (F.W.O.—Vlaanderen).

Portions of this article were presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, Louisiana, April 2000. I thank Michael Harris and Wilfried De Corte for their insightful suggestions on a previous version of this article. I also thank Julie Fuller for her editorial assistance with the manuscript.

Correspondence concerning this article should be addressed to Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. E-mail: filip.lievens@rug.ac.be

An additional explanation is that the typical construct validity findings reported (e.g., low convergence among ratings of the same dimension across different exercises) may also represent true cross-situational performance differences of assessees (Highhouse & Harris, 1993; Neidig & Neidig, 1984; Schneider & Schmitt, 1992). In fact, because assessees perform in a very diverse set of exercises, some assessees may simply perform better on the same dimensions in some exercises (e.g., individual exercises) than in other exercises (e.g., group exercises). Two recent studies have found some support for this view (Lance et al., 2000; Lievens, 2001b).

Prior studies have typically focused on one of these two explanations. It seems reasonable, however, that these two explanations are not mutually exclusive and that a combination of them is responsible for the poor construct validity findings in assessment centers. Therefore, this study examines the effects of assessor-related factors (i.e., type of assessor) as well as assessee-related factors (i.e., type of assessee profile) on the construct validity of assessment center ratings. Accordingly, this study aims to provide a more complete understanding of the different pieces of the construct validity puzzle in assessment centers.

Background

The Importance of Assessor-Related Factors

Lievens and Klimoski (2001) offered two theoretical models that might help in understanding how assessors and rating processes influence the quality of construct measurement in assessment centers. The first model posits that assessors possess limited information-processing capacities and therefore are not always able to meet the cognitive demands of the assessment center process. Many studies have implicitly or explicitly used this model to suggest assessment center design modifications for simplifying the cognitively challenging task of assessors. Examples of practical design recommendations have included limiting the number of dimensions rated per exercise (Gaugler & Thornton, 1989; Maher, 1990), asking assessors to rate conceptually distinctive dimensions (Kleinmann, Exler, Kuptsch, & Köller, 1995), providing assessors

with behavioral observation checklists (Donahue, Truxillo, Cornwell, & Gerrity, 1997; Reilly, Henry, & Smither, 1990), specializing rating tasks of assessors so that only one dimension across exercises is rated (Robie, Adams, Osburn, Morris, & Etchegaray, 2000), and enabling assessors to pause and rewind videotaped assessee performances (Ryan et al., 1995). With a few exceptions, these studies were generally effective in reducing assessor cognitive overload, as inferred by improvements in the quality of ratings.

A second model proposed by Lievens and Klimoski (2001) is the expert assessor model. According to this model, differences between novices and experts account for differences in rating quality (Chi, Glaser, & Farr, 1988). Specifically, experienced assessors are expected to possess and to use well-established cognitive structures when rating assessees. These organizing frameworks are helpful because they guide attention, categorization, integration, and recall processes (Fiske & Taylor, 1991; Srull & Wyer, 1989; Zedeck, 1986). Alternatively, novice assessors are not expected to possess such well-established cognitive structures when rating, which may result in poor quality of construct measurement. This model further posits that novice assessors can develop more expertise by abstracting from education (e.g., a degree in psychology), training (e.g., an assessor training program), and experience (e.g., rating experience; Lorenzo, 1984; Sagie & Magnezy, 1997).

In the performance appraisal field, this second model has received some research attention. Cardy, Bernardin, Abbott, Senderak, and Taylor (1987) found that personnel administrators' ratings were more accurate than those of master's in business administration (MBA) students, who, in turn, were more accurate than undergraduates. 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 (e.g., Kozlowski, Kirsch, & Chao, 1986; Kozlowski & Mongillo, 1992) also underscores the role of experience in promoting accurate ratings. In the assessment center field, few studies have used this model to suggest procedural interventions in assessment center design. Two studies reported better quality of construct measurement among assessors who followed a frame-of-reference training program (Lievens, 2001a; Schleicher, Day, Mayes, & Riggio, 1999). In another study, Sagie and Magnezy (1997) investigated the effects of type of assessor and reported that construct validity was higher among psychologist assessors than among managerial assessors. Thus, despite the scarcity of studies regarding the expert assessor model, the results generally indicate that this model shows promise in terms of formulating assessment center design changes, which can increase the quality of construct measurement. Therefore, this study aims to extend research on the expert assessor model by examining the effects of three groups of assessors (i.e., industrial/organizational [I/O] psychologists, managers, and university students) on construct validity. On the basis of prior empirical research (Sagie & Magnezy, 1997), I expected that convergent and discriminant validity would be more clearly established for psychologist assessors than for the other types of assessors.

The Importance of Assessee-Related Factors

Although most studies found improvements in assessment center construct validity by modifying the design of assessment centers to facilitate assessors' rating processes, it should be noted that in other studies these procedural interventions were less successful. For instance, although Chan (1996), Schneider and Schmitt (1992), and Fleenor (1996) carefully implemented many of the aforementioned design recommendations, they found little evidence for construct validity. This suggests that careful assessment center design may install necessary but insufficient conditions for ensuring assessment center construct validity.

In particular, an additional explanation for the lack of construct validity in assessment center ratings is that it results from variation in candidate performances across assessment center exercises. Basically, proponents of this explanation posit that assessors are only partially to blame for the typical construct validity findings because these findings are also due to candidates' real performance differences across situations (Highhouse & Harris, 1993; Neidig & Neidig, 1984; Schneider & Schmitt, 1992). For example, certain individuals may perform better in one-to-one exercises than in group exercises, diminishing the convergence of ratings across exercises.

Empirical research on this explanation is scarce. Two recent studies (Lance et al., 2000; Lievens, 2001b) provided some evidence that, apart from assessor and design factors, assessee factors also play a role in explaining the construct validity puzzle. First, Lance et al. examined whether the exercise effects (i.e., the high correlations among dimension ratings within exercises) represented bias or true cross-situational performance differences. Lance et al. reported on several studies in which they correlated latent exercise factors and external correlates such as cognitive ability measures. In general, their findings supported the hypothesized relationships between the exercise factors and external correlates, suggesting that the exercise factors captured true variance instead of bias. Therefore, Lance et al. argued that assessors are providing relatively accurate assessments of assessees. These assessees, however, do not show performance consistency across exercises, which may explain the construct validity findings typically found. Second, Lievens (2001b) asked assessors to rate videotaped assessees, whose performances varied across dimensions and were relatively consistent across exercises. Results showed that when assessors rated these assessees, they were reasonably able to differentiate among the various dimensions and to use these dimensions consistently across exercises. In other words, similar to Lance et al., assessor ratings of this study were also relatively veridical. Clearly, these two recent studies shed a different and more positive light on the construct validity puzzle. They also call for a deeper understanding and more research on the effects of assessee performances on assessment center construct validity. A limitation of these studies, however, was that assessee performances were not experimentally manipulated, which precluded drawing firm conclusions about the effects of assessee performances on the convergent and discriminant validity of assessment center ratings.

To examine the effects of assessee performances on convergent and discriminant validity, it is relevant to categorize assessee performances along two continua (see Figure 1). A first continuum refers to the degree of *cross-exercise consistency* in assessee performances. On the one hand, assessee performances may be conceptualized as being relatively consistent and stable across situations (i.e., exercises). Accordingly, it is believed that assessee

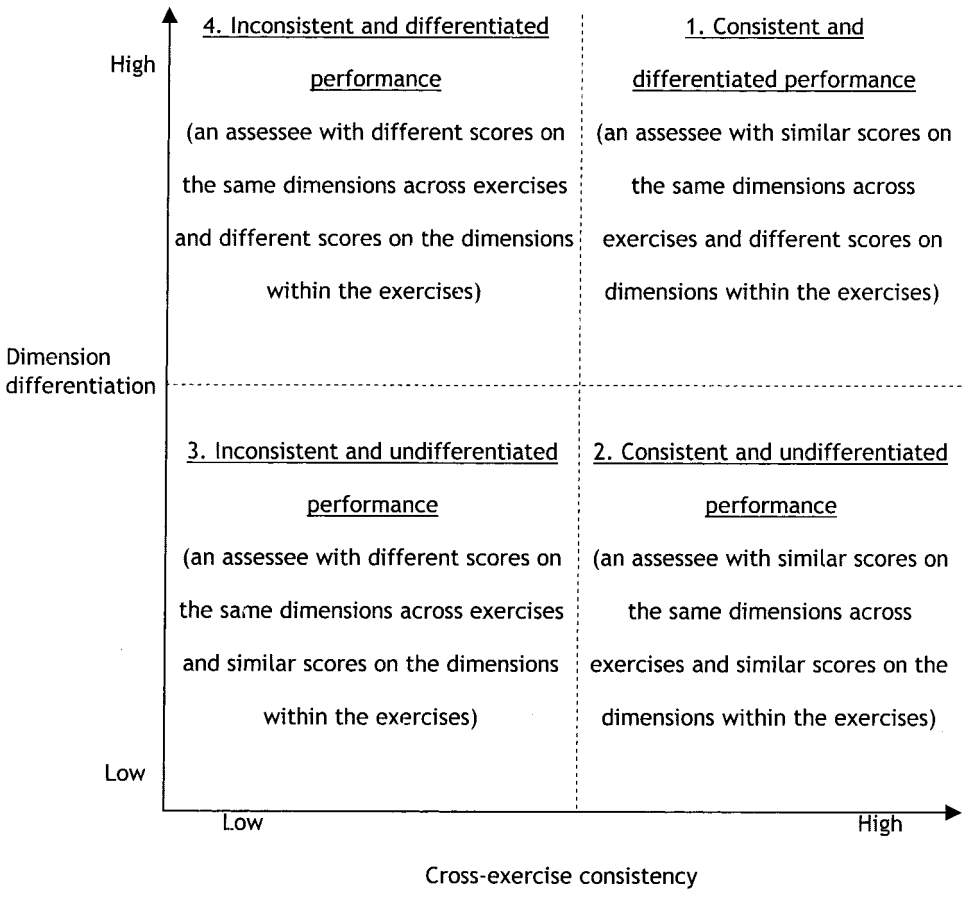

Figure 1. Categorization of assessee performances along two continua: cross-exercise consistency and dimension differentiation.

performances are primarily influenced by dispositional factors (i.e., stable personal characteristics). On the other hand, assessee performances may also be regarded as primarily dependent on the different demand characteristics of the various assessment center exercises. In this case, the underlying belief is that situational factors (e.g., exercise form) mainly determine assessee performances (Highhouse & Harris, 1993; Neidig & Neidig, 1984; Schneider & Schmitt, 1992).

Conceptually, the cross-situational consistency continuum hinges on a long-standing controversy about the stability of behavior between the so-called personalists and situationists in personality and social psychology (see Epstein & O'Brien, 1985; Johnson, 1997). The cross-exercise consistency continuum is especially relevant in light of the evidence, or lack thereof, for convergent validity. If assessors are required to rate candidates with relatively consistent performances across exercises (see right part of Figure 1), evidence for convergent validity should be established. However, if assessors are asked to rate candidates whose performances depend on the situational (exercise) demands (see left part of Figure 1), no evidence for convergent validity should be expected.

A second continuum refers to the degree of *dimension differentiation* in assessee performances. On the one hand, assessees may show relatively large performance variations across dimensions. In other words, assessees may perform well on some dimensions and poorly on other dimensions. On the other hand, assessee performances may also be mainly invariant across dimensions. This implies that assessee performance variability can be brought back to one dominant performance factor.

The dimension differentiation continuum is similar to the discussion of whether job performance is multidimensional (Campbell, McCloy, Oppler, & Sager, 1993) or unidimensional (Viswesvaran, 1996). Conceptually, it also reflects the debate with respect to holistic versus elementalistic perspectives in personality and social psychology (Magnusson & Toerestad, 1993). The holistic view posits that psychological phenomena such as personality can only be assessed globally without using separate variables and dimensions; elementalists, however, believe the opposite to be true. This dimension differentiation continuum is especially relevant for discriminant validity. If assessors are asked to rate candidates whose performances meaningfully differ across dimensions (see upper part of Figure 1), evidence for discriminant validity should be established. Yet, if assessors rate candidates without clear performance fluctuations across dimensions (see lower part of Figure 1), no discriminant validity evidence should be expected.

The Present Study

This study aims to integrate the two aforementioned research streams (i.e., research on assessors and research on assessees) by experimentally manipulating both assessor-related factors (i.e., type of assessor) and assessee-related factors (i.e., type of assessee profile) to determine their effects on convergent and discriminant validity. In particular, three types of assessors (I/O psychologists, managers, and students) were asked to rate assessees whose performances were designed to vary according to cross-exercise consistency (i.e., relatively inconsistent vs. relatively consistent) and dimension differentiation (relatively undifferentiated vs. relatively differentiated).

On the basis of the discussion above about assessor-related and assessee-related factors, I formulated the following hypotheses:

Hypothesis 1a: Evidence for convergent validity will be established for candidates whose performances are consistent across exercises (i.e., Candidate Profiles 1 and 2 of Figure 1).

Hypothesis 1b: Evidence for discriminant validity will be established for candidates whose performances meaningfully differ across dimensions (i.e., Candidate Profiles 1 and 4 of Figure 1).

Hypothesis 2a: Evidence for convergent validity will be more clearly established for I/O psychologists than for either line managers or students.

Hypothesis 2b: Evidence for discriminant validity will be more clearly established for I/O psychologists than for either line managers or students.

Method

Sample

Three different sub-samples were included in the study. The first sub-sample was composed of 26 I/O psychologists (15 women, 11 men; mean age = 34.1 years, *SD* = 4.4 years). All I/O psychologists had been working for several years (minimum of 3 years) as human resource officers or human resource managers in a private or public company. None of them worked for a consultancy firm. The I/O psychologists were enrolled in a special human resource management program about emerging trends and practices.

The second sub-sample was composed of 20 managers (2 women, 18 men; mean age = 34.4 years, *SD* = 4.9 years). All managers had several years of full-time working experience, came from a broad variety of organizations, and had different functional backgrounds (e.g., engineering, sales). They were enrolled in an executive MBA program, which aimed to provide them with broader managerial skills (in addition to their technical proficiency).

The third sub-sample consisted of 27 individuals, who followed a specialized 1-year full-time program in personnel management after graduating. The sample included 20 women and 7 men with a mean age of 26 years and 5 months (*SD* = 6.4 years). These students had a diversity of educational backgrounds (e.g., law, business). However, none of them had a degree in I/O psychology.

Assessment Center Simulation

Participants were told that they would participate as assessors in an assessment center simulation. This provided them with an opportunity to observe and rate assessment center candidates. In particular, they were asked to evaluate four videotaped candidates applying for the job of district sales manager. Assessors knew that afterwards they would be expected to explain their ratings to one another. This common assessment center practice served as an incentive to take the assessor task seriously.

Prior to serving as assessors, participants were given assessor training (2 hr). First, assessment centers were situated among other personnel selection techniques, assessment centers were defined, and their basic components were delineated. Next, the target job and the target organization were described to the participating assessors. This implied that assessors received details about the main tasks and qualifications required for successful district sales managers. They also knew the job context of the district sales manager (e.g., place in organizational tree, number of subordinates). Regarding the organization, assessors received information on the type of business, the level of decentralization, the workforce size, the market share, and the organizational culture. Pictures of the products were displayed. These job and organization details were extracted from a real job posting and an actual annual report of an organization. Next, the assessor training covered the process of observing, recording, classifying, and evaluating assessee behavior (see Byham, 1977). The trainer instructed assessors to make clear behavioral notes of assessee behavior instead of vague nonbehavioral interpretations. In particular, the principles behind careful observation were discussed and examples of behavioral versus nonbehavioral observations were presented. The trainer also taught assessors to categorize behavior by dimensions. To this end, assessors received the behaviorally anchored definitions of the six dimensions, which were relevant for the target job of district sales managers. The dimensions were the following: problem analysis, listening, planning and organizing, improvisation, initiative, and oral communication. For example, *planning* and *organizing* was defined as "the ability to systematically structure own and others' activities to achieve maximum work performance." Planning and organizing behaviors included setting a concrete agenda, managing the scarce time properly, not jumping from one subject to another, making concrete and specific (follow-up) agreements, and formulating concrete deadlines. The last concept described to the assessors was the rating of dimensions according to the behavior observed. In particular, assessors were familiarized with the rating scale and the performance standards. Finally, the exercises (i.e., sales presentation and group discussion), which were relevant for the target job, were reviewed.

After this assessor training, assessors were randomly assigned to small teams, which were placed in separate rooms. Next, they observed the videotaped performance of the first candidate in the sales presentation, recorded observations, and independently provided dimensional ratings. This process was repeated for the presentation performance of the other three candidates and for the group discussion (in which the four candidates performed together). This observation and rating session lasted for about 2 hr. To control for order effects, the study design included the development of four versions of the integral film (each with a different candidate order). The order of the exercises was the same in all four versions. The assessor groups were randomly assigned to a particular version of the film. An equal number of assessors viewed each version. Irrespective of the videotaped version, all assessors rated all candidates in every exercise, creating a fully crossed design (Jones, 1992). After observing and rating candidates, assessors met in their teams to share observations, discuss ratings, and write assessee reports (1.5 hr).

Taken together, this whole procedure (including several breaks) lasted for about 6 hr. This study's simulated assessor environment converged closely with current assessment center practices in organizations (Spychalski, Quinones, Gaugler, & Pohley, 1997) and with previous assessment center simulations (see Gaugler & Rudolph, 1992, for an example). Furthermore, this simulation met virtually all of the 10 essential elements of an assessment center delineated by the Guidelines and Ethical Considerations for Assessment Center Operations (Task Force on Assessment Center Guidelines, 1989). The only exception was that assessors were not sys-

Table 1
Mean Expert Assessor Ratings of Candidate Performances in the Assessment Center Exercises

	Profile 1: Consistent and differentiated performance		Profile 2: Consistent and undifferentiated performance		Profile 3: Inconsistent and undifferentiated performance		Profile 4: Inconsistent and differentiated performance	
Dimension	Presentation	Group	Presentation	Group	Presentation	Group	Presentation	Group
Problem analysis	3.0	3.2	1.8	1.8				
Listening	4.4	4.2	1.4	1.4	1.6	3.0		
Planning and organization			2.0	1.6				
Improvization					1.6	3.2	1.8	2.8
Initiative					1.8	3.4	2.8	4.2
Oral communication	3.4	3.6					3.0	3.8

Note. $N = 5$. A score of 1 indicates poor performance, and a score of 5 indicates excellent performance. Presentation = mean expert assessor ratings of the candidate in the sales presentation; group = mean expert assessor ratings of the candidate in the group discussion. Blank cells indicate that the dimension was not rated by the expert assessors because it was not manipulated (i.e., no behavioral incidents indicative of this dimension were built into the scripts).

tematically evaluated at the end of the training. However, this is also seldom done in operational assessment centers (Spychalski et al., 1997).

Videotaped Assessee Performances

Design. The design underlying the candidate performance profiles was a 2 (cross-exercise consistency) × 2 (dimension variation) design. The levels of the factor cross-exercise consistency included relatively inconsistent performance across exercises versus relatively consistent performance across exercises (job-related exercises were a sales presentation exercise and a leaderless group discussion). The levels of dimension variation were relatively undifferentiated performance across dimensions versus relatively differentiated performance across dimensions (these job-related dimensions were problem analysis, listening, planning/organizing, improvisation, initiative, and oral communication). Crossing these levels resulted in four candidate performance profiles (see four quadrants of Figure 1).

Development of videotapes. To ensure realism, scripts were developed with the help of two experienced professional assessors. The assessors qualified as experts because of their extensive practical experience as assessors and their theoretical knowledge of assessment centers. These experienced assessors were asked to construct assessment center performances around each of the candidate profiles (see four quadrants in Figure 1). To this end, they were encouraged to think of real candidates. It soon became clear that it was neither realistic to vary each of the six dimensions per candidate profile nor to vary the same dimensions per profile. Therefore, it was decided to vary only three of the six dimensions per candidate profile. In addition, the same three dimensions were not chosen for each of the candidate profiles. For example, it was decided to build behaviors indicative of listening, problem analysis, and oral communication into the performance of Candidate Profile 1 (i.e., differentiated and consistent performance; upper right quadrant of Figure 1). No behaviors indicative of the other three dimensions were built into Candidate Profile 1. As an operationalization of dimension differentiation, Candidate 1 was designed to perform highly on listening but moderately on both problem analysis and oral communication. As an operationalization of exercise consistency, these performance levels were designed to be similar in the first candidate's sales presentation and group discussion scripts. For the second candidate (i.e., undifferentiated and consistent performance; lower right quadrant of Figure 1) behaviors indicative of problem analysis (poor), listening (poor), and planning and organization (poor) were built into both the sales presentation and group discussion scripts. Again, the scripts of Candidate 2 contained no behaviors relevant to the other three dimensions. The same logic was used to operationalize the third and fourth candidate profiles. The final scripts depicted the word-for-word dialogue for each performance. The experienced assessors reread these scripts and made adjustments in light of realism. Semi-professional actors were filmed delivering the scripted assessment center performances. After professional editing, the videotaped performances ran between 7 min (presentation) and 14 min (discussion). The length of the whole set of videotapes was about 40 min.

Manipulation check. Procedures by Sulsky and Balzer (1988) were followed to verify whether the videotaped performances reflected the performance profiles of Figure 1. Five professional assessors viewed each videotaped performance. They could view the tape repeatedly and rewind it. All experts independently rated the dimensions, which were manipulated for each assessee performance, on a 5-point scale ranging from *poor* (1) to *excellent* (5). Interrater agreement among the expert ratings equaled .81 (intraclass correlation = 2.1; Shrout & Fleiss, 1979). The mean assessor expert ratings are presented in Table 1. On the whole, the expert ratings closely reflected the candidate profiles shown in Figure 1. The only exception was that the expert ratings reflected somewhat less exercise inconsistency in the performances of Candidate 4 (i.e., the performances on the dimensions of improvisation and oral communication) than anticipated.

Measures

Participants in the assessment center simulation recorded observations on an observation form and completed a rating form for each videotaped performance to evaluate six dimensions.[1] The dimensions were rated on a 5-point scale ranging from *poor* (1) to *excellent* (5). The behaviorally anchored definitions of the dimensions were available to the assessors.

Analyses

In this study, generalizability analysis (Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Marcoulides, 1989) was

[1] As previously mentioned, the videotaped assessee performances were built around three dimensions. Yet, assessors rated six dimensions. Hence, assessors also had to rate dimensions that were not a priori built into the assessee performances. This enhanced the realism of the assessor task. In operational assessment centers, assessors usually evaluate assessees on more than three dimensions. In addition, in operational assessment centers, exercises typically vary in the opportunity for behavior representing a dimension to be manifested (Reilly, Henry, & Smither, 1990). Therefore, it is not unusual for assessors to rate candidates on dimensions that are less observable.

used to partition the sources of variance in assessment center scores and accordingly provide information on assessment center construct validity. Although the multitrait–multimethod approach and confirmatory factor analysis traditionally have been used to examine assessment center construct validity, generalizability analysis has also been used in some studies (Arthur et al., 2000; Lievens, 2001b). In these studies, the generalizability analysis results were found to be very similar to the results of the multitrait–multimethod matrix and the confirmatory factor analysis.

In the present study, generalizability analyses were conducted within each assessee profile so that it was possible to compare the different assessee profiles in terms of the sources of variance, which were relevant to convergent and discriminant validity. The generalizability analyses within each assessee profile had three facets (i.e., factors affecting the measurement process): type of Assessor (T), Assessors (A), and Exercises (E). The Assessors and Exercises facets were completely crossed with each other. Dimensions (D) were not considered a facet but served as the object of measurement. Only the dimensions that were manipulated in the videotaped performances, were used in each of the generalizability analyses. Note also that the Assessor facet was nested in the Type of Assessor facet (an assessor was either a psychologist, a manager, or a student). Because the Assessor facet was nested in the Type of Assessor facet and because generalizability analysis requires a balanced design, the number of psychologist, managerial, and student assessors included in the generalizability analysis had to be the same. For instance, in the generalizability analysis within Candidate 1, six randomly chosen psychologists and seven randomly chosen students were not included, so as to result in an equal ($n = 20$) number of psychologist, managerial, and student assessors.

Variance components are estimated in generalizability analysis. A variance component reflects a facet's contribution to the total variance. Applied to this study's within-candidate generalizability analyses, the variance components represent the variances of the mean candidate ratings attributable to the Dimensions (D) (object of measurement), to the Type of Assessor (T), to the Assessors nested within Assessor Type (A:T), to the Exercises (E), and to the respective interactions among them. Estimated variance components depend on the scale of measurement (in this case a 5-point rating scale). Hence, it is important to interpret variance components by their relative magnitudes (Shavelson & Webb, 1991). To this end, the percentage contribution of each variance component was used. This percentage contribution refers to the percentage of the sum of the variance components (i.e., the total variance) accounted for by each variance component.

Some variance components estimated are especially relevant for examining construct validity (see Kane, 1982; Kraiger & Teachout, 1990, for detailed discussions) and therefore are relevant for testing the hypotheses proposed. Specifically, in this study, evidence for convergent validity is derived from the variance component of Exercises. A low value of this Exercises variance component suggests invariance of a specific candidate rating across exercises. Evidence for discriminant validity is derived from the variance component associated with the Dimensions. A high value of this Dimensions variance component indicates substantial differences in ratings of a specific candidate across dimensions.

Consistent with practice, generalizability analyses were performed with a random effects design (Shavelson & Webb, 1991). Data were analyzed with GENOVA (Version 2.2), a Fortran-based program developed for generalizability analyses (Crick & Brennan, 1983).

Results

The first set of hypotheses was related to the effects of the candidate profiles on convergent and discriminant validity. Table 2 presents the results of generalizability analyses within each of the four candidate profiles. Included are the degrees of freedom, the estimated variance components, and their 90% confidence intervals. These confidence intervals were computed by procedures outlined in Brennan (1992). Hypothesis 1a stated that evidence for convergent validity would be established for assessors' ratings of candidate profiles with relatively consistent performances across exercises. In line with Hypothesis 1a, convergent validity evidence was found for the "consistent" candidate profiles (i.e., Profiles 1 and 2) because these generalizability analyses yielded small Exercise variance components (both 3%). Similarly, no evidence for convergent validity (i.e., a substantial Exercise variance component) was found for candidate profiles with relatively inconsistent performances across exercises. For example, the variance component due to Exercises for the inconsistent Candidate Profile 3 explained 30% of the variance.

According to Hypothesis 1b, evidence for discriminant validity would be established for assessors' ratings of candidate profiles with relatively differentiated performances across dimensions (i.e., Candidate Profiles 1 and 4). Consistent with this hypothesis, a substantial variance component due to Dimensions (object of measurement) was found for Candidate Profile 1 (31%). The variance component due to Dimensions was also among the largest variance components of the generalizability analysis within Candidate Profile 4 (16%). Similarly, no evidence for discriminant validity was found for candidate profiles with relatively undifferentiated performances across dimensions. For example, the variance component due to Dimensions for the undifferentiated Candidate Profile 2 accounted for 0% of the variance.

Hypotheses 2a and 2b were related to differences between the assessor samples in terms of convergent and discriminant validity. Two interactions (T × E and T × D) provide information about this hypothesis. The variance component of the Type of Assessor × Exercises (T × E) interaction indicates whether Exercise variance varies according to the type of assessor and therefore informs whether the different types of assessors differ in terms of evidence for convergent validity. Similarly, the variance component of the Type of Assessor × Dimensions (T × D) interaction indicates whether Dimension variance (discriminant validity) varies according to the type of assessor. Table 2 shows that in the generalizability analysis within Candidate Profile 1, the variance component of the Type of Assessor × Exercises (T × E) interaction explained 4% of the variance, and the variance component of the Type of Assessor × Dimensions (T × D) explained 0%. In the other generalizability analyses, similar percentages for these variance components were obtained, suggesting that there were at best only small differences between assessor samples in terms of convergent and discriminant validity. The finding that assessor differences were smaller than candidate profile differences is not

Table 2
Generalizability Study Variance Components Within Candidate Profiles Across All Samples

Effect	*df*	VC	90% confidence intervals	Explained variance (%)
Candidate Profile 1 (consistent and differentiated performance)				
T(ype of Assessor)	2	.00[a]		
A(ssessors):T	57	.07	.03 < VC < .20	6
E(xercises)	1	.03	.02 < VC < .10	3
D(imensions)	2	.34	.18 < VC < 1.00	31
T × E	2	.04	.02 < VC < .12	4
T × D	4	.00	.00 < VC < .01	0
A × E:T	57	.08	.04 < VC < .22	7
A × D:T	114	.04	.02 < VC < .12	4
E × D	2	.05	.03 < VC < .14	4
T × E × D	4	.00[a]		
A × E × D:T	114	.46	.38 < VC < .58	42
Candidate Profile 2 (consistent and undifferentiated performance)				
T	2	.04	.02 < VC < .10	4
A:T	54	.15	.09 < VC < .34	18
E	1	.02	.01 < VC < .07	3
D	2	.00[a]		
T × E	2	.00[a]		
T × D	4	.04	.02 < VC < .12	5
A × E:T	54	.12	.07 < VC < .23	14
A × D:T	108	.13	.09 < VC < .24	16
E × D	2	.03	.02 < VC < .09	3
T × E × D	4	.06	.03 < VC < .19	8
A × E × D:T	108	.26	.21 < VC < .33	30
Candidate Profile 3 (inconsistent and undifferentiated performance)				
T	2	.00[a]		
A:T	57	.00	.00 < VC < .00	0
E	1	.50	.26 < VC < 1.48	30
D	2	.12	.06 < VC < .36	7
T × E	2	.04	.02 < VC < .13	3
T × D	4	.00[a]		
A × E:T	57	.33	.22 < VC < .54	19
A × D:T	114	.18	.12 < VC < .35	11
E × D	2	.01	.00 < VC < .02	0
T × E × D	4	.05	.03 < VC < .15	3
A × E × D:T	114	.43	.35 < VC < .54	26
Candidate Profile 4 (inconsistent and differentiated performance)				
T	2	.01	.01 < VC < .03	1
A:T	54	.08	.04 < VC < .23	7
E	1	.10	.05 < VC < .30	9
D	2	.19	.10 < VC < .55	16
T × E	2	.00[a]		
T × D	4	.03	.01 < VC < .08	2
A × E:T	54	.18	.11 < VC < .38	15
A × D:T	108	.11	.06 < VC < .32	9
E × D	2	.00	.00 < VC < .01	0
T × E × D	4	.02	.01 < VC < .05	2
A × E × D:T	108	.46	.37 < VC < .58	39

Note. VC = estimated variance components. Within each candidate profile, a generalizability analysis was conducted.
[a] Small negative estimates of variance components were reported as zero (see recommendations of Shavelson & Webb, 1991).

unexpected because candidate profiles were experimentally manipulated, whereas the assessor groups were existing groups.

To inspect the differences in terms of convergent and discriminant validity across assessor samples in more detail, generalizability analyses within the four candidate profiles were conducted per assessor sample. Accordingly, it was possible to examine which assessor samples displayed the highest convergent and discriminant validity (see Hypotheses 2a and 2b). Each of these general-

izability analyses had two facets: Assessors (A) and Exercises (E), which were completely crossed with each other. Again, Dimensions (D) served as the object of measurement. Because these analyses were conducted per sample, the Assessor facet was no longer nested within the Type of Assessor facet. Therefore, all assessors of each sample were included.

According to Hypothesis 2a, evidence for convergent validity would be more clearly established for I/O psychologists than for either line managers or students. As can be seen in Table 3, the generalizability analyses within the "consistent" Candidate Profile 1 show that Exercises explained more variance in the student sample than in the I/O psychologist and managerial assessor samples (16% as compared with 0% and 2% in the I/O psychologist and managerial samples, respectively). The same was also true to a lesser extent for Candidate Profile 2. Although this candidate performed consistently across exercises, the Exercise variance component equaled 7% in the student assessor sample (as compared with 0% in both the I/O psychologist and managerial assessor samples). These results are not consistent with Hypothesis 2a because both I/O psychologist and managerial assessors outperformed students in terms of convergent validity.

According to Hypothesis 2b, discriminant validity would be more clearly established in the ratings of I/O psychologist assessors than in the ratings of managerial and student assessors. As can be seen in Table 3, a substantial variance component (36%) due to Dimensions was found for the "differentiated" Candidate Profile 1 in both the I/O psychologist and managerial samples. Conversely, in the student assessor sample, Dimensions accounted for only 23% of the variance in ratings of Candidate 1. The generalizability analysis results of the other "differentiated" Candidate Profile 4

Table 3
Generalizability Study Variance Components Within Candidate Profiles for Psychologist, Managerial, and Student Samples

	Psychologists				Managers				Students			
Effect	*df*	VC	90% confidence intervals	Explained variance (%)	*df*	VC	90% confidence intervals	Explained variance (%)	*df*	VC	90% confidence intervals	Explained variance (%)
Candidate Profile 1 (consistent and differentiated performance)												
Assessors (A)	23	.05	.03 < VC < .15	5	19	.04	.02 < VC < .12	3	25	.1	.05 < VC < .29	8
Exercises (E)	1	.02	.01 < VC < .06	2	1	.00[a]		0	1	.2	.10 < VC < .59	16
Dimensions (D)	2	.33	.17 < VC < .97	36	2	.42	.22 < VC < 1.23	36	2	.29	.15 < VC < .85	23
A × E	23	.08	.04 < VC < .23	9	19	.05	.03 < VC < .15	4	25	.1	.05 < VC < .29	8
A × D	46	.06	.03 < VC < .18	7	38	.04	.02 < VC < .12	3	50	.07	.04 < VC < .20	6
E × D	2	.03	.02 < VC < .09	3	2	.07	.04 < VC < .20	6	2	.04	.02 < VC < .12	3
A × E × D	46	.35	.26 < VC < .50	38	38	.55	.40 < VC < .82	47	50	.44	.32 < VC < .66	35
Candidate Profile 2 (consistent and undifferentiated performance)												
A	23	.00[a]		0	18	.12	.06 < VC < .35	14	26	.39	.24 < VC < .79	41
E	1	.00	.00 < VC < .00	0	1	.00[a]		0	1	.07	.04 < VC < .20	7
D	2	.01	.01 < VC < .03	2	2	.05	.03 < VC < .15	6	2	.02	.01 < VC < .06	2
A × E	23	.12	.06 < VC < .35	20	18	.04	.02 < VC < .12	5	26	.08	.04 < VC < .23	8
A × D	46	.08	.04 < VC < .23	13	36	.21	.12 < VC < .51	25	52	.11	.06 < VC < .32	11
E × D	2	.09	.05 < VC < .26	15	2	.15	.08 < VC < .44	18	2	.01	.01 < VC < .03	1
A × E × D	46	.31	.23 < VC < .44	51	36	.27	.19 < VC < .41	32	52	.28	.21 < VC < .39	29
Candidate Profile 3 (inconsistent and undifferentiated performance)												
A	21	.01	.01 < VC < .03	1	19	.11	.06 < VC < .32	5	26	.00[a]		0
E	1	.76	.39 < VC < 2.23	46	1	.84	.43 < VC < 2.46	41	1	.18	.09 < VC < .53	14
D	2	.13	.07 < VC < .38	8	2	.06	.03 < VC < .18	3	2	.14	.07 < VC < .41	11
A × E	21	.1	.05 < VC < .29	6	19	.39	.22 < VC < .96	19	26	.31	.18 < VC < .73	24
A × D	42	.17	.09 < VC < .50	10	38	.14	.07 < VC < .41	7	52	.12	.06 < VC < .35	9
E × D	2	.00[a]		0	2	.05	.03 < VC < .15	2	2	.03	.02 < VC < .09	2
A × E × D	42	.49	.36 < VC < .71	30	38	.48	.34 < VC < .73	23	52	.52	.39 < VC < .74	40
Candidate Profile 4 (inconsistent and differentiated performance)												
A	21	.2	.10 < VC < .59	18	18	.00[a]		0	26	.02	.01 < VC < .06	1
E	1	.09	.05 < VC < .26	8	1	.01	.01 < VC < .03	1	1	.17	.09 < VC < .50	13
D	2	.19	.10 < VC < .56	17	2	.19	.10 < VC < .56	19	2	.27	.14 < VC < .79	20
A × E	21	.06	.03 < VC < .18	5	18	.19	.10 < VC < .56	19	26	.34	.19 < VC < .78	25
A × D	42	.13	.07 < VC < .38	12	36	.07	.04 < VC < .20	7	52	.07	.04 < VC < .20	5
E × D	2	.01	.01 < VC < .03	1	2	.05	.03 < VC < .15	5	2	.02	.01 < VC < .06	1
A × E × D	42	.45	.32 < VC < .68	40	36	.48	.34 < VC < .73	48	52	.47	.35 < VC < .66	35

Note. VC = estimated variance components. Within each candidate profile, a generalizability analysis was conducted. This was done for each sample. Hence, this table presents the results of 12 separate generalizability analyses.
[a] Small negative estimates of variance components were reported as zero (see recommendations of Shavelson & Webb, 1991).

showed no substantial differences in the variance component due to Dimensions (D) across samples. Taken together, these results are not in line with Hypothesis 2b.

In addition to these results, which were directly relevant to this study's hypotheses, it is also interesting to inspect other variance components. An example is the variance component due to Assessors, which indicates whether a specific candidate is rated differently by assessors (averaging over dimensions and exercises). Hence, small variance components attributable to assessors demonstrate high interrater reliability (Kane, 1982; Kraiger & Teachout, 1990). Likewise, the Assessors × Dimensions interaction or the Assessors × Exercises interaction indicate whether assessor ratings vary in terms of dimensions and exercises, respectively. As can be seen in Table 3, these assessor-related variance components varied considerably across candidate profiles and assessor samples, showing that for some candidate profiles interrater reliability among assessors was problematic.

Discussion

Construct Validity Puzzle

This study aimed to provide a more complete understanding of the different pieces of the construct validity puzzle in assessment centers. To this end, both assessor-related and assessee-related factors were manipulated to determine their effects on convergent and discriminant validity. The results show that evidence for convergent validity is established when assessors rate candidates who perform consistently across exercises. Similarly, evidence for discriminant validity is found when assessors rate candidates who perform differently on dimensions. No evidence for both convergent and discriminant validity is found for the other candidate profiles. In this study, evidence for discriminant and convergent validity also varies according to the type of assessor; however, the differences are smaller than the effects of the candidate profiles. In particular, evidence for discriminant and convergent validity is more clearly established in the I/O psychologist and managerial assessor samples than in the student assessor sample. In other words, this study demonstrates that type of assessee performance profile and type of assessor influence the construct validity of assessment center ratings. An additional finding is that for some candidate profiles, interrater reliability was often low among assessors.

The main contribution of this study is that these results may help to explain why construct validity is so difficult to establish in operational assessment centers. Actually, this study reveals that at least three conditions should be satisfied to establish construct validity in the field. First, practitioners should pay attention to careful assessment center design. In particular, this refers to the choice of assessors. In this study, for instance, asking inexperienced non-psychology students to serve as assessors resulted in reduced rating quality. On the basis of a recent large-scale evaluation of assessment center design interventions (Lievens & Conway, 2001), similar negative effects can be expected when assessment center users include too many dimensions per exercise, do not use behavioral checklists, or use a set of very different exercises. If the design of the assessment center is undermined by one of these factors, the quality of construct measurement will be seriously reduced.

As a second condition to establish construct validity evidence in operational assessment centers, there should be high interrater reliability among assessors. If interrater reliability among assessors of operational assessment centers is at best moderate (as illustrated by this study and many prior studies; Thornton, 1992, p. 114), this variance due to assessors will be necessarily confounded with variance due to exercises. This is because, in operational assessment centers, assessors typically rotate through the various exercises and therefore do not rate candidates in all exercises. Because of this confounding, part of the large exercise variance (exercise effects) usually observed in construct validity studies of operational assessment centers may in fact be assessor variance (Howard, 1997; Jones, 1992). In this study, all assessors rated all candidates in each exercise so that separate estimates of exercise and assessor variance were obtained.

Third and most important, this study shows that careful assessment center design and assessor reliability are necessary but insufficient conditions for establishing evidence for convergent and discriminant validity in operational centers. This is because the nature of candidate performances may be limiting factors to establish evidence for construct validity. This is evidenced by the generalizability analysis results of Candidate Profile 2 (i.e., consistently low-performing candidates), Candidate Profile 3 (i.e., candidates performing poorly on many dimensions in one exercise and excellently on many dimensions in another exercise), and Candidate Profile 4 (i.e., candidates performing poorly on some dimensions in one exercise and excellently on these same dimensions in another exercise). Apparently, if many assessment center candidates perform similarly to these three candidate profiles, even a well-designed assessment center will not guarantee evidence for convergent and discriminant validity.

A problem in prior studies on assessment center construct validity is that it was implicity assumed that candidates performed similarly to Candidate Profile 1, namely as candidates, who perform differently on dimensions and consistently across exercises. Consequently, when prior studies found no evidence for distinct constructs, the validity and the design of an assessment center was questioned. However, when candidates did not exhibit performance variation across dimensions, and no evidence for discriminant validity was found, this result was, in fact, correct. Stated differently, when assessors provided veridical ratings of assessee performance levels on the different dimensions in this condition, the discriminant validity should have been low. Similar examples might be given for (lack of) convergent validity.

These examples illustrate that in prior research the implicit assumptions about the true performance levels of assessees have not been realistic. Therefore, a methodological implication of this study's results is that researchers should reconsider these implicit assumptions when interpreting convergent and discriminant validity evidence. If one acknowledges that candidates often do not perform differently across dimensions and consistently across exercises, the results of prior construct validity studies in the assessment center field are less puzzling than was previously thought. In any case, it is crucial that researchers explicitly consider the nature of candidate performances when looking at convergent and discriminant validity results of assessment centers because "without knowledge of true performance levels, it is difficult to interpret the meaning of convergent and discriminant validity indices" (Smither, Barry, & Reilly, 1989, p. 149).

The present study's finding that assessee performance profiles may be key determinants of construct validity evidence also has practical implications. One important message to practitioners is that assessor ratings of candidates may be more veridical than previously thought. After all, for two assessor groups, the magnitude of the variance components closely paralleled the elements built into the candidate performances. Thus, despite considerable research attention on assessor biases, this study shows that substantial accuracy exists in assessor ratings. These results strengthen recent findings of Lance et al. (2000) and Lievens (2001b), who showed that assessor ratings might reflect true candidate performances rather than biases. Similar conclusions about accuracy on the part of raters have also been drawn in the broader performance appraisal field (Borman, 1978; Borman, White, Pulakos, & Oppler, 1991; Smither et al., 1989). Second, at a practical level, it is also important to understand that efforts to "fix" an assessment center and to increase its convergent and discriminant validity may be less effective when candidates' performance profiles are the primary reason behind the poor convergent and discriminant validity of that specific assessment center. This may also explain why prior field-based examinations of well-designed assessment centers (e.g., Chan, 1996; Fleenor, 1996; Schneider & Schmitt, 1992) showed weak evidence for convergent and discriminant validity.

Assessor Differences

In line with previous studies regarding the expert assessor model, an assessment center design factor, such as the type of people serving as assessors (e.g., Sagie & Magnezy, 1997), influences the quality of construct measurement. In particular, the generalizability analyses indicated that the convergent and discriminant validity was lower in the student sample than in the I/O psychologist and managerial samples.

Differing levels of rating experience may explain these rating quality differences. Recently, Dipboye and Jackson (1999) described rating experience in selection situations as the possession of more detailed, complex, and organized knowledge structures that allow for more accurate and efficient information gathering and processing. Specifically, in this study, the I/O psychologists had gained prior rating experience in personnel selection because they had been working for a minimum of 3 years in human resources departments. About 10% of them had also served as assessors in operational assessment centers. Similarly, it can be assumed that the line managers had acquired rating experience outside a selection context, namely, as part of the process of conducting performance appraisals. I do not know whether any of them may already have served as assessors. The students, however, lacked both selection and general rating experience. Therefore, they probably did not have the resulting frame-of-reference for rating (prospective) employees.

The fact that convergent and discriminant validity is established in both the I/O psychologist and line manager samples and that this may have resulted from different rating experiences in the past (selection experience vs. performance appraisal experience) shows that rating experience is not unidimensional. This is in line with the general model of work experience (Quinones, Ford, & Teachout, 1995), which specifies two dimensions along which experience can vary: measurement mode (amount of experience, time-based experience, and diversity of experience) and specificity (task, job, and organizational level). Future studies can use this multidimensional model to obtain a more fine-grained measure of assessors' rating experience. Researchers should also examine how rating experience, education, and participation in assessor training programs all contribute to rating expertise (see Ericsson & Lehmann, 1996, for a more general discussion of expertise).

Limitations

One limitation of this study is related to the interpretation of the generalizability analysis results. In particular, no statistical tests have been developed for comparing variance components across different generalizability analyses (Brennan, 1992). This limitation made it difficult to compare the generalizability analysis results across profiles. Despite this limitation, generalizability theory seemed the most appropriate technique for partitioning assessor variance and for testing this study's hypotheses. Another 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, Kraiger and Teachout suggested a priori predicting the relative size of effects within the design. In addition, the percentage contribution may serve as a way of interpreting the effects (Shavelson & Webb, 1991). In adherence with these suggestions, a priori predictions about the relative size of variance components were made, and the percentage contribution was used as a heuristic to interpret the magnitude of the variance components.

Although substantial efforts were undertaken in creating a realistic assessment center simulation, questions might also be raised regarding the realism of the candidate profiles used. First, some people may argue that the videotaped assessee performances are relatively short, Granted, the length of the various assessee performances (between 7 and 14 min) is somewhat shorter than in real assessment centers. This was done to keep the whole procedure, which already lasted for 6 hr, feasible. Second, other people may argue that the videotaped candidates are straightforward to rate. For example, it may be easier to rate a person who is uniformly low (e.g., Candidate Profile 2) on all dimensions than one who is uniformly mediocre (e.g., Candidate Profile 1). Therefore, it would have been preferable to have a mediocre mean performance level across all candidate profiles. Although I did not examine whether the simulation was perceived as easy, some student participants said that it was often difficult to rate the relatively short assessee performances. This anecdotal evidence was confirmed by looking at the generalizability analysis results in the student sample. Third, some assessment center researchers and practitioners may argue that assessment centers are devised to measure managerial abilities, which should be transferable from one exercise to another and, in the end, to the target job. Yet, others may posit that cross-situational consistency in assessee performances should not be expected because otherwise it would not be necessary to include multiple exercises. This study does not take a position in this controversy but operationalizes each of the different perspectives (see Figure 1). Furthermore, extreme candidate profiles are deliberately not constructed because the profiles are formulated in terms of "*relatively* differentiated performances across dimensions" and "*relatively* consistent performances across exercises." For example, relatively differentiated performances are operationalized by

performance levels between 3.0 and 4.4 (Candidate Profile 1 in presentation; see Table 1) or between 1.8 and 3.0 (Candidate Profile 4 in presentation) instead of by performance levels between 1 and 5. Although this may have made the manipulation check results less clear (e.g., Candidate Profile 4; see Table 1), the use of such more moderate profiles contributes to the realism of the videotaped performances.

Implications for Future Research

Given the results of this study, two routes seem particularly fruitful for future research on assessment center construct validity. The first and more traditional route consists of further investigating which assessment center design factors positively influence the quality of construct measurement in assessment centers. It is essential that this research proceed in a theory-driven manner. Along these lines, Lievens and Klimoski (2001) have offered various suggestions on how social cognition may advance the understanding of the assessment center process and the quality of construct measurement.

The second route does not focus on assessment center design but on assessee performances. Now that it is clear that assessee profiles can affect assessment center construct validity, the next research question should address which factors influence assessee profiles and cross-situational assessee consistency. In other words, future studies are needed to examine under which conditions candidates in operational centers adjust their behavior from one exercise to another. The answer to this question is likely to be complex, as individual differences variables (self-monitoring, impression management), trait-related variables (see Tett & Guterman, 2000), and situational variables (exercise characteristics) may lead candidates to perform differently across exercises. Future studies are needed to ascertain which of these variables are responsible for the cross-situationally inconsistent performances of assessees.

References

Arthur, W., Woehr, D., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management, 26,* 813–835.

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

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76,* 863–872.

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, J. P., McCloy, R. A., Oppler, S. H., & Sager C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco, CA: Jossey Bass.

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.

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organisational Psychology, 69,* 167–181.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A GENeralized analysis Of VAriance system* (ACT Technical Bulletin No. 43). Iowa City, IA: American College Testing Program.

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.

Dipboye, R. L., & Jackson, S. L. (1999). Interviewer experience and expertise effects. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 279–292). Thousand Oaks, CA: Sage.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behavior and Personality, 12,* 85–108.

Epstein, S., & O'Brien, E. J. (1985). The person–situation debate in historical and current perspective. *Psychological Bulletin, 98,* 513–537.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence for maximal adaptation to task constraints. *Annual Review of Psychology, 47,* 273–305.

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

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10,* 319–333.

Gaugler, B. B., & Rudolph, A. S. (1992). The influence of assessee performance variation on assessors' judgments. *Personnel Psychology, 45,* 77–98.

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.

Highhouse, S., & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology, 23,* 140–155.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality, 12,* 13–52.

Johnson, J. A. (1997). Units of analysis for the description and explanation of personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 73–93). London: Academic Press.

Jones, R. G. (1992). Construct validation of assessment center final dimension ratings: Definition and measurement issues. *Human Resource Management Review, 2,* 195–220.

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

Kleinmann, M., Exler, C., Kuptsch, C., & Köller, O. (1995). Unabhängigkeit und Beobachtbarkeit von Anforderungs dimensionen im Assessment Center als Moderatoren der Konstruktvalidität [Independence and observability of dimensions as moderators of construct validity in the assessment center]. *Zeitschrift für Arbeits- und Organisationspsychologie, 39,* 22–28.

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 for the validity of job performance ratings. *Human Performance, 3,* 19–35.

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors

represent cross-situational specificity, not method bias. *Human Performance, 13,* 323–353.

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

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

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait–multimethod studies. *Journal of Applied Psychology 86,* 1202–1222.

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 16., pp. 245–286). Chicester, England: Wiley.

Lorenzo, R. V. (1984). Effects of assessorship on managers' proficiency in acquiring, evaluating, and communicating information about people. *Personnel Psychology, 37,* 617–634.

Magnusson, D., & Toerestad, B. (1993). A holistic view of personality: A model revisited. *Annual Review of Psychology, 44,* 427–452.

Maher, P. T. (1990, March). *How many dimensions are enough?* Paper presented at the International Congress on the Assessment Center Method, Orange, CA.

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

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69,* 182–186.

Quinones, M. A., Ford, J. K., & Teachout, M. S. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology, 48,* 887–910.

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43,* 71–84.

Robie, C., Adams, K. A., Osburn, H. G., Morris, M. A., & Etchegaray, J. M. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance, 13,* 355–370.

Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., & McCormick, S. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology, 80,* 664–670.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67,* 401–410.

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., Mayes, B. T., & Riggio, R. E. (1999, May). *A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Schneider, J.R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77,* 32–41.

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

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

Smither, J. W., Barry, S. R., & Reilly, R. R. (1989). An investigation of the validity of expert true score estimates in appraisal research. *Journal of Applied Psychology, 74,* 143–151.

Spychalski, A. C., Quinones, M. A., Gaugler, B. B., & Pohley, K. A. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology, 50,* 71–90.

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

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.

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

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34,* 397–423.

Thornton, G. C., III (1992). *Assessment centers in human resource management.* Reading, MA: Addison-Wesley.

Viswesvaran, C. (1996, April). *Modeling job performance: Is there a general factor?* Paper presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

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

Received May 22, 2001
Revision received October 31, 2001
Accepted November 2, 2001 ■