5-2001

Assessors and use of assessment centre dimensions: A fresh look at a troubling issue

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

Citation

LIEVENS, Filip. Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. (2001). *Journal of Organizational Behavior*. 22, (3), 203-221.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/5590

Published in Journal of Organizational
Behavior, 2001 May, Volume 22, Issue 3,
Pages 203-221

Assessors and use of assessment centre dimensions: a fresh look at a troubling issue

FILIP LIEVENS[*]

Department of Personnel Management and Work and Organizational Psychology, Ghent University, Ghent, Belgium

Summary Previous studies on the construct validity of assessment centres have generally produced puzzling results. The premise of this study is that these prior studies were relatively one-sided. Actually, most previous studies were field studies, which typically used the multitrait–multimethod approach to distinguish between two sources of variance (i.e., exercises and dimensions). Therefore, this study aims to shed light on the issue of assessment centre construct validity by addressing substantive and methodological concerns inherent in previous research. In this study, 85 industrial and organizational psychology students and 39 managers rated videotaped assessment centre candidates in three exercises on six dimensions. Results from generalizability analyses showed that assessors' ratings were relatively veridical. In addition, when assessors rated candidates whose performances varied across dimensions and whose performances were relatively consistent across exercises, they were reasonably able to differentiate among the various dimensions. They also rated such candidate profiles similarly on the various dimensions across exercises. When assessors rated a candidate profile without clear performance fluctuations across dimensions, distinctions about dimensions were more blurred. Results from student and managerial assessors were similar, although managers distinguished somewhat less between the various dimensions. The research and practical implications of these findings are discussed. Copyright © 2001 John Wiley & Sons, Ltd.

Introduction

Over the past forty years assessment centres (ACs hereafter) have established themselves as popular procedures which can serve a variety of human resource functions such as selection and development. It is established in the literature that ACs possess good criterion-related validity (Gaugler *et al.*, 1987) and face validity (Macan *et al.*, 1994). Since the early 1980's, however, questions have been raised whether AC ratings did indeed represent meaningful constructs. Most studies (e.g., Chan, 1996; Fleenor, 1996; Joyce *et al.*, 1994; Robertson *et al.*, 1987; Sackett and Dreher, 1982) used the multitrait–multimethod matrix (Campbell and Fiske, 1959) to analyse the ratings which assessors made upon completion of each exercise. The general conclusion was that, within exercises, the distinctions

[*] Correspondence 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

Parts of this manuscript were presented at the Academy of Management Meeting, San Diego, August 1998.

Received 23 July 2000
Accepted 24 May 2000

between dimensions were blurred, as scores on one dimension in an exercise correlated highly with scores on other dimensions (i.e., low discriminant validity). When people were rated on the same dimension in more than one exercise, there was little correlation among the ratings obtained (i.e., low convergent validity). Similar results were obtained through a more powerful construct validation approach such as confirmatory factor analysis: Within-exercise dimension ratings in ACs primarily represented exercises instead of dimensions (e.g., Bycio *et al.*, 1987). Other studies placed the final dimension ratings in ACs in a nomological net with measures of personality and cognitive ability. Results here were equivocal, as some studies (Shore *et al.*, 1990; Thornton *et al.*, 1997) established most of the expected relationships, whereas other studies (Chan, 1996; Fleenor, 1996) failed to demonstrate the expected relationships.

The lack of clear construct valid dimensions in ACs has considerable practical implications (Klimoski and Brickner, 1987). For instance, (developmental) ACs use the dimensional ratings as starting points to provide feedback to the assessees and to formulate action plans. If the dimensions are no valid indicants of the managerial abilities, the feedback and subsequent action plans could have detrimental effects (Bycio *et al.*, 1987; Fleenor, 1996; Joyce *et al.*, 1994). The following example by Kudisch *et al.* (1997) further exemplifies these practical implications: *Telling a candidate that he or she needs to improve his or her overall leadership skills may be inappropriate if the underlying construct being measured is dealing with a subordinate in a one-on-one situation (i.e., tapping individual leadership as opposed to group leadership)'* (p. 131). In other words, although ACs for selection purposes (i.e., yes/no decisions) may not require construct validity ('they work for some unknown reason'), ACs purported to identify and develop managerial strengths and weaknesses do require construct validity. Therefore, the lack of construct valid measurement of the AC dimensions has been identified as the crux in the AC paradigm (Joyce *et al.*, 1994; Kauffman *et al.*, 1993; Klimoski and Brickner, 1987; Reilly *et al.*, 1990).

Because the issue of the construct validity of ACs is important, it requires further investigation. This study's premise is that prior research in this domain has been relatively one-sided. Actually, most previous studies were field studies, which typically used the multitrait–multimethod approach to distinguish between two sources of variance (i.e., exercises and dimensions). Therefore, this study aims to shed light on the troublesome issue of AC construct validity by addressing both substantive and methodological concerns inherent in prior research.

Concerns in Previous Research

Ambiguous test of rival explanations

Generally, two interpretations have been put forward to explain the lack of convergent validity found in within-exercise dimension ratings of ACs. A first interpretation is that these findings result from biases and inaccuracy of assessors. This 'biased assessors' explanation posits that due to cognitive overload (Gaugler and Thornton, 1989; Reilly *et al.*, 1990) or schema-based processing (Zedeck, 1986) assessors are not able to define and use the dimensions consistently across exercises. In the past this explanation has dominated construct validity research in ACs because various design and procedural interventions have been proposed to help assessors deal with their complex task. Examples include limiting the number of dimensions (Gaugler and Thornton, 1989), using behavioral checklists (Reilly *et al.*, 1990), or increasing the length of assessor training (Dugan, 1988). Some of these design interventions have resulted in significant increases in convergent validity. For instance, the use of behavioral checklists improved convergent validity from 0.24 to 0.43 (Reilly *et al.*, 1990). According to a second interpretation assessors are not to blame for the low convergent validity found. Instead, this finding is simply due to candidates' real performance differences across situations (Neidig and Neidig, 1984). For example, certain individuals may perform better in one-to-one exercises than in group situations, diminishing the convergence of ratings across exercises. This second interpretation is generally referred to as the 'true candidate performances' explanation.

Hence, because the lack of convergent validity may also represent candidates' true performances, it is still unclear whether assessors are indeed unable to provide consistent ratings (Brannick *et al.*, 1989; Harris *et al.*, 1993; Sackett and Dreher, 1982; Turnage and Muchinsky, 1982). Therefore, this study tests the biased assessor interpretation by holding the true candidate performances interpretation constant. More specifically, this study presents assessors with videotaped candidates who perform relatively consistently across exercises (see all candidate profiles of Table 1) and investigates the extent to which assessors will be able to see this. The answer to this question is important. If assessors are unable to give consistent across-exercise ratings, then their ratings are inherently biased. Alternatively, if assessors are able to use the dimensions consistently, this means that their ratings are veridical.

Similar interpretations have been proposed for the lack of discriminant validity found in within-exercise dimension ratings. According to the biased assessors' explanation, the low discriminant validity illustrates assessors' inability to differentiate among the various dimensions. According to the true candidate performances interpretation low discriminant validity results from the fact that candidates often do not exhibit a lot of performance variation on the dimensions. To disentangle these rival explanations related to discriminant validity, this study presents assessors with different types of candidates. On the one hand assessors are asked to rate candidates who exhibit performance fluctuations across dimensions within an AC exercise (see candidate profiles 1, 2, and 3 of Table 1). On the other hand assessors are also asked to rate a candidate whose performance does not vary across dimensions (see candidate profile 4 of Table 1). If these performance manipulations are not present in assessor ratings, then their ratings are inherently biased. Conversely, if assessor

Table 1. Estimated and intended performance profiles of candidates in exercises

	Candidate profiles			
Dimension	Profile 1	Profile 2	Profile 3	Profile 4
		Presentation exercise		
Problem analysis and solving	3.0 (3)	2.6 (1)	4.8 (5)	2.0 (1)
Interpersonal sensitivity	4.8 (5)	2.6 (3)	1.2 (1)	1.0 (1)
Planning and organizing	1.6 (1)	5.0 (5)	3.4 (3)	1.8 (1)
		Role-play exercise		
Problem analysis and solving	3.0 (3)	1.6 (1)	4.6 (5)	1.4 (1)
Interpersonal sensitivity	4.4 (5)	2.0 (3)	1.8 (1)	1.0 (1)
Planning and organizing	2.0 (1)	4.8 (5)	3.4 (3)	2.2 (1)
		Group discussion exercise		
Problem analysis and solving	3.2 (3)	2.0 (1)	4.6 (5)	1.2 (1)
Interpersonal sensitivity	4.8 (5)	2.2 (3)	1.2 (1)	1.2 (1)
Planning and organizing	2.0 (1)	4.8 (5)	4.0 (3)	1.6 (1)

Note. The average of the expert ratings is given. Intended scores are given in parentheses. A score of 1 indicates poor performance and a score of 5 indicates excellent performance.

ratings reflect these performance profiles, their ratings are veridical, refuting the biased assessors interpretation.

Incomplete design

In operational ACs all assessors typically do not rate all assessees in every exercise. Although this practice saves time and costs, it also leads to an important methodological limitation in research (Howard, 1997; Jones, 1992). In fact, in all previous studies assessors were nested within exercises. This means that the ratings studied share two potential sources of variance: They are ratings of different exercises, and the ratings in different exercises are made by different assessors (Robertson *et al.*, 1987; Sackett and Dreher, 1982).

Due to this confounding of exercise and assessor variance, some researchers have suggested using a fully crossed design (Assessors × Dimensions × Exercises) to investigate AC construct validity (Howard, 1997; Jones, 1992; Turnage and Muchinsky, 1982). In this study an AC environment was simulated in which all assessors viewed all assessees in every exercise. This enabled me to clearly separate the variance due to assessors and the variance due to exercises.

Multitrait–multimethod approach

As mentioned above, multitrait–multimethod analyses and confirmatory factor analysis have generally been employed to examine the construct validity of within-exercise dimension ratings in ACs. Unfortunately, previous studies using these approaches have only recognized two sources of variation in ACs, namely exercises and dimensions (Jones, 1992). In ACs, however, the assessors also play a vital role in the measurement process.

Given this drawback Brannick *et al.* (1989) recommended generalizability analysis (Brennan, 1992; Cronbach *et al.*, 1972) as an alternative approach for understanding sources of variance in AC scores (see also McHenry and Schmitt, 1994). Generalizability analysis aims to decompose, in any measurement, the observed variance into components attributable to the underlying attributes (real variance) or components attributable to measurement error (error variance) (Kane, 1982; Kraiger and Teachout, 1990; Marcoulides, 1989). As opposed to classical test theory, generalizability theory regards this measurement error as *multi*faceted. Accordingly, it permits the simultaneous estimation of many different sources of variance inherent in ratings. In this study generalizability analysis is used to provide a more complete partitioning of the sources of variance in AC scores (i.e., candidates, exercises, dimensions, and assessors).

Student samples

Recently, Lievens (1998) reviewed 21 studies, which manipulated specific variables to determine their impact on the quality of construct measurement in ACs. This review indicated that students served as assessors in all of the studies with videotaped assessees as stimulus material. This reliance on student samples limits the external validity of the results obtained because prior research demonstrated that students and managers differ substantially in making selection decisions (Barr and Hitt, 1986). Therefore, in this study both students and managers serve as assessors.

Method

Sample

Two samples were used. The first sample consisted of 85 industrial and organizational psychology students who participated in the study to receive credit for a human resource management course. The sample included 53 women and 32 men with a mean age of 22.6 years (*SD* = 2.7 years). Ages ranged from 20 to 35 years. Twenty-nine per cent of the subjects were seniors, 71 per cent had been in college for more than four years and had already worked as interns in a psychological consulting firm or in a company's personnel department.

The second sample was composed of 39 managers (30 men, 9 women). The average age of the managers was 34.6 years old (*SD* = 4.5 years). Ages ranged from 27 to 48 years. The managers had an average of 10.8 years full-time working experience (*SD* = 4.3 years, range = 4–23 years). All managers were enrolled in an executive MBA programme. The managers came from a broad variety of organizations and had different functional backgrounds (e.g., engineering, sales, etc.).

Assessment centre simulation

Funder (1987) advocated incorporating into laboratory research representations of real life by faithfully reconstructing all of the important elements and sources of information that actually are found in a particular real-life situation. Consistent with Funder's (1987) suggestions, a major thrust of this study consisted in simulating both rating task and rating context of assessors. On average, this whole AC simulation ran seven hours.

Firstly, students and managers were familiarized with the purpose of the whole simulation. Specifically, the participants were told to assume the role of assessors in an AC simulation and to evaluate four videotaped candidates applying for the job of district sales manager. This provided them with an opportunity to further acquire practical experience in observing and rating candidates. Assessors knew that afterwards they had to explain their observations and ratings to one another. This practice guaranteed that they took their task seriously.

Next, participants received an assessor training. This training started with an introductory lecture, which covered the origin, the basics, and the rating process of ACs. After the lecture, the assessors received details about the main tasks and qualifications required for successful district sales managers, the working context (e.g., place in organizational tree, number of subordinates, etc.), and the organization (e.g., the type of business, the size, pictures of products made, the organizational culture, etc.). This information originated from real materials (i.e., actual job posting and annual report of an organization). Besides this organization and job information, the assessor training also included presentation of AC dimensions, AC exercises, and rating scales. In particular, assessors received a list of 12 performance dimensions, which a job analysis had identified to be crucial for effective district sales managers. According to the job analysis the three most important performance dimensions were: (a) problem analysis and solving, (b) interpersonal sensitivity, and (c) planning and organizing. Descriptions of these dimensions were given. Assessors were told that three exercises were chosen because a job analysis had determined them to be relevant for the target job: (a) a sales presentation, (b) a role-play with a disgruntled employee, and (c) an assigned-role group discussion. These exercises were reviewed (e.g., general purpose, context, role-player task, etc.). Additionally, assessors were informed that the four candidates for the position would be presented on videotape and that videotaping assessees was relatively widespread in ACs (Ryan *et al.*, 1995).

Immediately following the assessor training, assessors were randomly assigned to small teams, which were placed in separate rooms. Next, assessors observed the videotaped performance of the first candidate in the sales presentation exercise, recorded observations, and provided independently dimensional ratings. This process was repeated for the presentation performance of the other three candidates, for the role-play performances of each of the four candidates, and for the group discussion, in which the four candidates performed together. To control for order effects, I developed four versions of the integral film for which I varied the candidate order. The order of the exercises was the same in all four versions. The assessor teams were randomly assigned to a particular version of the film. An equal number of assessor teams viewed each version. Irrespective of the videotaped version, all assessors rated all candidates in every exercise. After observing and rating candidates, assessors met in their respective teams to share observations, discuss ratings, and write assessee reports.

In sum, this study's simulated assessor environment converged closely to current AC practices in organizations (Spychalski *et al.*, 1997) and to previous AC simulations (see Gaugler and Rudolph, 1992, for an example). Furthermore, my simulation met virtually all of the ten essential elements of an AC delineated by the Guidelines and Ethical Considerations for AC Operations (Task Force on Assessment Center Guidelines, 1989): (a) a thorough job analysis was conducted to determine relevant dimensions and exercises, (b) assessors classified behavioral observations into meaningful and relevant categories, (c) the exercises were designed to elicit information for evaluating dimensions, (d) multiple assessment techniques were used and pretested in real organizations, (e) a sufficient number of job-related exercises were used to allow multiple opportunities to observe dimension-related behavior, (f) multiple assessors observed and rated each assessee, (g) assessors received thorough training, which followed the majority of recommendations of the Guidelines and Ethical Considerations for Assessment Center Operations (Task Force on Assessment Center Guidelines, 1989), (h) assessors systematically recorded specific behavioral statements at the time of their occurrence, (i) assessors had some report of the observations made in each exercise in preparation for the integration discussion with the other assessors, and (j) the integration of behaviors was based on a pooling of information from assessors. The only exception was that assessors were not systematically evaluated at the end of the training. However, this is also seldom done in operational ACs (Spychalski *et al.*, 1997).

Videotaped assessee performances

Underlying performance profiles

The candidate performances were designed to vary along three dimensions: problem analysis and solving, interpersonal sensitivity, and planning and organizing. In addition, as already noted, the performance profile of each candidate was designed to be relatively consistent across the exercises. In addition, three of the candidate profiles varied across the dimensions within exercises and one candidate profile (candidate profile 4) did not vary across the dimensions. The parenthetical values of Table 1 present the intended performance profiles for each of the four candidates.

Check of external validity of performance profiles

An important question was whether the four candidate profiles selected were based on realistic assumptions about AC performance. In fact, the four profiles selected were only a sample from the total set of 27 candidate profiles (given that there were three performance levels for each of the three dimensions). Therefore, 42 professional assessors (25 men, mean age = 34 years; mean assessor experience = 5 years) were asked to rate the realism of all 27 profiles on a 9-point scale (1 = *not at all realistic*, 9 = *very realistic*). Each assessor determined the realism of five randomly chosen profiles, which were presented in written form. On average, each profile was rated by eight assessors. The assessors found five (19 per cent) candidate profiles to be not realistic. The mean realism ratings for the four candidate profiles selected in this study were 8.00 (profile 1), 7.50 (profile 3), 7.43 (profile 4), and 5.45 (profile 2). Hence, the candidate profiles used in this study were considered to be realistic.

Development of videotapes

Firstly, a representative pool of assessee behaviors for each dimension in each of the three AC exercises was gathered. On the one hand, 20 professional assessors (15 men, mean age = 36 years) were asked to provide behaviors that cause them to judge an assessee as being higher or lower on a specific dimension. These assessors qualified as experts due to (a) their practical experience as assessors (mean assessor experience = 6 years), (b) their theoretical knowledge of ACs, and (c) their familiarity with AC research. They generated 765 dimension-specific behaviors across the three exercises. On the other hand, rating forms of five psychological consulting firms were scrutinized. These rating forms yielded 121 behaviors. After eliminating redundancies, I reduced the total list of 886 behaviors to a list of 310 behaviors.

Secondly, scripts of each candidate's performance in the three exercises were written. These scripts were based on the critical candidate behaviors gathered and on the performance profiles. The scripts depicted the word-for-word dialogue for each performance. Nine scripts were written: four scripts of a candidate delivering a sales presentation, four scripts of the same four candidates talking to a disgruntled employee (role-player), and one script of a group discussion between these four candidates. Two experienced assessors (two women, mean age = 33.5 years, mean assessor experience = 4 years) tested the scripts for realism and made adjustments.

Thirdly, semi-professional actors were filmed delivering their scripted AC performances. The videotaped performances ran between 5 min (role-play) to 14 min (group discussion). The total length of the whole set of videotapes was approximately 1 hour.

Finally, I followed procedures by Sulsky and Balzer (1988) to verify whether the videotaped performances reflected the scores built into the scripted performances. In particular, five professional assessors (3 men, mean age = 30 years; mean assessor experience = 4 years) viewed each videotaped performance under optimal conditions. This meant, for instance, that they could view the tape repeatedly and rewind it. All experts independently rated each performance on a 5-point scale, with 1 indicating *poor* and 5 indicating *excellent*. The average of these expert ratings correlated highly ($r = 0.94$) with the intended scores, demonstrating that the videotaped performances carefully reflect the intended scores. Table 1 presents the average of the expert scores for the various candidate performances.

Measures

Participants in the AC simulation completed an observation form and a rating form for each videotaped performance of the four candidates to evaluate six dimensions.[1] Three of these dimensions (i.e., problem analysis and solving, interpersonal sensitivity, and planning and organizing) were standard in every exercise. The other dimensions were specific for a particular exercise. Oral communication, for instance, was included only in the rating form of the sales presentation. The dimensions were rated on a 5-point scale, ranging from *poor* (1) to *excellent* (5).

[1] As already mentioned, the videotaped assessee performances were built around three dimensions. Yet, I asked assessors to rate six dimensions. Hence, assessors also had to rate dimensions, which were not *a priori* built into the videotapes. This was done to enhance the generalizability of the assessor task. In operational ACs assessors usually evaluate assessees on more than three dimensions. In addition, in operational ACs exercises typically vary in the opportunity for behavior representing a dimension to be manifested (Reilly *et al.*, 1990). So, it is not unusual for assessors to rate candidates on dimensions, which are less observable.

Table 2. Results of generalizability study: full design

	Students				Managers			
Effect	*df*	*VC*	90% Confidence intervals	Explained variance (%)	*df*	*VC*	90% Confidence intervals	Explained variance (%)
A(ssessors)	82	0.00	$0.00 < VC < 0.01$	0	37	0.0*		
E(xercises)	2	0.0*			2	0.01	$0.00 < VC < 0.02$	0.4
C(andidates)	3	0.08	$0.04 < VC < 0.24$	6.1	3	0.31	$0.16 < VC < 0.91$	17.3
D(imensions)	2	0.0*			2	0.0*		
AE	164	0.0*			74	0.03	$0.02 < VC < 0.09$	1.7
AC	246	0.09	$0.06 < VC < 0.14$	6.4	111	0.16	$0.10 < VC < 0.28$	8.7
AD	164	0.0*			74	0.0*		
CE	6	0.01	$0.01 < VC < 0.04$	1.2	6	0.03	$0.01 < VC < 0.08$	1.6
ED	4	0.02	$0.01 < VC < 0.05$	1.2	4	0.01	$0.01 < VC < 0.04$	0.7
CD	6	0.41	$0.21 < VC < 1.21$	30.9	6	0.34	$0.18 < VC < 1.01$	19.1
AEC	492	0.17	$0.14 < VC < 0.21$	13.0	222	0.26	$0.21 < VC < 0.34$	14.6
AED	328	0.01	$0.00 < VC < 0.02$	0.5	148	0.02	$0.01 < VC < 0.07$	1.3
ACD	492	0.10	$0.08 < VC < 0.14$	7.6	222	0.15	$0.11 < VC < 0.21$	8.1
ECD	12	0.02	$0.01 < VC < 0.06$	1.7	12	0.03	$0.01 < VC < 0.08$	1.4
AECD	984	0.42	$0.39 < VC < 0.45$	31.3	444	0.45	$0.23 < VC < 1.32$	25.1

Note. Due to missing data the generalizability analyses of the full design were based on a student sample of 83 assessors and a managerial sample of 38 assessors respectively; VC = estimated variance components. *Consistent with recommendations of Shavelson and Webb (1991), small negative estimates of variance components were reported as zero.

Analyses and Results

Descriptive statistics

Means and standard deviations of the ratings made by the two assessor samples are available from the author. Generally, managerial assessors gave lower ratings than student assessors. This was particularly true for candidate profile 4, with managers evaluating this candidate significantly less favorably.

Generalizability analysis of full design

As already noted, generalizability analysis decomposes an observed score into a component for the universe score and error components affecting the measurement process (Marcoulides, 1989). Accordingly, prior to each generalizability analysis the researcher typically specifies the factors affecting the measurement process. In generalizability theory these factors are referred to as facets. Applied to the present assessment centre study, the generalizability analysis of the full design had three facets: assessors (A), exercises (E), and dimensions (D). These three facets were completely crossed[2] with each other. Candidates (C) were treated as the object of measurement (i.e., universe score).

In this study GENOVA (version 2.2), which is a Fortran-based program specifically developed for generalizability analyses (Crick and Brennan, 1983), was used. The results of the generalizability analysis of both samples are presented in Table 2. Included are the degrees of freedom, the estimated variance components, and their 90 per cent confidence intervals. These confidence intervals were computed by procedures outlined in Brennan (1992). Variance components reflect each facet's contribution to the total variance. For example, in this generalizability analysis (full design) the variance components represent the variances of the mean candidate ratings attributable to the candidates, to the assessors, to the dimensions, to the exercises, 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 and Webb, 1990). To this end, I used the per cent contribution of each variance component. This percentage contribution, which is also given in Table 2, refers to the percentage of the sum of the variance components (i.e., the total variance) that each variance component accounts for.

Some variance components estimated are especially relevant in light of this study's purpose. Specifically, evidence of convergent validity is derived from the variance component of the Candidates × Exercises interaction (Kane, 1982; Kraiger and Teachout, 1990). A low value of this Candidates × Exercises variance component suggests invariance of candidate ratings across exercises. As shown in Table 2, the Candidates × Exercises (CE) interaction made only minor contributions to the total variance (1.2 per cent in the student sample and 1.6 per cent in the managerial sample). In addition, the variance component due to the Exercises × Candidates × Dimensions (ECD) interaction was negligible. In short, these results show no substantial variation in ratings of candidates across exercises. Or to put it differently, when assessors rated candidates whose performances were designed to be relatively consistent across exercises, evidence of convergent validity was established.

Evidence of discriminant validity is derived from the variance component associated with the Candidates × Dimension interaction. A high value of this variance component indicates substantial

[2] Because the generalizability analysis required a balanced design, the analysis included only those three dimensions which assessors had to rate in all three exercises.

Table 3. Generalizability study variance components within dimensions

	Students				Managers			
Effect	*df*	*VC*	90% Confidence intervals	Explained variance (%)	*df*	*VC*	90% Confidence intervals	Explained variance (%)
				Problem analysis and solving				
A(ssessors)	83	0.02	0.01 < VC < 0.06	1.5	37	0.0*		
E(xercises)	2	0.04	0.02 < VC < 0.11	2.7	2	0.06	0.03 < VC < 0.18	3.4
C(andidates)	3	0.40	0.21 < VC < 1.17	28.9	3	0.63	0.32 < VC < 1.84	34.5
AE	166	0.0*			74	0.01	0.01 < VC < 0.03	0.6
AC	249	0.11	0.07 < VC < 0.21	8.2	111	0.16	0.09 < VC < 0.38	8.7
CE	6	0.05	0.03 < VC < 0.14	3.6	6	0.07	0.03 < VC < 0.19	3.6
AEC	498	0.76	0.69 < VC < 0.85	55.2	222	0.90	0.77 < VC < 1.05	49.3
				Interpersonal sensitivity				
A(ssessors)	84	0.0*			37	0.0*		
E(xercises)	2	0.01	0.01 < VC < 0.03	0.6	2	0.01	0.00 < VC < 0.02	0.3
C(andidates)	3	0.78	0.40 < VC < 2.27	46.4	3	0.95	0.49 < VC < 2.79	46.5
AE	168	0	0.00 < VC < 0.00	0.2	74	0.06	0.03 < VC < 0.18	2.9
AC	252	0.33	0.23 < VC < 0.49	19.5	111	0.44	0.33 < VC < 0.62	21.4
CE	6	0	0.00 < VC < 0.01	0.1	6	0.05	0.02 < VC < 0.14	2.3
AEC	504	0.55	0.50 < VC < 0.62	33.2	222	0.55	0.47 < VC < 0.64	26.6
				Planning and organizing				
A(ssessors)	83	0.01	0.00 < VC.02	0.7	37	0.0*		
E(xercises)	2	0.0*			2	0.0*		
C(andidates)	3	0.32	0.17 < VC < 0.95	32.6	3	0.38	0.20 < VC < 1.12	25.1
AE	166	0.0*			74	0.09	0.05 < VC < 0.26	5.9
AC	249	0.12	0.09 < VC < 0.18	12.4	111	0.31	0.22 < VC < 0.49	20.2
CE	6	0.07	0.03 < VC < 0.20	6.8	6	0.05	0.03 < VC < 0.14	3.2
AEC	498	0.47	0.43 < VC < 0.53	47.5	222	0.70	0.60 < VC < 0.82	45.5

*Small negative estimates of variance components were reported as zero (Shavelson and Webb, 1991).

differences in candidate ratings across dimensions (Kane, 1982; Kraiger and Teachout, 1990). Table 2 shows that in the student sample besides the residual term (AECD), which contains both random error and error due to the four-way interaction, the most variance was attributed to this Candidates × Dimensions (CD) interaction (30.9 per cent), indicating differences between dimensions in evaluations of candidates. In other words, when assessors rated candidates designed to vary in their relative performance qualities across dimensions, evidence of discriminant validity was found. It is striking that in the managerial sample the Candidates × Dimensions (CD) interaction accounted for 19.1 per cent of the variance, revealing that managerial assessors differentiated somewhat less between the dimensions than student assessors.

A last variance component of interest to this study is the variance component of the Assessors × Candidates (AC) interaction. A low value of this variance component suggests little variation in candidate ratings across assessors and, therefore, is indicative of inter-rater reliability. Table 2 shows that in the student sample 6.4 per cent of the variance was attributed to this interaction effect. In the managerial sample this value mounted to 8.7 per cent. In other words, inconsistencies due to assessors accounted for some error variance. Other variance components, namely those associated with the Assessors × Exercises × Candidates interaction and the Assessors × Candidates × Dimensions interaction, also suggest that inconsistencies among assessors exist, both for rating exercises and for rating dimensions.

Other variance components, though informative, are not of interest to this study. For example, the negligible variance component associated with the assessor main effect in Table 2 illustrates that, averaging over candidates, exercises, and dimensions, ratings of assessors do not differ from each other. In other words, some assessors do not give higher (more lenient) or lower (more stringent) ratings than other assessors. Similarly, the negligible variance component associated with the dimension main effect shows that, averaging over exercises, candidates, and assessors, one dimension does not receive higher or lower ratings than another dimension (something which is also illustrated by Table 1). Otherwise, the moderate candidate main effect suggests that, averaging over assessors, exercises, and dimensions, ratings of candidates differ somewhat from each other. Indeed, as illustrated by Table 1, candidate 4 should normally be rated lower than the other candidates.

Within dimension generalizability analysis

The relatively large variance component for the Candidates × Dimensions (CD) interaction in the previous generalizability analysis (full design) revealed differences between dimensions in assessor evaluations of candidates. In line with recommendations of Shavelson and Webb (1991), I conducted separate generalizability analyses within each dimension. Because these analyses were conducted within each dimension, there were only two completely crossed facets: assessors (A) and exercises (E). Candidates (C) were again the object of measurement. The added informative value of these analyses is that they may point out whether the results of the generalizability analysis of the full design are also found for each of the dimensions studied.

Table 3 presents the results for both samples. Below I concentrate only on the variance components of interest to this study. Generally, the findings confirm the results of the full design. First, for each of the three dimensions, exercises (E) were again a minor source of error variation. This provides further support for convergent validity. For the dimension problem analysis and solving, however, exercises (E) accounted for some variance (2.7 per cent in the student sample and 3.4 per cent in the managerial sample), indicating that ratings of this dimension converged somewhat less across exercises. Second, for each of the three dimensions, the two largest variance components were those associated with the residual term (AEC) and with candidates (C). This latter variance component was not considered a

Table 4. Generalizability study variance components within profiles

	Students				Managers			
Effect	*df*	*VC*	90% Confidence intervals	Explained variance (%)	*df*	*VC*	90% Confidence intervals	Explained variance (%)
				Candidate profile 1				
A(ssessors)	84	0.04	$0.02 < VC < 0.12$	3.6	38	0.20	$0.11 < VC < 0.51$	13.8
E(xercises)	2	0.01	$0.01 < VC < 0.04$	1.0	2	0.05	$0.03 < VC < 0.16$	3.9
D(imensions)	2	0.29	$0.15 < VC < 0.84$	25.0	2	0.28	$0.14 < VC < 0.82$	19.7
AE	168	0.18	$0.13 < VC < 0.26$	15.3	76	0.34	$0.24 < VC < 0.53$	24.1
AD	168	0.18	$0.13 < VC < 0.26$	15.6	76	0.07	$0.03 < VC < 0.19$	4.7
ED	4	0.04	$0.02 < VC < 0.10$	3.1	4	0.01	$0.01 < VC < 0.04$	1.0
AED	336	0.42	$0.37 < VC < 0.48$	36.5	152	0.46	$0.39 < VC < 0.57$	32.7
				Candidate profile 2				
A(ssessors)	82	0.04	$0.02 < VC < 0.13$	4.0	37	0.09	$0.05 < VC < 0.26$	6.2
E(xercises)	2	0.00	$0.00 < VC < 0.00$	0.1	2	0.06	$0.03 < VC < 0.17$	4.1
D(imensions)	2	0.37	$0.19 < VC < 1.08$	34.0	2	0.29	$0.15 < VC < 0.84$	20.1
AE	164	0.19	$0.14 < VC < 0.27$	17.4	74	0.39	$0.27 < VC < 0.59$	26.9
AD	164	0.05	$0.03 < VC < 0.15$	4.8	74	0.16	$0.10 < VC < 0.31$	11.1
ED	4	0.01	$0.00 < VC < 0.03$	0.8	4	0.01	$0.01 < VC < 0.03$	0.7
AED	328	0.42	$0.37 < VC < 0.48$	38.8	148	0.44	$0.37 < VC < 0.54$	31.0
				Candidate profile 3				
A(ssessors)	84	0.07	$0.04 < VC < 0.20$	4.5	38	0.11	$0.06 < VC < 0.32$	6.2
E(xercises)	2	0.03	$0.01 < VC < 0.08$	1.8	2	0.03	$0.02 < VC < 0.09$	1.8
D(imensions)	2	0.68	$0.35 < VC < 1.99$	43.8	2	0.54	$0.28 < VC < 1.58$	30.7
AE	168	0.17	$0.13 < VC < 0.25$	11.1	76	0.30	$0.21 < VC < 0.49$	17.2
AD	168	0.14	$0.10 < VC < 0.22$	9.2	76	0.17	$0.11 < VC < 0.34$	9.9
ED	4	0.03	$0.01 < VC < 0.08$	1.8	4	0.08	$0.04 < VC < 0.23$	4.4
AED	336	0.43	$0.38 < VC < 0.49$	27.8	152	0.52	$0.44 < VC < 0.64$	29.7
				Candidate profile 4				
A(ssessors)	84	0.21	$0.14 < VC < 0.32$	21.0	38	0.15	$0.09 < VC < 0.33$	16.6
E(xercises)	2	0.00*			2	0.0*		
D(imensions)	2	0.09	$0.05 < VC < 0.27$	9.3	2	0.10	$0.05 < VC < 0.28$	10.9
AE	168	0.13	$0.09 < VC < 0.21$	13.7	76	0.13	$0.08 < VC < 0.27$	15.1
AD	168	0.02	$0.01 < VC < 0.07$	2.5	76	0.02	$0.01 < VC < 0.07$	2.7
ED	4	0.09	$0.05 < VC < 0.26$	9.1	4	0.05	$0.03 < VC < 0.15$	5.7
AED	336	0.44	$0.39 < VC < 0.50$	44.4	152	0.43	$0.36 < VC < 0.53$	48.9

*Small negative estimates of variance components were reported as zero (Shavelson and Webb, 1991).

source of error variation because it represented desirable universe score variance (i.e., candidates differ in terms of their performance) (Brennan, 1992; Marcoulides, 1989). For the dimension interpersonal sensitivity the variance component due to candidates was the highest (46.4 per cent in student sample and 46.5 per cent in managerial sample). This implied that ratings of candidates varied most on this dimension. Third, it was striking that, for each of the three dimensions, the Assessors × Candidates (AC) interaction was always the third largest variance component. In the student sample this Assessors × Candidates (AC) interaction accounted for 8.2 per cent, 19.5 per cent, and 12.4 per cent of the total variance of problem analysis and solving, interpersonal sensitivity, and planning and organizing respectively. In the managerial sample the percentages for this Assessors × Candidates (AC) interaction mounted to 8.7 per cent for problem analysis and solving, 21.4 per cent for interpersonal sensitivity, and 20.2 per cent for planning and organizing. All of this indicates variation in candidate ratings across assessors and confirms that, for both student and managerial assessors, inter-rater reliability was only moderate.

To investigate this further I computed generalizability coefficients (ρ^2) per dimension. A generalizability coefficient is an intraclass correlation coefficient similar in form to the classical reliability coefficient. It is defined as the ratio of the universe score variance to the expected observed score variance (Brennan, 1992). In this study the generalizability coefficient reflects generalizability of results generalizing over random samples of different number of assessors and exercises. In general, a generalizability coefficient equal or higher than 0.80 is considered to be acceptable. For the dimension interpersonal sensitivity three exercises and three assessors gave an estimated generalizability level of 0.82 in the student sample. The estimated generalizability coefficients for problem analysis and solving ($\rho^2 = 0.74$), and planning and organizing ($\rho^2 = 0.74$) were somewhat lower than the conventionally acceptable level of 0.80. In the managerial sample estimated generalizability coefficients equalled 0.81 for interpersonal sensitivity, 0.78 for problem analysis and solving, and 0.66 for planning and organizing.

It is also possible to examine the effects of varying the number of conditions of each facet on the generalizability coefficient estimated. For instance, when the number of exercises was reduced from three to one, the estimated generalizability level of the dimension sensitivity dropped only slightly from 0.82 to 0.72 (student sample) and from 0.81 to 0.72 (managerial sample). However, reducing the number of assessors from three to one had a more serious impact on the generalizability coefficient as it dropped from 0.82 to 0.60 (student sample) and from 0.81 to 0.60 (managerial sample). For the dimensions problem analysis and solving, and planning and organizing similar trends were found. Hence, the number of assessors had a larger effect on generalizability than the number of exercises.

Within candidate generalizability analysis

Separate generalizability analyses were also conducted within each candidate profile. In these generalizability analyses the dimensions served as the object of measurement (see Cardinet *et al.*, 1976). Assessors and exercises were the two completely crossed facets. As noted previously, these analyses were conducted to disentangle the rival explanations related to discriminant validity. In particular, these analyses enabled to compare ratings of candidates, who exhibited performance fluctuations across dimensions (see candidate profiles 1, 2, and 3 in Table 1) to ratings of candidates whose performance did not vary across dimensions (see candidate profile 4 in Table 1) in terms of discriminant validity. If assessors provide unbiased ratings of the candidates, then discriminant validity evidence should be established for the first three candidates but not for the fourth candidate profile.

As indicated by Table 4, the results of the separate generalizability analyses within each of the first three profiles were similar to previous analyses: Exercises (E) merely contributed to the total variance and the variance component of dimensions (D) was substantial, suggesting evidence of convergent and discriminant validity. The generalizability analysis within the fourth profile yielded different results because the variance component of dimensions (D) was not impressive. In fact, this variance component of dimensions for the fourth profile (9.3 per cent in student sample) was much lower than the variance component of dimensions for the first three profiles (respectively 25 per cent, 34 per cent, and 43.8 per cent in student sample).

Finally, Table 4 reveals meaningful differences between the student sample and the managerial sample. In fact, the desirable variance due to dimensions (D) was smaller in the managerial sample than in the student sample. This was true for candidate profile 1 (25 per cent in student sample versus 19.7 per cent in managerial sample), for profile 2 (34 per cent versus 20.1 per cent), and for profile 3 (43.8 per cent versus 30.7 per cent).

Discussion

Main conclusions

Several contrasts made this study different from previous research on AC construct validity. For instance, I held the true candidate performances explanation constant to enable a test of the biased assessor explanation. Next, a fully crossed design was used. Generalizability analyses were then conducted to provide a more complete partitioning of the variance. Finally, to address external validity issues both students and managers served as assessors.

The results support three conclusions. A first conclusion is that assessors' ratings of candidates are relatively veridical. This is indicated by the large size of variance components, which represent true variance built into the videotaped performances (i.e., variance due to candidates and dimensions). These results are not supportive of the biased assessors thesis. Nevertheless, the various generalizability analyses also reveal some undesirable bias in assessor ratings. In fact, moderate values for the Assessors × Candidates interaction term are found, suggesting only moderate inter-rater reliability among assessors. Other findings also point in this direction. Only the generalizability coefficient for the dimension sensitivity is higher than the acceptable level, and the number of assessors has a larger effect on generalizability than the number of exercises. There are at least three possible explanations for these results. First, industrial and organizational psychology students and managers served as assessors. Although, most of the students had already worked as interns in psychology consulting firms or in personnel departments, significant lower values for the Assessors × Candidates interaction may be expected for more experienced and professional assessors. Second, in this study assessors provided ratings upon completion of each exercise. Somewhat lower reliability values are generally reported for these so-called within-exercise dimension ratings (Thornton, 1992, p. 129). Third, the training provided to assessors was primarily information-oriented. I expect higher inter-rater reliability values for a more comprehensive and practice-oriented training (Lievens, in press). Regardless which of these explanations is correct, the moderate inter-rater reliability found in this study illustrates that the common AC practice of rating candidates across exercises by different assessors may contribute to the exercise effects reported in construct validity research in operational ACs.

As a second conclusion this study demonstrates that assessors are reasonably able to provide differentiated within-exercise ratings (discriminant validity) and similar across-exercise ratings on dimensions (convergent validity) when they evaluate candidates whose performances vary across dimensions and whose performances are relatively stable across exercises. The ability of assessors to provide relatively differentiated ratings on dimensions is indicated by the large contribution of the Candidates × Dimension interaction in the generalizability analysis of the full design. As shown by the generalizability analyses within each dimension, ratings of assessees vary most on the dimension interpersonal sensitivity. This result is not surprising because interpersonal sensitivity is considered to be a well observable construct, providing assessors with plenty of opportunities to make fine-grained ratings. The ability of assessors to rate the candidate profiles similarly across exercises is supported by the negligible contributions of the Exercises × Candidates interaction to the total variance in the generalizability analysis of the full design. Basically, this means that there does not exist much variation in ratings of candidates across the different exercises. The separate generalizability analyses within each dimension and within the first three candidate profiles also reveal that exercise effects are virtually absent. Only for the dimension problem analysis and solving a small exercise effect in both student and managerial sample is noted.

An important question is how these findings relate to the rather disappointing results of previous studies on AC construct validity. Probably, the contrasting results are due to a combination of the following explanations. On the one hand, the design of this AC builds on virtually all of the recommendations for maximizing the quality of construct measurement in ACs (see Fleenor, 1996; Lievens, 1998; Schneider and Schmitt, 1992, for reviews). It is possible that some important design considerations were not implemented in previous studies, decreasing construct validity evidence. In fact, the present study confirms the importance of design factors such as type of assessor because managerial assessors discriminated somewhat less between dimensions than psychology student assessors. On the other hand, the diverging results between this study and previous studies may not be due to design factors alone. In this study assessors rate candidates whose performance levels are relatively consistent across exercises and relatively different across dimensions. It is possible that the majority of assessees of previous studies exhibited other performance levels, resulting in exercise effects and lack of AC construct validity. The results of the separate generalizability analyses within candidate profiles lend at least some support to this possibility. These separate generalizability analyses demonstrate that evidence of discriminant validity varies according to the candidate profile rated. After all, evidence of distinct ratings across dimensions is demonstrated for the first three candidate profiles but not for the fourth candidate profile. The only difference between this fourth candidate and the others is that this candidate performs rather similarly across dimensions. These results show that to establish evidence of differentiated constructs in ACs the nature of assessee performances may be a limiting factor, which has been overlooked in previous studies. However, all of this should not be interpreted as if in operational ACs weak evidence of construct validity was found because real candidates typically resembled candidate profile 4. Instead, all of this should be interpreted in the sense that the weak evidence of construct validity found in operational ACs is determined by a host of factors such as AC design, assessor training, and candidates' performance levels.

A third conclusion from the results is that questions should be raised with respect to the use of managers as the only assessors in ACs. Past research revealed that criterion-related validities were lower when both managers and psychologists served as assessors than for professional psychologists alone (Gaugler *et al.*, 1987) and that managers had more difficulties in using AC constructs differentially than professional psychologists (Sagie and Magnezy, 1997). The results of this study extend these previous findings to differences between managers and industrial and organizational psychology students, because managers distinguish somewhat less between the various dimensions than industrial and organizational psychology students. This is indicated by the lower percentage of variance due to the Candidates × Dimension interaction in the managerial sample (19.1 per cent) than in the student sample (30.9 per cent). In addition, the separate generalizability analyses within each candidate profile indicate that managerial ratings were more subject to different sources of error. Perhaps managerial assessors make more holistic ratings because they were less motivated or because they tend to use fewer factors in making selection decisions than students (Barr and Hitt, 1986). Otherwise, it is also possible that on the basis of their managerial experience they implicitly use a management behavior schema for evaluating candidate performances (Cardy *et al.*, 1987; Zedeck, 1986). With respect to the latter, Zedeck (1986) suggests that AC exercises, which are basically replicas of managerial situations, actualize schemata of managerial behavior (e.g., the ideal way of organizing a meeting in a group discussion exercise) from managerial assessors. The exercise-specific nature of these schemata may then produce relatively high correlations between different dimension ratings within the same exercise.

Limitations

This study has several limitations. First, some may argue that the results are due to the use of generalizability analysis. Hence, for comparison reasons with previous studies I also applied the multitrait–multimethod approach (Campbell and Fiske, 1959) to the data. In this multitrait-multimethod analysis the three dimensions served as traits and the three exercises as methods.[3] Results showed that in the student sample the mean monotrait-heteromethod correlation (indicative of convergent validity) was 0.46. In the managerial sample this mean correlation was very similar because it equalled 0.47. The mean heterotrait-monomethod correlation in the student sample was 0.28, providing evidence of discriminant validity. Alternatively, this mean heterotrait-monomethod correlation was higher (0.45) in the managerial sample, indicating somewhat less evidence of discriminant validity for the manager assessors. In short, these findings from multitrait–multimethod analyses closely parallel the results of the generalizability analyses.

Second, the external validity of the study is also an important issue. The external validity of research results depends on various methodological aspects such as the sample, the research setting, and the stimuli used. Regarding the sample, strictly speaking the results only generalize to student and managerial assessors. Yet, if psychology students are reasonably able to use the dimensions differentially, it seems reasonable to assume that this will also be the case for professional psychologists. With respect to the research setting, I undertook substantial efforts to simulate an assessor environment, which included most of the Guidelines and Ethical Considerations for AC operations (Task Force on Assessment Center Guidelines, 1989). A last series of efforts (i.e., help of experienced assessors, use of common AC dimensions and exercises) ensured that the videotaped candidate performances contained the kind of stimuli assessors might encounter in actual ACs.

Implications for practice

On the basis of this study I am able to specify several guidelines which may improve AC practice in general and the quality of construct measurement in particular. Practitioners should realize that when a detailed report of a participant's strengths and weaknesses is at stake (as is the case in developmental assessment centers), managerial assessors have more difficulty in providing distinct ratings on the dimensions. Hence, I suggest that psychologists play a key role in assessor teams of assessment centres conducted for developmental purposes. Psychologists may then serve as coach of managerial assessors or as chair of the discussion session.

[3] Given the design characteristics of this study multitrait–multimethod correlations are less appropriate for two reasons. First, because in this study a large group of assessors rated a small number of candidates (i.e., four), correlations are based on only four candidates. Second, because of the fully crossed design assessors rated four candidates per exercise instead of one candidate per exercise. To sidestep this, the multitrait–multimethod correlations reported are computed by only using the ratings of one single candidate randomly selected for each assessor.

Besides the composition of assessor teams, assessment centre users and designers should also pay more attention to the number of assessors observing and rating assessees in an exercise. In this study the number of assessors had a greater impact on reliability than the number of exercises. Practically speaking, users have to weigh the investment of time and resources in the development of AC exercises, against using more assessors. This study's results do not suggest that a small number of exercises should be preferred. However, my findings do suggest that practitioners may get more out of their investment by adding more assessors.

Implications for future research

The present study represents a step in deciphering the enigma surrounding the internal workings of ACs. However, many questions remain to be answered. A first issue is whether it is possible to improve on the results obtained in this study. In particular, it may be interesting to look for procedural interventions, which may increase the extent to which candidates are differentially rated on dimensions (Candidates × Dimensions interaction). In the student sample this interaction contributed 30.9 per cent to the total variance, in the managerial sample 19.1 per cent. On the basis of previous research this percentage of variance due to the Candidates × Dimensions interaction may increase, if professional assessors (Sagie and Magnezy, 1997) and observational checklists are used (Reilly *et al.*, 1990).

Second, future studies may try to answer why managerial assessors had more difficulty distinguishing between the various dimensions. With this respect, the implicit theories which managers hold about performances in leaderless group discussions, role-plays, or presentations are an important and overlooked area of investigation. Specifically, studies are needed in the AC domain to determine whether managers are indeed using different management behavior schemata per exercise (see Zedeck, 1986), how these performance schemata are related to each other, and how they affect criterion-related and construct validity. In addition, it is worthwhile to ascertain whether training is a viable strategy to alter these performance schemata. To this end, frame-of-reference training (Bernardin and Buckley, 1981) may be fruitfully used. In performance appraisal this specific rater training has emerged as the method of choice to impose the same evaluative standards to raters as a reference for judging performance (Woehr and Huffcutt, 1994).

Third, future research may examine why candidates in operational ACs differ in their performance across exercises. For instance, Schneider and Schmitt (1992) identified variance due to the form of the exercise (e.g., role-play vs. group discussion) as the most important exercise factor to bolster performance differences across exercises. Future studies are needed to investigate whether, besides these situational characteristics, specific assessee characteristics may predict cross-situational inconsistent behavior. Examples of such assessee characteristics, which may lead candidates to adjust their behavior from one exercise to another, include impression management tactics, self-monitoring or tacit knowledge of how to behave in an ACs. No studies have addressed these issues in relation to AC performance.

Acknowledgements

Filip Lievens is postdoctoral research fellow of the Fund of Scientific Research, Flanders (Belgium) (F.W.O). He would like to acknowledge Wilfried De Corte, Terry Dickinson, and Paul Spector for their valuable suggestions on an earlier version of this manuscript. He is also especially grateful to Luc

Drieghe and Herlinde Pieters for their assistance in conducting the assessment centre simulation sessions.

References

Barr SH, Hitt MA. 1986. A comparison of selection decision models in managers versus student samples. *Personnel Psychology* **39**: 599–617.

Bernardin HJ, Buckley MR. 1981. Strategies in rater training. *Academy of Management Review* **6**: 205–212.

Brannick MT, Michaels CE, Baker DP. 1989. Construct validity of in-basket scores. *Journal of Applied Psychology* **74**: 957–963.

Brennan RL. 1992. *Elements of Generalizability Theory.* American College Testing Program: Iowa City, IA.

Bycio P, Alvares KM, Hahn J. 1987. Situational specificity in assessment centre ratings: a confirmatory factor analysis. *Journal of Applied Psychology* **72**: 463–474.

Campbell DT, Fiske DW. 1959. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin* **56**: 81–105.

Cardinet J, Tourneur Y, Allal L. 1976. The symmetry of generalizability theory: applications to educational measurement. *Journal of Educational Measurement* **13**: 119–136.

Cardy RL, Bernardin HJ, Abbott JG, Senderak MP, 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.

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

Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* John Wiley: New York.

Dugan B. 1988. Effects of assessor training on information use. *Journal of Applied Psychology* **73**: 743–748.

Fleenor JW. 1996. Constructs and developmental assessment centers: further troubling empirical findings. *Journal of Business and Psychology* **10**: 319–333.

Funder DC. 1987. Errors and mistakes: evaluating the accuracy of social judgment. *Psychological Bulletin* **101**: 75–90.

Gaugler BB, Rosenthal DB, Thornton GC, Bentson C. 1987. Meta-analysis of assessment centre validity. *Journal of Applied Psychology* **72**: 493–511.

Gaugler BB, Rudolph AS. 1992. The influence of assessee performance variation on assessors' judgments. *Personnel Psychology* **45**: 77–98.

Gaugler BB, Thornton GC. 1989. Number of assessment centre dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology* **74**: 611–618.

Harris MM, Becker AS, Smith DE. 1993. Does the assessment centre scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology* **78**: 675–678.

Howard A. 1997. A reassessment of assessment centers, challenges for the 21st century. *Journal of Social Behavior and Personality* **12**(5): 13–52.

Jones RG. 1992. Construct validation of assessment centre final dimension ratings: definition and measurement issues. *Human Resource Management Review* **2**: 195–220.

Joyce LW, Thayer PW, Pond SB. 1994. Managerial functions: an alternative to traditional assessment centre dimensions? *Personnel Psychology* **47**: 109–121.

Kane MT. 1982. A sampling model for validity. *Applied Psychological Measurement* **6**: 125–160.

Kauffman JR, Jex SM, Love KG, Libkuman TM. 1993. The construct validity of assessment centre performance dimensions. *International Journal of Selection and Assessment* **1**: 213–223.

Klimoski RJ, Brickner M. 1987. Why do assessment centers work? The puzzle of assessment centre validity. *Personnel Psychology* **40**: 243–260.

Kraiger K, Teachout MS. 1990. Generalizability theory as construct-related evidence of the validity of job performance ratings. *Human Performance* **3**: 19–35.

Kudisch JD, Ladd RT, and Dobbins GH. 1997. New evidence on the construct validity of diagnostic assessment centers: the findings may not be so troubling after all. *Journal of Social Behavior and Personality* **12**: 129–144.

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

Lievens F. in press. Assessor training strategies and their effects on accuracy, inter-rater reliability, and discriminant validity. *Journal of Applied Psychology.*

Macan TH, Avedon MJ, Paese M, Smith DE. 1994. The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology* **47**: 715–738.

Marcoulides GA. 1989. The application of generalizability analysis to observational studies. *Quality and Quantity* **23**: 115–127.

McHenry JJ, Schmitt N. 1994. Multimedia testing. In MJ Rumsey, CD Walker, J Harris (eds.) *Personnel Selection and Classification Research*, 193–222. Mahwah, NJ, Lawrence Erlbaum.

Neidig RD, Neidig PJ. 1984. Multiple assessment centre exercises and job relatedness. *Journal of Applied Psychology* **69**: 182–186.

Reilly RR, Henry S, Smither JW. 1990. An examination of the effects of using behavior checklists on the construct validity of assessment centre dimensions. *Personnel Psychology* **43**: 71–84.

Robertson I, Gratton L, Sharpley D. 1987. The psychometric properties and design of managerial assessment centres: dimensions into exercises won't go. *Journal of Occupational Psychology* **60**: 187–195.

Ryan AM, 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 PR, and Dreher GF. 1982. Constructs and assessment centre 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.

Schneider JR, Schmitt N. 1992. An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology* **77**: 32–41.

Shavelson RJ, Webb NM. 1991. *Generalizability theory: A primer.* Sage: Newbury Park, CA.

Shore TH, Thornton GC III, Shore LM. 1990. Construct validity of two categories of assessment center ratings. *Personnel Psychology* **43**: 101–116.

Spychalski AC, Quinones MA, Gaugler BB, Pohley KA. 1997. A survey of assessment center practices in organizations in the United States. *Personnel Psychology* **50**: 71–90.

Sulsky LM, Balzer WK. 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 centre operations. *Public Personnel Management* **18**: 457–470.

Thornton GC III. 1992. *Assessment Centers in Human Resource Management.* Addison-Wesley: Reading, MS.

Thornton GC III, Tziner A, Dahan M, Clevenger JP, Meir E. 1997. Construct validity of assessment centre judgments. *Journal of Social Behavior and Personality* **12**: 109–128.

Turnage JJ, Muchinsky PM. 1982. Transsituational variability in human performance within assessment centers. *Organizational Behavior and Human Performance* **30**: 174–200.

Woehr DJ, Huffcutt AI. 1994. Rater training for performance appraisal: a quantitative review. *Journal of Occupational and Organisational Psychology* **67**: 189–205.

Zedeck S. 1986. A process analysis of the assessment centre method. *Research in Organizational Behavior* **8**: 259–296.