Australian Journal of Psychology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713701010

An investigation into assessment centre validity, fairness, and selection drivers

K. V. Petrides[a]; Yana Weinstein[a]; Jenny Chou[a]; Adrian Furnham[a]; Viren Swami[a]
[a] Department of Psychology, University College London, London, United Kingdom

Online publication date: 05 November 2010

To cite this Article Petrides, K. V. , Weinstein, Yana , Chou, Jenny , Furnham, Adrian and Swami, Viren(2010) 'An investigation into assessment centre validity, fairness, and selection drivers', Australian Journal of Psychology, 62: 4, 227 — 235
To link to this Article: DOI: 10.1080/00049531003667380
URL: http://dx.doi.org/10.1080/00049531003667380

Australian Journal of Psychology, Vol. 62, No. 4, December 2010, pp. 227–235.

An investigation into assessment centre validity, fairness, and selection drivers

K. V. PETRIDES, YANA WEINSTEIN, JENNY CHOU, ADRIAN FURNHAM & VIREN SWAMI

Department of Psychology, University College London, London, United Kingdom

Abstract
This study examined the construct-related validity of an assessment centre (AC) developed by a national distribution company for the selection and development of lower-grade managers. In five locations throughout Britain, 487 individuals were observed on nine dimensions, each of which was measured through six distinct exercises. Multitrait-multimethod analyses conducted to investigate the convergent and discriminant validity of the AC revealed strong exercise (''method'') effects. This finding was corroborated by an exploratory factor analysis showing that AC ratings clustered into factors according to exercises, rather than according to performance dimensions. A series of MANOVAs and chi-squared tests demonstrated that neither the exercise ratings nor the selection decision were biased by sex, ethnicity, or training location, and a logistic regression determined which exercises had most impact on the final decision.

Keywords: *Assessment centres, industrial/organisational psychology, personality assessment, personnel selection*

The use of assessment centres (ACs) is widespread both in Europe and in North America (Woodruffe, 1993), their extensive appeal often attributed to their high predictive validity (Arthur, Day, McNelly, & Edens, 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987) and perceived fairness (Thornton & Byham, 1982). As regards the former, meta-analytic and primary studies have shown that ACs are superior to other selection methods in predicting performance ratings, status change, and person-organisation fit (Garavan, 2007; Klimoski & Brickner, 1987; but see Schmidt & Hunter, 1998, for an alternative view). With respect to the latter, a considerable body of evidence suggests that ACs produce few subgroup differences (Robertson, Iles, Gratton, & Sharpley, 1991) and thus do not disproportionately reject candidates from any particular social group (Huck & Bray, 1976; Ritchie & Moses, 1983), which is one factor that can contribute to ''adverse impact.'' Despite these advantages, the construct validity of ACs has not been convincingly established (Chen, 2006; Woehr & Arthur, 2003). A recent meta-analysis by Dean, Bobko, and Roth (2008) found that ACs do not fare as well as previously suggested as regards subgroup differences.

Taking a unitarian view (Landy, 1986), the construct validity of an AC might comprise three specific subtypes: content validity, criterion validity, and construct-related validity (Woehr & Arthur, 2003). Content validity refers to whether the activities which applicants have to undertake during their assessment adequately reflect the skills and tasks they will encounter on the job, and it is accepted that ACs do well in this respect (Klimoski & Brickner, 1987; Russell & Domm, 1995). Criterion validity (or predictive validity) refers to how well the outcome of the assessment procedure predicts future performance of the selected employees. It has generally been shown that ACs do well in terms of predicting job performance (Arthur et al., 2003; Gaugler et al., 1987), possibly thanks to their high content validity. Construct-related validity (not to be confused with overall construct validity) refers to how well the AC measures the dimensions it sets out to measure, and leads to correct inferences (Woehr & Arthur, 2003). To achieve high construct-related validity, ratings for the same dimension must be

Correspondence: Dr. K. V. Petrides, Department of Psychology, University College London, London, United Kingdom. E-mail: k.petrides@ucl.ac.uk

ISSN 0004-9530 print/ISSN 1742-9536 online © The Australian Psychological Society Ltd
Published by Taylor & Francis
DOI: 10.1080/00049531003667380

similar between different exercises (convergent validity), and ratings for different dimensions within an exercise should differ (discriminant validity). Findings relating to the construct-related validity of ACs have been equivocal, but meta-analyses by Lievens and Conway (2001) and Bowler and Woehr (2006) have shown that AC studies (e.g., Lievens & Klimoski, 2001; Sackett & Dreher, 1982) tend to produce low convergent and discriminant validity, with exercises, rather than dimensions, driving AC factor structure.

Lack of construct-related validity may not directly impact on the hiring process, but could have negative consequences at a later stage, producing feedback that is both misleading and potentially detrimental to a candidate's well-being (Fleenor, 1996; Joyce, Thayer, & Pond, 1994). As a result of the assessment procedure, candidates may be told to focus on improving performance in a particular dimension, which may not reflect the skills necessary for the job. Traditionally, it has been suggested that ACs are effective because the performance dimension ratings they produce capture individual differences in candidate skills and abilities (Byham, 1980). According to this view, the ratings of different dimensions obtained through the same exercises are expected to be orthogonal, thus displaying discriminant validity. In contrast, cross-exercise ratings pertaining to the same dimension are expected to correlate relatively strongly, thus displaying convergent validity. Contrary to theoretical expectations, empirical studies based on multitrait-multimethod matrices and factor analyses consistently reveal ''exercise effects.''

Multitrait-multimethod matrices (Campbell & Fiske, 1959) have revealed strong correlations between different performance dimensions assessed through the same method or exercise, and weak correlations between different methods or exercises intended to assess the same performance dimension (Bycio, Alveres, & Hahn, 1987; Chan, 1996; Fleenor, 1996; Robertson, Grattong, & Sharpley, 1987). For instance, Archambeau (1979) examined an AC with five exercises, designed to measure 10 performance dimensions and found that only 30 of the 81 monotrait-heteromethod correlations were statistically significant, which was interpreted as evidence of a lack of convergent validity. The same data also revealed a lack of discriminant validity, manifested through higher heterotrait-monomethod correlations than monotrait-heteromethod correlations (see also the meta-analysis by Lance, Lambert, Gewin, Lievens, & Conway, 2004).

Factor analytic studies have also produced equivocal findings with respect to the construct-related validity of AC dimensions. Thus, studies based on exploratory factor analyses (EFAs) have regularly reported that AC ratings produce fewer factors than their a priori hypothesised performance dimensions (Kauffman, Jex, Love, & Libkuman, 1993). In addition, some EFA studies have shown that assessor ratings tend to cluster according to exercises, rather than according to the performance dimensions that the ACs attempt to capture (Kauffman et al., 1993). For example, Sackett and Dreher (1982; see also Lance et al., 2004) found that the factor structure of AC ratings in three different organisations represented exercises and not performance dimensions. Several possible explanations have been offered to explain such findings, including actual exercise effects, low inter-rater reliabilities, flawed exercise designs, and the impact of situational determinants on behaviour.

The findings indicating relatively poor construct-related validity for the dimensions used in ACs have led many researchers to propose alternative explanations for their predictive validity (for an overview, see Lievens, 1998). One such explanation is that assessors may actually rate candidates in terms of their general performance in each exercise and not in terms of their specific behaviours in each performance dimension. In turn, these ratings predict job performance because the exercises included in ACs tend to be simulations of tasks occurring on the job (Hoeft & Schuler, 2001; Russell & Domm, 1995). This suggestion agrees with the conclusion that ACs mainly measure situation-specific performance (Bycio et al., 1987; Jackson, Stillman, & Atkins, 2005; Neidig & Neidig, 1984). Other explanations of the effectiveness of ACs revolve around assessors' inside knowledge of the organisation, range restriction effects, the use of background information about the candidates, self-fulfilling prophecy processes (it is often high-performing participants who get selected to attend), and the hypothesis that ACs may actually be measuring general intellectual ability (Klimoski & Brickner, 1987).

The present study

Our investigation aims to contribute to the validity debate by presenting data from a large AC developed by a national distribution company for the selection and development of lower-grade managers. The primary goal of our study was to determine whether an AC that was run in accordance with best practice recommendations would achieve construct-related validity. In relation to this goal, we implemented the following suggestions from the scientific literature: all assessors completed a general training course in which the errors and biases that are frequently made during the rating process were discussed, along with specific techniques that could be used to eradicate them (Woehr & Arthur, 2003). In order to limit the cognitive load placed on assessors, a maximum of

five dimensions were rated within each exercise (Gaugler & Thornton, 1989). Furthermore, assessors were provided with rating aids in order to reduce the cognitive demands of the rating task by prespecifying relevant behaviours (Reilly, Henry, & Smither, 1990). Special efforts were made to keep the form of the various exercises in this AC identical, and all assessments were carried out in a one-to-one context to reduce the likelihood of exercise effects. A rotation system, whereby all candidates were rated by at least two assessors, was introduced in order to minimise the likelihood of rating biases. Finally, as recommended by Kolk, Born, and van der Flier (2003), all measured dimensions were made transparent to participants, who received the relevant information two weeks prior to the assessment.

In addition to the primary goal outlined above, we had two additional goals: to examine issues of subgroup differences and fairness, and to determine which type of exercise is given the most weight in the final selection decision. The following hypotheses were formulated in relation to the three goals:

H1a. A multitrait-multimethod analysis would provide evidence of convergent and discriminant validity in the form of higher monotrait-heteromethod correlations than heterotrait-monomethod correlations.
H1b. Assessor ratings would cluster according to the performance dimensions of the AC, rather than according to the exercises used to assess those dimensions.
H2. The AC would not show any evidence of subgroup or training location differences.
H3. Assessors would weigh some exercises more than others in their final selection decision. We did not have a theory as to which exercises would have most impact on the final decision, thus this was an exploratory hypothesis.

Method

Participants

Participants were 487 individuals who had applied for lower or assistant management positions. Of the total sample, 72.3% were male and 9.9% were female (17.9% did not report their sex)[1]. Age ranged between 21 and 58 years ($M = 36.05$, $SD = 7.82$). As regards ethnicity, 76.4% were European Caucasian, 2.9% were Asian, 1.4% were African Caribbean, 1.0% described themselves as ''other'' and 18.3% did not report ethnic information. Finally, 65% of participants were internal employees and the rest were external candidates.

Assessment centre procedure

The main functions of the AC were to select employees for lower- and middle-management positions, as well as to develop new skills in internal employees. The selection process lasted a full day and was conducted in five locations across Britain (London, Birmingham, Bristol, Edinburgh, and Newcastle) over a period of two years. The exercises and performance dimensions were based on job analyses, involving interviews with job incumbents, managers, and experts from the target organisation. Two weeks prior to the assessment day, participants were provided with extensive information explaining the purpose of the centre and what was expected of them. The information pack included a detailed description of the exercises and dimensions.

Each AC had two assessors and one chairperson. Assessor 1 assessed the briefing exercise and interview 1. Assessor 2 assessed the role-play exercise and interview 2. The chairperson conducted a chair interview, acted as the role-player, and chaired the ''wash-up'' (consensus) session during which assessors met to discuss their final ratings and selection decisions.

Assessor training

All assessors attended a one-week general-purpose course providing basic training about the purpose of the assessment, their assigned roles, discrimination legislation, and rater errors. In addition, assessors received training in observation, evaluation, interviewing, and report-writing skills. Two weeks before the assessment day, assessors were provided with detailed information about the exercises, the dimensions and dimension-relevant behaviours, and the rating procedures.

Dimensions

Nine distinct performance dimensions were measured, namely ''managing resources,'' ''oral communication,'', ''managing others,'' ''building relationships,'' ''managing self,'' ''making decisions,'' ''written communication,'' ''improving business,'' and ''satisfying customers and managing commercially.'' Table I presents a brief description of each dimension.

Exercises

Six exercises were used to measure the performance dimensions: an in-tray and scheduling exercise, a briefing exercise, a role-play exercise, a numerical test, and two semi-structured capability interviews. Interview 1 assessed capabilities pertaining to ''improving business,'' ''managing resources,'' ''making decisions,'' and ''managing commercially.'' Interview 2 assessed capabilities pertaining to ''managing others,'' ''building relationships,'' ''managing self,''

Table I. Descriptive statistics for the assessment centre ratings

Source of rating	Dimension	Brief description of dimension	*M*	*SD*
In-tray exercise				
	Managing resources	Gathers relevant information by utilising the resources available; analyses simple data from one or two sources to schedule activities; anticipates future needs and plans for contingencies.	2.34	0.81
	Making decisions	Makes rational decisions based on accurate data analysis; evaluates different courses of action and considers the implications and impact of decisions.	2.21	0.75
	Written communication	Presents information in a clear and concise format with appropriate structure; uses correct spelling and grammar; presents ideas succinctly.	2.74	0.65
	Improving business	Considers impact of own activities on other business functions; identifies the components of a problem; proposes and implements solutions successfully.	2.18	0.71
	Satisfying customers	Understands customer needs; strives to maximise customer satisfaction; deals effectively with customer complaints and considers the impact of own work on customers.	2.42	0.76
Numerical test				
	Raw numerical test score	Possesses basic number skills (e.g., arithmetical operations like addition and subtraction) that are needed to make correct decisions or inferences from numerical data.	21.57	6.06
Briefing exercise				
	Oral communication	Accurately and clearly expresses ideas and information; appreciates the target audience and tailors a message appropriately; demonstrates good listening skills and encourages interaction with audience.	2.81	0.77
	Managing others	Manages others to achieve targets; prioritises and effectively delegates workload; encourages others to contribute; addresses unacceptable behaviour and poor performance in an appropriate manner.	2.72	0.74
	Managing self	Is resilient and copes effectively in demanding situations; keeps clam, holds the goal in sight and maintains credibility under pressure; seeks personal feedback and takes responsibility for own actions.	2.70	0.79
	Improving business	Described above.	2.46	0.78
Role-play exercise				
	Oral communication	Described above.	2.83	0.73
	Managing others	Described above.	2.64	0.76
	Building relationships	Establishes rapport and develops good working relationships; works well with people from a range of backgrounds; demonstrates respect and trust and shows appropriate empathy and sensitivity.	2.76	0.78
	Managing self	Described above.	2.75	0.73
	Improving business	Described above.	2.47	0.74
Interview 1				
	Managing resources	Described above.	2.84	0.64
	Making decisions	Described above.	2.79	0.65
	Improving business	Described above.	2.80	0.70
	Satisfying customers	Described above.	2.71	0.72
Interview 2				
	Managing others	Described above,	2.90	0.69
	Building relationships	Described above.	3.06	0.63
	Managing self	Described above.	2.90	0.66
	Satisfying customers	Described above.	2.92	0.64

Note: The numerical test has been measured on a different scale (ranging from 0 to 40) than the other variables in the table.

and "satisfying the customer." The numerical test was part of the Personnel Test Battery developed by an international test development company and certified by the British Psychological Society. It assessed basic number skills that are needed to make correct decisions or inferences from numerical data (e.g., arithmetical operations).

In addition to the six exercises, the assessment day included a chair interview, which aimed to ensure that candidates had a realistic picture of the job for which they were applying and also to explore the reasons for their wanting to become lower-level or assistant managers. The chair interview was not part of the assessment and did not affect the final decision.

Dimension ratings

Assessors were required to record candidate performance on a piece of paper first and then to identify and rate the observed behaviours. Each exercise was designed to measure between three and five performance dimensions. Within each dimension, there were several behavioural examples. Accompanying each dimension within each exercise was a rating scale based on behavioural descriptions for a range of possible responses (Behavioural Assessment and Research System; BARS). BARS outlined these behaviours in rows, with each row describing three different levels of behaviour.

The first description, under heading 1 (Grade 1), corresponded to very poor performance; the description under heading 3 (Grade 3) corresponded to adequate performance with minor weaknesses that may be rectifiable with training; last, the description under heading 5 (Grade 5) corresponded to excellent performance. Assessors had to select a description that best fits the candidate's behaviour and subsequently to assign a preliminary grade. The same procedure was carried out for all dimensions across all exercises.

At the end of the assessment day, assessors gathered for the consensus session during which they discussed and assigned the final grades. In order to pass, candidates had to achieve no more than one Grade 2, with the rest of the marks at Grade 3 or above. In those cases where the successful candidates exceeded the positions available, they were rank-ordered based on their total scores. Once selection decisions were made, written feedback reports with suggestions on how to improve dimension-related behaviours were produced for all internal candidates.

Results

Construct-related validity

Descriptive statistics for the dimension ratings are shown in Table I. Mean ratings ranged from 2.18 ("improving business" in the in-tray exercise) to 3.06 ("building relationships" in interview 2). *SD*s ranged from.63 ("building relationships" in interview 2) to .81 ("managing resources" in the in-tray exercise). These values indicate that the ability ratings did not vary greatly.

The multitrait-multimethod correlations are presented in Table II. Although the mean monotrait-heteromethod correlation coefficient was positive ($r = .29$), the corresponding value for the heterotrait-monomethod matrix was significantly larger ($r = .57$; $z = 5.43$, $p < .01$), thus failing to support *H1a*. These values are indicative of an "exercise effect" influencing the ratings and potentially compromising convergent and discriminant validity.

Exercise/dimension clustering

To determine whether ratings clustered according to exercises or performance dimensions, we conducted an EFA. Both the scree plot and the Kaiser-Guttman "eigenvalue > 1" rule suggested a five-factor solution (first six eigenvalues were 7.87, 2.39, 1.83, 1.29, 1.03, and.74). Five factors, accounting for 56% of the total variance, were extracted through principal components analysis and rotated via the PROMAX algorithm ($\delta = 4$). The factor loadings, reported in Table III, confirmed the presence of strong exercise effects, thus failing to support *H1b*.

More specifically, the five-dimension ratings from the role-play exercise loaded strongly onto Factor 1, the five-dimension ratings from the in-tray exercise loaded onto Factor 2, the four-dimension ratings from the briefing exercise loaded onto Factor 3, and

Table II. Mean Monotrait-Heteromethod and Heterotrait-Monomethod correlation coefficients for key dimensions in the assessment centre

Key dimensions	Correlation coefficient
Inter-exercise items	
Managing resources	.21
Oral communication	.35
Managing others	.28
Building relationships	.43
Managing self	.29
Making decisions	.26
Improving business	.22
Satisfying customers	.24
Mean *r*	.29
Intra-exercise items	
In-tray/scheduling	.47
Briefing	.64
Role-play	.65
Interview 1	.53
Interview 2	.54
Mean *r*	.57

Note: All correlations are significant at $p < .01$.

Table III. Factor pattern matrix

Dimensions and (exercises)	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Managing others (role-play)	**.88**	−.04	−.02	−.06	.05
Oral communication (role-play)	**.82**	−.01	.03	.11	−.07
Managing self (role-play)	**.79**	−.04	.01	.05	−.02
Building relationships (role-play)	**.77**	.02	−.02	−.05	.08
Improving business (role-play)	**.67**	.07	.00	.01	−.01
Making decisions (in-tray)	.00	**.79**	−.04	.04	.00
Managing resources (in-tray)	−.02	**.74**	.00	−.09	.02
Improving business (in-tray)	.00	**.68**	.01	.01	−.01
Satisfying customers (in-tray)	−.08	**.63**	−.02	.06	.04
Written communication (in-tray)	.12	**.60**	.07	−.03	−.05
Oral communication (briefing)	.04	.02	**.91**	−.12	.01
Managing others (briefing)	−.01	−.07	**.81**	.01	.01
Managing self (briefing)	−.01	.02	**.76**	.04	.02
Improving business (briefing)	−.04	.07	**.51**	.22	−.06
Improving business (interview 1)	−.03	−.06	.00	**.85**	.02
Making decisions (interview 1)	.00	−.01	−.04	**.70**	.08
Managing resources (interview 1)	−.01	.01	.11	**.67**	−.04
Satisfying customers (interview 1)	.08	.05	−.03	**.62**	−.01
Managing others (interview 2)	.06	−.02	.10	−.02	**.73**
Managing self (interview 2)	.00	.02	−.03	.00	**.73**
Building relationships (interview 2)	.07	−.04	−.04	.02	**.67**
Satisfying customers (interview 2)	.05	.09	.00	.13	**.48**

Note: Loadings greater than |.40| are in bold.

dimension ratings for the two interviews loaded onto Factors 4 and 5. Table IV presents the factor intercorrelation matrix, which reveals considerable variance overlap, with factor correlations ranging between .35 and .67.

Table IV. Factor intercorrelation matrix

Factors	1	2	3	4
1. Role-play	–			
2. In-tray	.35	–		
3. Briefing	.41	.47	–	
4. Interview 1	.49	.45	.60	–
5. Interview 2	.67	.42	.48	.60

Subgroup differences

Four multivariate analyses of variance were performed with scores on the numerical test, the in-tray, briefing, role-play, and the two interviews as dependent variables, and sex, age, ethnicity, and training location (London versus outside) as the independent variables. Neither the multivariate nor the *post-hoc* univariate tests revealed significant differences in ratings by sex (Wilks' $\lambda_{(6,\ 356)} = 1.96$, $p > .05$), age (Wilks' $\lambda_{(18,\ 959)} = 1.19$, $p > .05$), ethnic group (Wilks' $\lambda_{(6,\ 355)} = 1.32$, $p > .05$), or training location (Wilks' $\lambda_{(6,\ 353)} = .70$, $p > .05$).

Sex, age, ethnicity, and training location were also modelled as independent variables in four separate chi-square analyses with the outcome of the selection process as the dependent variable. Sex, $\chi^2_{(1)} = .70$, $p > .05$, ethnicity, $\chi^2_{(6)} = 10.91$, $p > .05$, and training location, $\chi^2_{(1)} = .12$, $p > .05$, all yielded non-significant results. However, the overall chi-square was significant for age, $\chi^2_{(3)} = 8.38$, $p < .05$, and follow-up tests revealed that significantly more 30- to 39-year-olds were offered positions than either 20- to 29-year-olds, $\chi^2_{(1)} = 6.23$, $p < .05$, or 40- to 49-year-olds, $\chi^2_{(1)} = 3.96$, $p < .05$. With the exception of the finding concerning age, these results provide solid support for *H2*.

Predictors of final outcome

One aim of the data analysis was to establish which type of exercise was given the most weight in the final selection decision. Given that the selection process resulted in a dichotomous decision (offer of position or rejection), the data were analysed through a binary logistic regression (an extension of multiple regression, which allows for the fact that the dependent variable is dichotomous; Agresti, 2007).

A model with the six exercises as predictors (in-tray, role-play, briefing, interview 1, interview 2, and the numerical test) was significantly better than the null model, $\chi^2_{(6)} = 184.12$, $p < .01$. The pseudo-R^2s ranged between .40 (Cox & Snell statistic; see Hosmer & Lemeshow, 2000) and .64 (Negelkerke statistic; see Hosmer & Lemeshow, 2000). The model was better at predicting negative (i.e., position not offered) than positive decisions (95.1% and 63.9% hit rate, respectively). In total, 88.8% of decisions were correctly classified. As regards individual predictors, interview 1 ($expB = 1.82$; Wald = 15.60, $p < .01$), interview 2 ($expB = 1.68$;

Wald = 13.45, $p < .01$), role-play ($expB = 1.38$; Wald = 14.60, $p < .01$), and the in-tray exercises ($expB = 1.37$, Wald = 11.12, $p < .01$) reached significance levels. In contrast, neither the briefing exercise (Wald = 2.31, $p > .05$) nor the numerical test (Wald = .62, $p > .05$) were reliable predictors in the equation.

The odds ratios in the parentheses above (*expB*) indicate the change in odds of being offered a position when the value of a predictor increases by one unit, while holding all other predictors in the equation constant. Thus, a one-unit increase in the in-tray exercise score multiplies the odds of being offered a position 1.37 times. Similarly, a one-unit increase in the role-play exercise score multiplies the odds 1.38 times, a one-unit increase in the first interview score multiplies the odds 1.82 times, and finally, a one-unit increase in the second interview score multiplies the odds 1.68 times. In summary, when deciding whether or not to offer a position, assessors weighted interview 1 as most important, followed by interview 2, the role-play exercise, and, last, the in-tray exercise, results that are in line with the exploratory hypothesis *H3*.

Discussion

This study is rare in providing recent British data from a large and multifaceted AC, while simultaneously considering the issue of adverse impact in relation to gender, age, ethnicity, and geographical location. As such, it makes a substantial empirical contribution to the ongoing debate about the construct validity and fairness of ACs (Dean et al., 2008; Gibbons & Rupp, 2009; Lance, Woehr, & Meade, 2007).

Contrary to *H1a*, the multitrait-multimethod matrix showed higher heterotrait-monomethod than monotrait-heteromethod correlations. This pattern indicated salient exercise (''method'') effects and performance dimensions that were insufficiently distinguishable from each other. These results were corroborated by the EFA wherein the factors were defined by exercises, rather than by performance dimensions, disconfirming *H1b*. The analyses revealed that assessor ratings tended to reflect performance on exercises and not on dimensions as intended, which constitutes evidence of lack of discriminant and convergent validity (Bycio et al., 1987; Jackson et al., 2005).

It has been suggested that lack of discriminant validity reflects either a ''halo effect'' or a true relationship between the dimensions, while lack of convergent validity reflects either low inter-rater reliability or the multifaceted nature of certain performance dimensions (Robertson et al., 1987). It seems unlikely that ''illusory halo'' (Cooper, 1981) or inter-rater unreliability effects are the sole or even the main reasons for the results obtained in the present study because all assessors employed in the AC had followed a standardised training course designed to inform them about these problems and how to address them. However, since this training course was not formally evaluated, it is not prudent to rule out the possibility that such effects may have played a role.

An alternative explanation is that different exercises tap into different aspects of the same dimension (Hoeft & Schuler, 2001; Robertson et al., 1987; see also Lievens, 2001). In other words, candidate behaviour in the AC may be largely exercise-specific, thus producing salient exercise effects. This implies that the construct-related validity of AC dimensions may partly depend on the relative importance of traits versus situations in determining behaviour in general (Epstein & O'Brien, 1985; Fleeson, 2004; Mischel, 1968). Another alternative explanation posits that the different demands placed on applicants by each exercise require them to engage in behaviours that cannot be grouped into dimensions (Lievens, Chasteen, Day, & Christiansen, 2006). Such a view is inconsistent with the idea that more than one exercise taps into a unifying construct. Finally, questions remain about the extent to which assessor ratings in ACs actually measure the dimensions that they claim to measure. Like other ACs, the present one appears to have limited construct-related validity. In light of such results, organisations should either avoid using performance dimension ratings to select and develop candidates or predicate their feedback reports on exercise-specific behaviours.

The analysis of subgroup differences showed that neither the exercise ratings nor the final outcome were influenced by sex, ethnicity, or training location. These results support *H2* and corroborate the prevailing view that well-designed ACs generally lead to fair and unbiased decisions (Ritchie & Moses, 1983; Shore, Tashchian, & Adams, 1997) and are unlikely to have adverse impact. However, this conclusion must be qualified, since age did have a small impact on the final outcome, with 30- to 39-year-olds being offered more positions than other age groups. One explanation is that assessors valued the fact that 30- to 39-year-olds have substantial working experience, while still being flexible and malleable. It should be noted that the age effect was evident only in the final outcome and not in any of the exercise ratings. However, as Anderson, Lievens, van Dam, and Born (2006) point out, sophisticated statistical techniques, such as latent means analysis, may be more sensitive to group differences than conventional univariate comparisons (see also Petrides, Jackson, Furnham, & Levine, 2003). Furthermore,

Dean et al.'s (2008) meta-analysis revealed that ACs are sometimes biased with respect to particular intergroup differences. The data at hand do not allow for a more detailed examination of the effect of age, since crucial information about the candidates (e.g., work experience and educational qualifications) was not available, but this could be a topic for future investigations.

Given that the underlying structure of the ratings can best be summarised in terms of exercises, we investigated which exercises assessors took most into account when deciding whether to offer a position to a candidate. A logistic regression revealed that, as predicted by *H3*, not all exercises contributed to the selection decisions. Instead, assessors appeared to place greater weight on the results of the two interviews, the role-play, and the in-tray exercises. Conversely, the least weight was placed on the numerical test and the briefing exercise. These findings are an interesting starting point for investigating the selection process, but other factors we did not measure (e.g., individual differences in personality and emotions; Lievens, de Fruyt, & van Dam, 2001; Petrides, Pita, & Kokkinaki, 2007) may also have contributed to the final decision.

In conclusion, our findings suggest that the AC in this study does not have strong convergent and discriminant validity. This is at odds with the view that ACs are effective because their dimension-related ratings capture individual differences in candidate skills and abilities. Alternative explanations seem to be more plausible, including that ACs mainly measure situation-specific performance and that assessors rate the candidates according to their general performance in each exercise, rather than according to specific dimension-related behaviours. In turn, as noted by Klimoski and Brickner (1987) and Russell and Domm (1995), these ratings predict job performance because the exercises included in the ACs tend to be accurate simulations of tasks occurring on the job. In our exploratory analysis, we showed that exercise ratings had considerable explanatory power in relation to the final selection decision. Future research should focus on improving our understanding of the impact of the different types of exercises used in ACs with a view to optimising the selection process and the feedback provided to candidates.

References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.

Anderson, N., Lievens, F., van Dam, K., & Born, M. (2006). A construct-driven investigation of gender differences in a leadership-role assessment center. *Journal of Applied Psychology*, *91*, 555–566.

Archambeau, D. J. (1979). Relationships among skill ratings assigned in an assessment center. *Journal of Assessment Center Technology*, *2*, 7–20.

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *56*, 125–154.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension ad exercise factors on assessment center ratings. *Journal of Applied Psychology*, *91*, 1114–1124.

Bycio, P., Alveres, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology*, *72*, 463–474.

Byham, W. C. (1980). Starting an assessment center the correct way. *Personnel Administrator*, *25*, 27–32.

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

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

Chen, H-C. (2006). Assessment center: A critical mechanism for assessing HRD effectiveness and accountability. *Advances in Developing Human Resources*, *8*, 247–264.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, *90*, 218–244.

Dean, M. A., Bobko, P., & Roth, P. L. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology*, *93*, 685–691.

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

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

Fleeson, W. (2004). Moving personality beyond the person-situation debate - The challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, *13*, 83–87.

Garavan, T. N. (2007). Using assessment centre performance to predict subjective person-organisation (P-O) fit: A longitudinal study of graduates. *Journal of Managerial Psychology*, *22*, 150–167.

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

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

Gibbons, A. M., & Rupp, D. E. (2009). Dimension consistency as an individual difference: A new (old) perspective on the assessment center construct validity debate. *Journal of Management*, *35*, 1154–1180.

Hoeft, S., & Schuler, H. (2001). The conceptual basis of assessment centre ratings. *International Journal of Selection and Assessment*, *9*, 114–123.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Huck, J. R., & Bray, D. W. (1976). Management assessment center evaluations and subsequent job performance of Black and White females. *Personnel Psychology*, *29*, 13–30.

Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance*, *18*, 213–241.

Joyce, L. W., Thayer, P. W., & Pond, S. B. (1994). Managerial functions: An alternative to traditional assessment centre dimensions? *Personnel Psychology*, *47*, 109–121.

Kauffman, J. R., Jex, S. M., Love, K. G., & Libkuman, T. M. (1993). The construct validity of assessment centre performance dimensions. *International Journal of Selection and Assessment*, *1*, 213–223.

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

Kolk, N. J., Born, M. P., & van der Flier, H. (2003). The transparent assessment centre: The effects of revealing dimensions to candidates. *Applied Psychology*, *52*, 648–668.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center post-exercise dimension ratings. *Journal of Applied Psychology*, *89*, 377–385.

Lance, C. E., Woehr, D. J., & Meade, A. W. (2007). A Monte Carlo investigation of assessment center construct validity models. *Organizational Research Methods*, *10*, 430–448.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*, 1183–1192.

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

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

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology*, *91*, 247–258.

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*, *89*, 377–385.

Lievens, F., de Fruyt, F., & van Dam, K. (2001). Assessors' use of personality traits in descriptions of assessment centre candidates: A five-factor model perspective. *Journal of Occupational and Organizational Psychology*, *74*, 623–636.

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment centre process: Where are we now? *International Review of Industrial and Organizational Psychology*, *16*, 245–286.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

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

Petrides, K. V., Jackson, C. J., Furnham, A., & Levine, S. Z. (2003). Exploring issues of personality measurement and structure through the development of a revised short form of the Eysenck Personality Profiler. *Journal of Personality Assessment*, *81*, 272–281.

Petrides, K. V., Pita, R., & Kokkinaki, F. (2007). The location of trait emotional intelligence in personality factor space. *British Journal of Psychology*, *98*, 273–289.

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.

Ritchie, R. J. & Moses, J. L. (1983). Assessment center correlates of women's advancement into middle management: A 7-year longitudinal analysis. *Journal of Applied Psychology*, *68*, 227–231.

Robertson, I. T., 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.

Robertson, I. T., Iles, P. A., Gratton, L., & Sharpley, D. (1991). The impact of personnel selection and assessment methods on candidates. *Human Relations*, *44*, 693–982.

Russell, C. J., & Domm, D. R. (1995). Two field tests of an explanation of assessment center validity. *Journal of Occupational and Organizational Psychology*, *68*, 25–47.

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

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–273.

Shore, T. H., Tashchian, A., & Adams, J. S. (1997). The role of gender in a developmental assessment center. *Journal of Social Behavior and Personality*, *12*, 191–203.

Thornton, G. C., & Byham, W. C. (1982). *Assessment centers and managerial performance.* New York: Academic Press.

Woehr, D. J., & Arthur, W. Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management*, *29*, 231–258.

Woodruffe, C. (1993). *Assessment centres: Identifying and developing competence.* London: Institute of Personnel Management.