Investigating the construct validity of a development assessment centre

Authors:
Nadia M. Brits[1]
Deon Meiring[2]
Jürgen R. Becker[3]

Affiliations:
[1]Department of Industrial Psychology, Stellenbosch University, South Africa

[2]Department of Human Resource Management, University of Pretoria, South Africa

[3]Department of Industrial Psychology and People Management, University of Johannesburg, South Africa

Correspondence to:
Deon Meiring

Email:
deon.meiring@up.ac.za

Postal address:
Private Bag X20, Pretoria 0028, South Africa

Dates:
Received: 04 Feb. 2013
Accepted: 20 Aug. 2013
Published: 14 Nov. 2013

How to cite this article:
Brits, N.M., Meiring, D., & Becker, J.R. (2013). Investigating the construct validity of a development assessment centre. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde*, *39*(1), Art. #1092, 11 pages. http://dx.doi.org/10.4102/sajip.v39i1.1092

Orientation: The assessment centre (AC) is a prominent measurement tool for selection and development.

Research purpose: The aim of this study was to determine the construct validity of a one-day development assessment centre (DAC) using a convenience sample of 202 managers in a large South African banking institution.

Motivation for the study: Although the AC method is popular, it has been widely criticised as to whether it predominantly measures the dimensions it is designed to measure.

Research design, approach and method: The fit of the measurement models implied by the dimensions measured was analysed in a quantitative study using an ex post facto correlation design and structural equation modelling.

Main findings: Bi-factor confirmatory factor analysis was used to assess the relative contribution of higher-order exercise and dimension effects. Empirical under-identification stemming from the small number of exercises designed to reflect designated latent dimensions restricted the number of DAC dimensions that could be evaluated. Ultimately, only one global dimension had enough measurement points and was analysed. The results suggested that dimension effects explained the majority of variance in the post-exercise dimension ratings.

Practical/managerial implications: Candidates' proficiency on each dimension was used as the basis for development reports. The validity of inferences holds important implications for candidates' career development and growth.

Contribution/value-add: The authors found only one study on construct validity of AC dimensions in the South African context. The present study is the first use the bi-factor approach. This study will consequently contribute to the scarce AC literature in South Africa.

Introduction

Over the past 60 years assessment centres (ACs) have become a popular method for evaluating individual behaviour (performance) for both selection and development purposes. This popularity is attributable to the method's numerous strengths, which include the fact that ACs demonstrate little adverse impact (Iles, 1992; Thornton & Rupp, 2006) and predict a variety of performance criteria (Thornton & Rupp, 2006) with impressive predictive validity correlations ranging from 0.37 to 0.52 (Gaugler, Rosenthal, Thornton & Bentson, 1987; Rupp, Thornton & Gibbons, 2008; Thornton & Gibbons, 2009). In addition, the method has been shown to have high criterion-related validity (Schmitt, Gooding, Noe & Kirsch, 1984) as well as content validity (Gaugler *et al*., 1987; Iles, 1992). Furthermore, simulations, which are a necessary component for a centre to be acknowledged as an AC, receive high scores on fidelity as they present job-related situations to candidates (Lievens & Schollaert, 2011; Thornton & Rupp, 2006). Although ACs are usually expensive in terms of time, effort and money, their excellent predictive validity (Krause, Kersting, Heggestad & Thornton, 2006) and other strengths mean that this assessment technique is generally regarded as financially worthwhile. However, the utility of ACs has been somewhat tainted by the inconsistent and contradictory construct validity evidence found in applied research (Meriac, Hoffman, Woehr & Fleisher, 2008).

Although the AC method has earned a reputation as a relatively fair and unbiased selection technique (Thornton & Gibbons, 2009), its construct validity has been debated for years due to contradictory research results. For example, authors such as Rupp *et al*. (2008) and Arthur, Woehr and Maldegen (2000) have claimed that the extensive collection of research evidence accumulated over more than five decades clearly confirms the method as a valid assessment and development tool for job-related performance dimensions. However, some research findings suggest that ACs do not in fact display satisfactory construct validity (Thornton & Rupp, 2003).

Murphy (2010) suggested that evidence to show that the necessary constructs are actually measured is scarce. The construct-related validity of the assessed dimensions has been described as the method's weakness (Bowler & Woehr, 2009) and its Achilles' heel (Lance, 2008b). Clearly, there is a great deal of controversy surrounding the construct-related validity of the performance dimensions of ACs (Bowler & Woehr, 2009; Lievens, 2001; Sackett & Dreher, 1982).

Purpose of the study

The 'so-called construct validity problem' (Howard, 1997, p. 21) served as the foundation for the present research study within the South African context. According to the traditional AC model, AC ratings should cluster together according to dimensions, as opposed to exercises. Thus, AC performance ratings on the same dimension within different exercises should show a high degree of correspondence (convergent validity) whilst the ratings of different dimensions measured within the same exercise should show relatively low levels of congruence with each other (discriminant validity) (Lievens, Chasteen, Day & Christiansen, 2006). However, the bulk of AC evidence does not reflect cross-dimensional discrimination (Jackson, Atkins & Stillman, 2005; Lance, 2008a; Lievens, 2009). Despite the honourable design intentions of most ACs, weak empirical support is consistently found for the cross-exercise dimension ratings of performance constructs. This has resulted in multiple authors (e.g. Bowler & Woehr, 2009; Crawley, Pinder & Herriot, 1990) investigating the 'construct validity puzzle' (Jackson, Barney, Stillman & Kirkley, 2007, p. 415) and 'construct-related validity paradox' (Arthur, Day & Woehr, 2008, p. 105) in relation to ACs. However, the majority of the research findings related to the construct validity of ACs stem from international research and South Africa-related research is almost non-existent.

Research objectives

The main research objective of the present study was to examine the construct validity of singular dimensions measured in a DAC. The DAC initially consisted of 12 primary dimensions that were grouped into five global dimensions (performance motivation competence, decision-making skills, leadership skills, communication skills and administration skills) measured by four exercises (analysis problem, group discussion, one-on-one interview and persuasive presentation). In order to run confirmatory factor analysis (CFA) procedures on AC data, every dimension should be measured by a minimum of three exercises (Kline, 2011). Regrettably, only one of the original five global dimensions had sufficient ratings (items) to fit the CFA model. For this reason, only the results of one global dimension, leadership skills, were considered[1], instead of the whole DAC model.

The first step was to determine the construct validity of the single dimension single exercise leadership measurement model. Based on whether or not a tenable model fit was found for this step, the next step would involve analysing the amount of additional variance explained when adding the conglomerated exercise effect to the single dimension CFA model. The resulting model can be described as a bi-factor model with a single conglomerated dimension and exercise effect (1D1E measurement model).

The following overarching questions and objectives guided this study:

- To what extent does the proposed theoretical model of the leadership dimension reasonably correspond to the empirical data?
- How much additional variance is explained by the method effect in a model that already contains the leadership dimension?

In conjunction, the answers to these two questions indicated whether valid inferences can be made from the specific leadership dimension of the DAC.

This research has both conceptual and practical importance. Conceptually this study emphasises a movement in the AC field away from either task-based models or dimension-based models to the new mixed-model approach. On a practical level the empirical results of this research will contribute to the scarce AC literature in South Africa and provide recommendations to AC practitioners on designing and validating ACs.

The study by Greyling, Visser and Fourie (2003) appears to be the only South African study that has contributed to the construct validity debate. In accordance with the bulk of international research findings, the Greyling *et al*. study also reported strong support for exercise effects as opposed to dimension effects. The present study was the first South African study to use CFA on AC ratings used for development purposes.

Review of the literature

Construct validity

The origins of the construct validity debate can be traced back to Sackett and Dreher (1982), who discovered that low correlations existed amongst ratings of a single dimension across exercises, and high correlations existed amongst ratings of various dimensions within a single exercise. Further analysis indicated differences in candidates' behaviours attributable to variance in their performance in exercises rather than variance in their behaviour on the dimensions measured (i.e. exercise effect). This landmark article led to a flurry of research that confirmed Sackett and Dreher's findings. Most research studies found low (or absent) discriminant validity (Huysamen, 1996; Schneider & Schmitt, 1992; Spector, Schneider, Vance & Hezlett, 2000), low convergent validity (Robertson, Gratton & Sharpley, 1987) and a significant level of exercise effect (Bowler & Woehr, 2006; Robertson *et al*., 1987; Schneider & Schmitt, 1992). These consistent findings of low construct validity instigated heated debates in the AC field, with a shared concern in most

1.Please contact the authors for more information on the remaining global and primary dimensions.

of these debates about the consequences of using invalid AC post-exercise dimension ratings (PEDRs).

Firstly, failure to identify stable characteristics that may be advantageous in determining which individual is best suited for a job position limits the usefulness of the AC process for recruitment and selection (Lowry, 1996). Secondly, when used as development tools AC ratings provide information about a person's strengths and weaknesses on the dimensions measured and developmental feedback is typically formulated around these dimensions (Lievens & Christiansen, 2010). Thirdly, if dimensions fail to measure a person's proficiency on the specific constructs they were intended to measure then the AC method fails to be useful and will not realise equitable return on investment when compared to alternative assessment techniques (e.g. psychometric tests and reference checks). The quality of decisions made based on individuals' performance on AC dimensions is clearly dependent on the construct validity of these dimensions. Several solutions, which became known as design fixes, were proposed in an attempt to enhance construct validity results for ACs.

In light of the prevailing inconsistent construct validity evidence, Lance (2008b) asserted that ACs do not work in the way in which they were designed to work. He suggested a number of redesign solutions (design fixes) to overcome what he saw as inherent problems with the construct validity of ACs. These design fixes included the definitions of dimensions, the number of dimensions observed and recorded, assessor training, as well as the type of evaluation approach. It was concluded that design fixes have unfortunately resulted in only small improvements in construct validity findings (Jackson *et al.*, 2007). Exercise factors continue to predominate despite alternative scoring methods and dimension definition strategies designed specifically to emphasise cross-exercise consistency in PEDRs (Harris, Becker & Smith, 1993). In addition, Lievens (1998) and Woehr and Arthur (2003) indicated that although some design fixes have resulted in slight improvements in AC construct validity, the basic pattern of findings remains unchanged and suggests that PEDRs substantially reflect the effects of the exercises in which they were measured and not the behavioural or performance dimensions they were designed to assess.

Assessment centres at a crossroad

Following decades of conflicting research findings concerning the construct validity of AC dimensions, the popular method has reached a crossroad: dimension-based ACs or exercise-based ACs. Repeated research results suggest that candidates perform inconsistently across exercises due to method variance or bias. As a result of findings such as these and the presence of persistent exercise effects, Lievens (2002) conducted two studies to determine the extent of cross-situational candidate behaviour. The results of Lievens's studies suggest that variation in results across exercises is not due to inaccurate judgments of candidate behaviour or to lack of construct validity, but instead occurs because candidates' behaviour changes in response to different situations. In addition, Hoeft and Schuler (2001) found that candidates' performances were more situation specific (57%) than situation consistent (43%). Neidig and Neidig (1984) reported similar findings and suggested that instead of blaming assessor ratings for the lack of construct validity, researches should focus on candidates' real performance differences across situations. This argument is supported by the fact that different exercises are designed to carefully uncover job-related competencies that place different psychological demands on the candidates (Lievens, 2009). AC exercises are designed to elicit behaviour, skills and abilities related to specific job tasks and it is therefore likely that candidates will perform better in some exercises than in others (Arthur *et al.*, 2008).

The AC industry faces two distinct lines of thought concerning the longstanding construct validity debate. The first line of thought views exercise effects as a serious threat to the construct validity of AC ratings. In contrast, the situation-specific interpretation regards exercise effects as a reflection of true cross-situational specificity of (relevant) performance in the AC across different exercises and suggests that these effects should be included in the design, interpretation and scoring of ACs (Hoffman & Baldwin, 2012). However, a new line of research has recently emerged with the proposal of a mixed-model AC design focusing on AC behaviour and its determinants (Lievens & Christiansen, 2010). In this model exercises are viewed as behaviour-triggering situational indicators or cues and dimensions are seen as conditional dispositions. Borman (2012) indicated that the mixed-model approach assumes an interactionist position, suggesting that behaviour in ACs is a function of both individual differences in behavioural tendencies and situational influences on behaviour. According to this model individual differences interact with exercise influences and demands, resulting in behaviour relevant to both task and dimension. According to Lance (2012), the mixed-model perspective is increasing in popularity, as it acknowledges the importance of both dimension and exercise information and tries to assign appropriate weight to each of these factors (Melchers, Wirz & Kleinmann, 2012).

Context of the present study – Development assessment centre for banking personnel

The present study aimed to address the scarcity of South African AC research by examining the construct validity of the performance dimension measured in a development assessment centre (DAC). The DAC investigated in this study was developed for a South African banking institution. In the 2010 World Competitiveness Report, the South African banking sector was rated first out of the 139 countries that participated in the study. South Africa has a developed and well-regulated banking system which compares favourably with the banking systems of industrialised countries (Banking Association South Africa, 2010). However, despite factors such as advanced technology, deregulation and globalisation that are causing a revolution in the financial services industry,

mergers and consolidations as well as demanding customers require South African banking companies to concentrate on the strategic value and competitive advantage of their employees (Kock, Roodt & Veldsman, 2002) as a means to address challenges exclusive to the sector.

The nature of the financial industry, which deals with products and services that are complicated, risky and of a long-term nature, results in customers being in a high involvement relationship with their financial service providers (Howcroft, Hewer & Durkin, 2003). In addition, customers have high expectations regarding banks' service delivery. In order to succeed, banks have to adopt proactive approaches to maintain standards of service delivery (Ackermann & Van Ravesteyn, 2006). If banks are to achieve their goal of excellent customer service they need to attract and retain high-quality employees who will deliver exceptional service.

The one-day DAC investigated in the present study was used to assess employees who work in the welcoming zone of the bank. Their personal skills when dealing with clients entering the bank was measured in an attempt to improve the quality of customer service delivered by these employees. The DAC had three main purposes: firstly, to identify candidates who fit the role of a new job position, secondly, to reposition the remaining employees into more appropriate roles and, thirdly, to provide a development experience for all participants taking part in the centre. Each participant received an individual development report as guidance for long-term growth.

Research design

Research approach

The present study falls within the quantitative research paradigm. A non-experimental research design was used focusing on an ex post facto correlation design. Results obtained from a competency-based AC were used as a level of measurement of participants' proficiency on predetermined behavioural dimensions.

Research strategy

In the first phase of the research strategy, an in-depth literature review on the construct validity of ACs was conducted. Subsequently a priori hypotheses regarding the construct validity of an existing DAC were examined with a quantitative correlation design.

Research method

Participants and sampling

The data were drawn from a DAC utilised by a banking institution to assess 202 branch managers. The participants were both male and female. These participants were formally invited to join the assessment process as part of a large organisational change strategy. The researcher did not have any control over the size and characteristics of the sample. The sample can therefore be characterised as a convenience sample. Due to the confidentiality agreement between the researcher and the consulting agency that provided the AC data, no information regarding the sample's demographic characteristics or identities was included in the analysis in order to maintain the participants' confidentiality. Consequently, no further demographic information regarding the sample group is disclosed in this study.

Measuring instruments

Information derived from a job analysis was combined with input from the client organisation and subject experts to arrive at five primary dimensions that were conglomerated into one global factor: leadership skills. Each participant was assessed on the four primary factors using three exercises (illustrated in Table 1).

Dimensions: The primary dimensions assessed (see Table 1) included utilisation and development, task structuring, impact and conflict resolution and sensitivity. These served as primary factors of the global leadership skills dimension.

Exercises: Three exercises were used to measure the related dimensions: Analysis Problem, One-on-one Interview and Group Discussion:

> **Exercise 1:** Analysis problem service improvement plan
> This exercise required the participant to review a large amount of information concerning the branch and then make formal recommendations as to how to improve the service levels and effectiveness.
>
> **Exercise 2:** Service improvement team group discussion
> In this exercise participants were divided into service improvement teams of 10. Each participant was given the opportunity to act as team leader when presenting their recommendations and interventions to improve service delivery.
>
> **Exercise 3:** One-on-one conflict resolution and coaching interview
> In this scenario each participant had to meet one-on-one with an irritated customer and try to resolve the conflict and enhance the customer service experience.

Two different values were used to indicate the reliability of the exercises: Cronbach's alpha and the reliability coefficient rho. The overall values for the exercises were 0.78 and 0.83 (Brits, 2011).

Assessors: Selected assessors and administrators were trained. Two groups of assessors and administrators were trained by the consultant agency over a period of five days to observe, administer and conduct the assessments. The training included thorough discussions on the five-point Likert scale to be used as rating scale for all exercises as well as all competencies and their behavioural indicators, instructions to facilitators and in-depth group discussions

TABLE 1: Behavioural matrix: Leadership skills.

Dimension	Analysis problem	Group discussion	One-on-one	Final score
Utilisation and development	X	X	X	X
Task structuring	-	X	X	X
Impact and conflict resolution	-	X	-	X
Sensitivity	X	X	X	X

about the simulations As part of the training, all assessors and administrators had to complete all of the DAC exercises themselves. Assessors used a behavioural matrix (see Table 1) to evaluate each participant's performance on the relevant dimensions.

Data collection procedure: The data were received from a private consultant company in the form of the AC ratings of 202 individuals who were assessed in a one-day DAC. A convenient sample was used to pursue the listed research objectives.

Statistical analysis

Structural equation modelling (SEM) based on EQS (6.1) (Bentler, 2005) was used to test the correspondence (fit) of the proposed leadership dimension measurement model with the empirical data. As with other multivariate linear statistical procedures, CFA requires that certain assumptions must be met with regard to the sample. Therefore, prior to formally fitting the CFA model to the data, the assumptions of multivariate normality, linearity and adequacy of variance were evaluated. In general, no serious violations of these assumptions were detected in the data. However, the data did not follow a multivariate normal distribution and therefore robust maximum likelihood (ML) was specified as the estimation technique. Respondents with extreme scores ($z > 3.00$) were removed from the dataset and missing values were estimated with the ML estimation technique.

The CFA was used to gain unconfounded estimates of exercise and dimension variance in the AC ratings. This technique allows for the disentanglement of exercise and dimension variance because CFA partitions the variance into dimension, exercise and error variance components. Although critics of the CFA technique argue that it results in an overly simplistic view of AC functioning as it only refers to variance caused by exercises and dimensions (Bowler & Woehr, 2009), a substantial amount of AC research over the last 20 years has been based on the application of this technique. According to Maas, Lansvelt-Mulders and Hox (2009), the CFA technique offers many opportunities for the examination of AC construct validity due to its flexibility and statistical rigour.

The present study was based on the assumption that each dimension loads predominantly on one trait and one method factor, and that the covariances between trait and method factors are zero. Prima facie claims of construct validity would therefore be tenable when dimension-related factor loadings exceed factor loadings associated with the single dimension exercise effect. In addition, the change in R^2 values between the main effect model and the bi-factor model (i.e. including the single method effect) informed the degree to which the main leadership dimension or the single method effect explains proportionally more variance in overall AC ratings. From a methodological point of view the bi-factor model can be regarded as a nested variation of a main effect model fitted in consecutive steps analogous to stepwise hierarchical regression. Figure 1 graphically depicts the bi-factor CFA model that was applied to the empirical data.

Several fit indices were used to assess the amount of congruence between the proposed bi-factor CFA model and the empirical data. The following prominent fit indices were utilised to evaluate the tenability of the proposed theoretical model: Satorra-Bentler scaled chi-square statistics, root mean square error of approximation (RMSEA), the root mean square residual (RMR), and the comparative fit index (CFI). Robust maximum likelihood estimation (RML) was used to specify the model.

The two main objectives of the study would be supported if:

- The basic CFA model containing only the single factor leadership dimensions fitted the empirical data satisfactorily. More specifically, the fit indices and model parameters reflected a well-fitting model.
- The completely standardised factor loadings related to the single dimension leadership dimension exceeded the factor loadings related to the single factor method effect. In addition, the model R^2 should not increase significantly when the single dimension method effect is included in the main effect leadership model.

If these conditions were satisfied, prima facie evidence of construct validity could be assumed. However, these results would not suggest that the proposed model is the best possible fitting model and alternative model configurations should be investigated.

Results

Confirmatory factor analysis (CFA) was performed in EQS (6.1) on the hypothesised leadership dimension measurement model. CFA focuses on how, and the extent to which, the observed variables are linked to their underlying latent factors (Byrne, 2006). The measurement model (see Figure 1) describes how each variable is operationalised by corresponding manifested indicators and provides information about the validity and reliability of the observed indicators (Diamantopoulos & Sigauw, 2000). More specifically, measurement model fit refers to the extent to which a hypothesised model is consistent with or explains the data. A number of different fit indices exist that can be used to evaluate model fit.

The tenability of CFA models is assessed on both global (via fit indices) and molecular (via model parameters) levels of

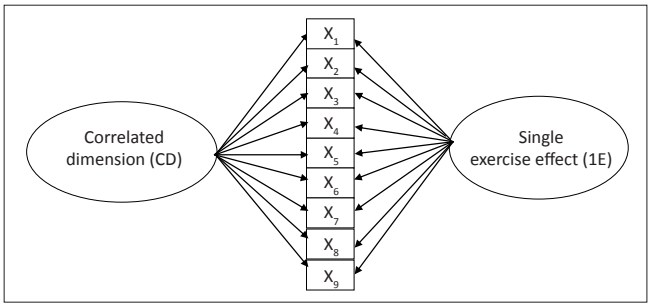

FIGURE 1: Leadership skills (1D1E model).

observation (Kline, 2011). The validity of the single factor leadership dimension model was initially assessed on a global level through the examination of various fit indices, specifically Satorra-Bentler scaled chi-square statistics, RMSEA, RMR and standardised RMR (SRMR) and model residuals. A summary of the fit indices is provided in Table 2.

A statistically significant chi-square leads to the rejection of the null hypothesis and implies an imperfect model fit. The aim is consequently not to reject H_0 (Diamantopoulos & Siguaw, 2000). The Satorra-Bentler scaled chi-square (p = 0.00) indicated that the model was not reproducing the data perfectly. Consequently, the null hypothesis of exact fit was rejected (p = 0. 00).

The RMSEA expresses the difference between the observed and estimated sample covariance matrices and expresses the population discrepancy function value in terms of the degrees of freedom of the model. This is a measure of closeness of fit. RMSEA values less than 0.05 indicate a good fit and RMSEA over 0.08 indicate a reasonably good fit. RMSEA values between 0.08 and 0.1 indicate an average fit and values greater than 0.10 indicate a poor fit (Diamantopoulos & Siguaw, 2000). A RMSEA value of 0.128 (see Table 2) was obtained from the data, which illustrates that the model does not fit the data well. The 90% confidence interval for RMSEA suggests that the true RMSEA value for the population falls between 0.105 and 0.152, which is a mediocre degree of precision. The upper bound of the confidence interval exceeds the critical cut off value of 0.05 and therefore resulted in the rejection of the null hypothesis of close fit.

The RMR is the square root of the mean of the squared discrepancies between the implied and observed sample covariance matrices (Oehley & Theron, 2010). Standardised residuals are often interpreted to avoid problems relating to unstandardised residuals that may vary with the unit of measurement (Diamantopoulos & Siguaw, 2000). An SRMR with values under 0.05 indicates that the data fits the theoretical model with acceptable levels of discrepancy (Oehley, 2007). RMR values ranging between 0.05 and 0.08 are indicative of good fit. The RMR reported for the leadership skill measurement model was 0.120, which did not fall within the acceptable range indicative of good fit.The same value (0.120) was reported for the SRMR, which is the standardised index of the RMR.

The last index to be considered is the CFI. This index measures the proportionate improvement in fit by comparing a hypothesised model with a more restricted, nested baseline model (Byrne, 2006). The CFI assumes a baseline model in which all latent variables are uncorrelated. Values typically range between 0 and 1 (Hooper, Coughlan & Mullen, 2008), with higher values (> 0.95) generally considered indicative of a well-fitting model. The CFI reported for the leadership dimension measurement model is 0.855, which is less than the cut-off value of 0.95. Considered collectively, these results suggest that the single factor leadership dimension does not fit the model well. However, as the model fit did not deviate significantly from the normative guidelines, the authors looked for further corroborative evidence by examining specific model parameters. Table 3 contains the completely standardised lambda factor loadings of the single factor leadership dimension model.

Kline (1999) suggested that completely standardised factor loadings should be statistically significant and range from 0.50 to 0.70 since standardised lambda loadings must be squared in order to express the proportion of variance in the indicator variables that can be explained by each dimension constituting the DAC. For example, a standardised factor loading of 0.71 squared equates to 0.50. Thus, 50% of the variance reflected in the specific indicator is due to the latent variable. The remaining 50% of the variance that is left unexplained can obviously be attributed to the influence of systematic and unsystematic sources of extraneous variance. When lambda loadings fall below 0.70, more than half of the variance in the measure is due to error variance (systematic and random). Standardised loading estimates of 0.50 and higher are viewed as acceptable (Becker, 2009) and were considered sufficiently large for the purposes of the current study. When this criterion was strictly adhered to, five of the nine indicators failed to meet the minimum lambda loading criteria (indicated in Table 3).

Based on the results discussed above it was concluded that the single factor leadership CFA model did not fit the data well. Both the global and molecular statistical indices suggested that the discrepancy between the sample covariance matrix and the reproduced matrix was substantial. However, although the basic CFA model did not fit the data well, a

TABLE 2: Goodness of fit statistics for the 1D measurement model.

Leadership skills dimension	Result
Degrees of freedom	27
Satorra-Bentler scaled chi-square	116.4625 (p = 0.00)
Root mean square error of approximation (RMSEA)	0.128
90% confidence interval for RMSEA	(0.105, 0.152)
Independence Akaike information criterion (AIC)	580.670
Model AIC	62.462
Independence consistent AIC	425.582
Model consistent AIC	-53.861
Comparative fit index (CFI)	0.855
Root mean square residual (RMR)	0.120
Standardised RMR	0.120

TABLE 3: Confirmatory factor analysis model parameters of 1D measurement model.

Item	Standardised lambda factor loadings (λy)
LS_UD_AP_V1	0.499
LS_UD_GD_V2	0.831
LS_UD_ONE_V3	0.276
LS_TS_GD_V4	0.750
LS_TS_ONE_V5	0.303
LS_ICR_GD_V6	0.717
LS_SEN_AP_V7	0.368
LS_SEN_GD_V8	0.537
LS_SEN_ONE_V9	0.343

λy, Lambda; LS, leadership skills; UD, utilisation development subscale; AP, analysis problem exercise; GD, group discussion; ONE, one-on-one exercise; TS, task structuring subscale; ICR, impact and conflict resolution subscale; SEN, sensitivity subscale.

nested variation of this model, namely the 1D1E model, was specified to determine whether any meaningful improvement in model fit occurred. The researchers posed the following question: Does the 1D1E model provide a better account of the empirical data than the single leadership dimension CFA model?

A summary of the most important fit indices for the 1D1E model (single dimension single exercise leadership dimension model) is presented in Table 4. The Satorra-Bentler scaled chi-square ($p > 0.05$) indicates that the model fitted the empirical data exactly. The null hypothesis of exact fit could consequently not be rejected.

The RMSEA value of 0.030 falls below the critical cut-off value of 0.05 and therefore H_0: RMSEA $\leq$ 0.05 could not be rejected. The model was therefore regarded as fitting the empirical data well. The 90% confidence interval used for RMSEA (0.000 to 0.072) instilled further confidence in the assumption that the true RMSEA value in the population falls between the bounds of 0.000 and 0.072.

The RMR reported for the leadership skill measurement model boasted a value of 0.051, which falls within the acceptable range indicative of good fit. The same value (0.051) was reported for the SRMR, which is the standardised index of the RMR.

The CFI reported for the leadership Skill dimension 1D1E model (0.995) is indicative of a well-fitting model, implying good fit of the theoretical model to the empirical data.

Considered collectively, the fit indices for the 1D1E model suggest that the model fits the data relatively well. In order to further assess the tenability of the model, specific model parameters were examined. Table 5 provides a summary of the completely standardised factor loadings relevant to the 1D1E model.

Table 6 shows that although the majority of the factor loadings pertaining to the single leadership dimension factor were robust, strong loadings were also reported for the single exercise dimension. In order to investigate the amount of additional unique variance accounted for by the exercise effect over and above the single dimension effect, it was necessary to investigate the squared multiple correlations (R^2 communalities).

A high squared multiple correlation was indicative of the dimension or exercise explaining a substantial amount of true variance in the indicator (Moyo, 2009). In line with findings reported in previous research, the results of this study found that the squared multiple correlation values increased substantially with the inclusion of the exercise effect in the bi-factor model of leadership. This implies that considering factor loadings and communality values in a single dimension model may mask at least some of the variance attributable to the method effect. The one-on-one exercises (role-plays) appeared to reflect the most exercise effect, probably due to the bias inherent in the exercise methodology. With the exception of the analysis problem exercises there was an increase in the amount of variance explained for all the exercises. This is somewhat to be expected since the mixed-model approach to AC design presumes that at least some of the variance in dimension ratings will be due to the method effects in which the dimensions are framed. These results are explained in more detail in the subsequent section.

TABLE 4: Goodness of fit statistics for the 1D1E measurement model.

Leadership skills dimension	Result
Degrees of freedom	18
Satorra-Bentler scaled chi-square	21.1506 (p = 0.27188)
Root mean square error of approximation (RMSEA)	0.030
90% confidence interval for RMSEA	(0.000, 0.072)
Independence Akaike Information Criterion (AIC)	580.679
Model AIC	-14.849
Independence consistent AIC	425.582
Model consistent AIC	-92.398
Comparative fit index (CFI)	0.995
Root mean square residual (RMR)	0.051
Standardised RMR	0.051

TABLE 5: Completely standardised factor loadings of 1D1E measurement model.

Item	Dimension	Exercise
LS_UD_AP_V1	0.419	0.264
LS_UD_GD_V2	0.622	0.551
LS_UD_ONE_V3	0.654	0.388
LS_TS_GD_V4	0.525	0.568
LS_TS_ONE_V5	0.675	0.347
LS_ICR_GD_V6	0.533	0.490
LS_SEN_AP_V7	0.265	0.252
LS_SEN_GD_V8	0.402	0.375
LS_SEN_ONE_V9	0.599	0.213

LS, leadership skills; UD, utilisation development subscale; AP, analysis problem exercise; GD, group discussion; ONE, one-on-one exercise; TS, task structuring subscale; ICR, impact and conflict resolution subscale; SEN, sensitivity subscale.

TABLE 6: Squared multiple correlations (R^2) of indicators.

Item	1D model	1D1E model	ΔR^2
	Dimension effect size (R^2)	Method effect size (R^2)	
LS_UD_AP_V1	0.249	0.245	-0.004
LS_UD_GD_V2	0.691	0.691	-
LS_UD_ONE_V3	0.076	0.579	0.503
LS_TS_GD_V4	0.562	0.598	0.036
LS_TS_ONE_V5	0.092	0.552	0.460
LS_ICR_GD_V6	0.514	0.524	0.01
LS_SEN_AP_V7	0.135	0.134	-0.001
LS_SEN_GD_V8	0.289	0.289	-
LS_SEN_ONE_V9	0.118	0.404	0.289

LS, leadership skills; UD, utilisation development subscale; AP, analysis problem exercise; GD, group discussion; ONE, one-on-one exercise; TS, task structuring subscale; ICR, impact and conflict resolution subscale; SEN, sensitivity subscale.

Discussion

Summary of the results

The dimensions versus exercises debate remains an important on-going theme in the AC literature. The current investigation aimed to contribute to the debate by investigating results from a typical DAC implemented and developed in the banking sector in the South African economy. The research study was designed to answer the

following question: 'Do AC ratings predominantly reflect dimension or exercise effects?'. The study found that in the absence of the single exercise effect the single dimension leadership measurement model seemed to fit the data well when fit indices as well as the completely standardised factor loadings were considered. However, when the single factor dimension effect was specified along with the single dimension effect in the form of a bi-factor CFA model, the single exercise effect appeared to account for non-negligible proportions of the true variance. The pattern of results was quite clear: role-playing type exercises reflected large proportions of method effects, whilst analysis problem type exercises predominantly reflected the dimensions effects. It is likely that a similar pattern exists for the group discussion exercises, which predominantly reflected dimension effects.

Possible reasons for the significant exercise effect found in the one-on-one exercise include candidates' perception of this type of interactive exercise, and the effect that role-players and even assessors have on candidates' true performance. Role-playing exercises are more susceptible to bias ratings from raters. The fact that the majority of the variance in role-playing exercises is attributable to method effects suggests that the true performance of candidates on the dimension accounts for less variance than the specific method that is being used, in this case role-plays. In current AC practice, role-players are trained to perform realistically and consistently across candidates in order to evoke behaviour from candidates (Thornton & Mueller-Hanson, 2004). The One-on-one form of exercise is widely used within ACs. In their review of about 500 operational assessment centres Thornton and Byham (1982) reported that the interview simulation (the One-on-one exercise in the current study) was used in 75% of the centres. In this specific role-play exercise candidates worked alone with an associate who was trained to play a standardised role (Schneider & Schmitt, 1992). This is similar to the scenario candidates faced in the present DAC's One-on-one exercise.

Schneider and Schmitt (1992) also found that the type of exercise (i.e. exercise form) was the most significant exercise factor that contributed to candidates performing differently across exercises. So-called exercise or method effects are therefore not necessarily caused by the measurement of invalid constructs but instead by the trend of ratings assigned by assessors to role-play behaviour or by the fact that people's actions and behaviour vary across situations, depending on both personal and situational variables (Lievens, 2002; Lievens & Christiansen, 2010). The present study's findings seem to corroborate the results reported by Arthur *et al*. (2000), Rupp *et al*. (2008) and Hoffman and Meade (2012), which suggest that the mixed-model AC approach may be the most plausible explanation for candidate behaviour in ACs. The mixed-model perspective suggests that elements of role behaviour specific to both dimensions and exercises are represented in AC ratings. It is therefore possible to conclude that both dimensions and exercises (which elicit dimension-relevant behaviour) should be acknowledged in the design, scoring, interpretation and reporting of ACs (Hoffman & Baldwin, 2012).

It appears to be time to acknowledge that, in accordance with Walter Mischel's (1968) explanation of human behaviour, both dimensions and exercises are the currency of ACs (Hoffman & Baldwin, 2012). Candidate behaviour in ACs should thus be conceptualised in terms of a recent interactionist theory such as Trait Activation Theory (TAT) (Lievens, Tett & Schleicher, 2009; Tett & Burnett, 2003), which explains behaviour as responses to trait-relevant cues found in situations (Tett & Burnett, 2003).

According to this theory, situation trait relevance and situation strength are both important factors in understanding the situations in which a trait is likely to manifest itself in behaviour. A situation is considered relevant to a trait if it provides cues for the expression of trait-relevant behaviour (Tett, Guterman, Bleier & Murphy, 2000). Situation strength is conceptualised as existing on a continuum that relates to how much clarity exists with regard to the way in which situations are perceived. Strong situations involve unambiguous behavioural demands and are therefore likely to contradict almost all individual differences in behaviour without regard to any specific traits. Conversely, weak situations are characterised by more ambiguous expectations, enabling more variability in behavioural responses to be observed (Meyer, Dalal & Hermida, 2010).

Trait Aactivation Theory is relevant to AC exercises because it highlights the importance of building multiple stimuli into the AC exercises. These exercises can thus be explicitly designed to increase their situation trait relevance in order to increase behaviour observability (Lievens & Schollaert, 2011). These authors suggest that the use of situational stimuli to elicit a higher number of behaviours in AC exercises also results in dimensions being better measured in AC exercises. This could have an advantageous effect on the construct validity of AC exercises.

Recommendations

The first set of recommendations concerns the design phase of ACs. Woehr and Arthur (2003, p. 251) summarised this perspective by noting that 'assessment centres as measurement tools are probably only as good as their development, design and implementation'. Focused attention should be paid to the definition of dimensions. Merely labelling data as a reflector of a particular construct does not mean that this construct is actually being assessed (Collins *et al*., 2003). According to Van der Bank (2007), the nature of the relationship between the competency chosen and job outcome should ideally be tested and proven in structural equation competency models. Despite the high premium placed on competencies in ACs, researchers have not given sufficient attention to models reflecting the relationship between competencies and results (Van der Bank, 2007). This may be a result of practitioners' desires to satisfy the specific needs of their client organisations, such that they prefer to simply alter and label espoused constructs that will meet the client's requirements instead of redesigning actual constructs with the necessary theoretical evidence of

construct presentation (Arthur *et al*., 2008). In many instances the problem starts when practitioners inherit an established, generic competency model and fail to challenge or investigate the inherent rationale of this model.

In relation to dimensions, Hoffman and Baldwin (2012) recommended not only personalising the dimensions for the client organisation and relevant job, but also personalising the exercises used to elicit these dimensions. Exercises that generate sufficient behavioural evidence to measure a particular dimension should be developed and statistically validated in a careful manner. If this process is not followed the exercises should be excluded from the AC in order to prevent any unnecessary cognitive demands being placed on assessors (Greyling *et al*., 2003).

In relation to the design phase of ACs it is also important that the number of dimensions be taken into consideration (Lievens, 2009). It is recommended that a small number of dimensions (Campbell & Fiske, 1959; Iles, 1992; Krause, 2010) with a large number of short exercises be measured in order to run comprehensive statistical analyses on AC data and to generate more evidence of behaviour before assigning a final dimension rating (Lievens *et al*., 2009). Brannick (2008) recommended using five six-minute role-plays instead of a single 30-minute role-play. This would enable the generation of samples of performance on a large number of independent tasks exclusively designed to elicit behaviour related to a specific dimension. A further recommendation is that when interpersonal exercises are used, role-player cues or prompts could be used to serve as additional means of eliciting job-related behaviour. Prompts are defined as predetermined verbal and non-verbal cues that a trained role-player consistently provides during AC exercises to elicit job-related behaviour (Schollaert & Lievens, in press). Through using prompts (which could be based on TAT), a situational stimulus for evoking behaviour could be created, thus increasing candidates' opportunities to demonstrate dimension-related behaviour and assessors' opportunities to observe this behaviour. Lievens and Schollaert (2011) found that the use of prompts led to greater observablity of behaviour, which, in turn could lead to higher levels of consistency in candidate behaviour and therefore greater discriminant and convergent validities. For example, Schollaert and Lievens (in press) found that Problem-solving and Interpersonal Sensitivity dimensions were better measured in AC exercises when role-players used prompts designed to evoke these dimensions. Their results showed that construct-related validity (convergent and discriminant correlations) was highest when role-players used prompts for eliciting behaviour and when assessors were familiar with these prompts (Lievens & Schollaert, 2011).

Other recommendations for increasing the situation trait relevance of exercises, specifically within the present context of one-on-one exercises, include adapting the content of the exercise as well as emphasising the instructions that provide information and expectations to candidates about what behaviour to show or not show. Based on the recommendations discussed above it is safe to conclude that it would be beneficial to AC research and practice if future researchers endeavour to design ACs in such a way as to ensure that AC exercises activate the desired behavioural traits that align with the relevant job competencies as derived from a thorough job analysis and, ultimately, the compilation of a comprehensive competency model.

Limitations of the study

The most significant limitation of the present study is that only one of the initial five global dimensions was considered acceptable for statistical analysis. Ideally the researchers would have been able to study the effect of all the individual exercises on all five global dimensions, as well as on all 12 sub-dimensions' ratings. Due to the lack of sufficient indicators it was not possible to investigate discriminant and convergent validity. Due to the hierarchical structured nature of applied ACs data, it is often impossible to investigate the psychometric properties of dimensions due to the clinical integration of dimension ratings into overall ratings. Practitioners should aim to capture behavioural indicators at each hierarchical level when assessing candidates on ACs and DACs (i.e. behavioural indicator level, dimension level and across exercise level). This will enable researchers to use sophisticated multivariate statistical techniques to investigate the accuracy of ratings at various levels of observation. Time and financial constraints endemic to DAC development are admittedly major contributing factors in this regard (Lievens & Conway, 2001). As a result of these constraints empirical under-identification remains a serious methodological limitation in AC research (Lance, Lambert, Gewin, Lievens & Conway, 2004) and was also a limitation of this study.

Another limitation of the present study concerns the generalisation of the findings. The DAC data was generated from a singular sample group. Results can therefore not be generalised with great certainty. Ideally a second sample should be assessed using the same DAC in order to compare validity results.

Proposed for future research

Firstly, the results of the present study should serve as impetus for South African practitioners and researchers to focus their attention on new developments and research findings concerning the mixed-model approach. In addition, practitioners and researchers need to conduct further research concerning the AC method. Despite the important groundwork provided by the assessment centre study group, a gap remains within the South African literature concerning the validity of ACs. The present study is only the second South African study that has focused on the construct validity of ACs. Repeated low construct validity findings, which have been reported regardless of the statistical technique applied (multitrait-multimethod correlations, EFA, CFA, analysis of variance, variance component analysis, etc.), suggest that further investigation of the construct validity of ACs is justified (Lievens & Christiansen, 2010). Perhaps it is time for researchers and practitioners to not only start asking

new and different questions about the internal structure of ACs but also experiment with the mixed-model approach in designing ACs.

The second recommendation also concerns the use of the mixed-model approach. A need exists for a more sophisticated understanding of how person variables, and the dimensions tapped by these variables, interact with situational factors. Future research should aim to use interactionist theories such as the TAT in a proactive and prescriptive way to change and improve AC practices (Lance, 2008a; Lievens & Christiansen, 2010).

Thirdly, researchers are encouraged to use variance partitioning to analyse AC data. Variance partitioning, also known as the generalisability theory, examines the different sources of variance associated with AC dimension ratings and estimates the relative impact that each source has on the ratings (Bowler & Woehr, 2009). According to Borman (2012), the main advantage of using the multivariate case is that it allows for non-zero correlations between dimensions, which is normally a realistic expectation for AC data. Bowler and Woehr (2008) indicated that this method is well suited for examining the construct validity of ACs since it generates results that are representative of the population data and it acknowledges sources of variance that cannot be assessed with the traditional CFA of a multitrait-multimethod matrix. In their study on the effect of using generalisability theory to examine AC ratings, Bowler and Woehr (2009) noted that the single largest source of variance in the ratings was person by exercise interaction, which suggests the presence of differential patterns of performance across exercises. This corroborates the idea that exercises are a systematic cause of AC performance.

Conclusion

This study was the second study conducted in South Africa focusing on the construct validity of an AC. The results of this study concur with the results found in international studies. The literature shows that new theoretical, empirical and even philosophical issues have been introduced to the debate on the construct validity of ACs. The stage is therefore set for productive discussions and debates around new models of ACs and analysis strategies. South African researchers and practitioners should rise to the challenge of exploring the new mixed-model approach to designing ACs.

Acknowledgement

The authors would like to thank Experiential Technologies for providing the data sample that was used in this study.

Competing interests

The authors declare that they have no financial or personal relationship(s) that may have inappropriately influenced them in writing this article.

Authors' contributions

N.M.B. (Stellenbosch University) conducted the research as part of her Master's degree dissertation, prepared the data for analysis and was responsible for the majority of the article writing. D.M. (University of Pretoria) acted as study leader and promoter and reviewed, made recommendations on and presented the final article. J.R.B. (University of Johannesburg) assisted with the data analysis and the writing of the article, focusing specifically on the empirical results and interpretation thereof.

References

Ackermann, P.L.S., & Van Ravesteyn, L.J. (2006). Relationship marketing: The effect of relationship banking on customer loyalty in the retail business banking industry in South Africa. *Southern African Business Review, 10*(3), 149–167.

Arthur, W.A., Day, E.A., & Woehr, D.J. (2008). Mend it, don't end it: An alternative view of assessment centre construct-related validity evidence. *Industrial and Organisational Psychology: Perspectives on Science and Practice, 1*(1), 105–111. http://dx.doi.org/10.1111/j.1754-9434.2007.00019.x

Arthur, W.A., Woehr, D.J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. *Journal of Management, 26*(4), 813–835. http://dx.doi.org/10.1016/S0149-2063(00)00057-X, http://dx.doi.org/10.1177/014920630002600410

Banking Association South Africa. (2010). *South African banking sector overview*. Retrieved December 7, 2010, from http://www.banking.org.za

Becker, J.R. (2009). *Influence of values on the attitude towards cultural diversity.* Unpublished master's dissertation, Stellenbosch University, Stellenbosch, South Africa.

Bentler, P.M. (2005). *EQS 6 structural equations program manual.* Retrieved March 7, 2011, from http://www.mvsoft.com

Borman, W.C. (2012). Dimensions, rask and mixed models: An analysis of the three diverse perspectives on assessment centers. In D. Jackson, C. Lance, & B. Hoffman (Eds.), *The psychology of Assessment Centres* (pp. 309–320). New York, NY: Routledge.

Bowler, M.C., & Woehr, D.J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91*(5), 1114–1124. http://dx.doi.org/10.1037/0021-9010.91.5.1114, PMid:16953772

Bowler, M.C., & Woehr, D.J. (2008, April). Evaluating assessment center construct-related validity via variance partitioning. In B.J. Hoffman (Chair), *Reexamining assessment centers: Alternate approaches*. Symposium conducted at the 23rd annual meeting of the Society for Industrial and Organisational Psychology, San Francisco, CA.

Bowler, M. C., & Woehr, D.J. (2009). Assessment center construct-validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior, 74*(2), 173–182. http://dx.doi.org/10.1016/j.jvb.2009.03.008

Brannick, M. T. (2008). Back to basics of test construction and scoring. *Industrial and Organisational Psychology: An Exchange on Science and Perspectives, 1*, 131-133.

Brits, N.M. (2011). *An explorative investigation into the construct validity of a development assessment centre.* Unpublished master's thesis, University of Stellenbosch. Retrieved June 21, 2012, from http://www.scholar.sun.ac.za

Byrne, B.M. (2006). *Structural equation modelling with EQS: Basic Concepts, Applications and Programming*. (2nd edn.). Mahwah, NJ: Lawrence Erlbaum Associates. http://dx.doi.org/10.1207/s15328007sem1302_7

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. http://dx.doi.org/10.1037/h0046016, PMid:13634291

Collins, J.M., Schmidt, F.L., Sanchez-Ku, M., Thomas, L., McDaniel, M.A., & Le, H. (2003). Can basic individual differences shed light on the construct meaning of assessment center evaluations? *International Journal of Selection and Assessment, 11*(1), 17–29. http://dx.doi.org/10.1111/1468-2389.00223

Crawley, B., Pinder, R., & Herriot, P. (1990). Assessment centre dimensions, personality and aptitudes. *Journal of Occupational Psychology, 63*(3), 211–216. http://dx.doi.org/10.1111/j.2044-8325.1990.tb00522.x

Diamantopoulos, A., & Siguaw, J.A. (2000). *Introducing LISREL*. London, UK: Sage Publications. PMCid:PMC1298088

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C., & Bentson, C. (1987). Meta-analysis of Assessment Center validity. *Journal of Applied Psychology Monograph, 72*(3), 493–511. http://dx.doi.org/10.1037/0021-9010.72.3.493

Greyling, L., Visser, D., & Fourie, L. (2003). Construct validity of competency dimensions in a team leader assessment centre. *SA Journal of Industrial Psychology, 29*(2), 10–19. http://dx.doi.org/10.4102/sajip.v29i2.97

Harris, M.M., Becker, A.S., & Smith, D.E. (1993). Does the assessment center scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology, 78*(4), 675–678. http://dx.doi.org/10.1037/0021-9010.78.4.675

Hoeft, S., & Schuler, H. (2001). The conceptual basis of assessment centre ratings. *International Journal of Selection and Assessment, 9*(1/2), 114–123. http://dx.doi.org/10.1111/1468-2389.00168

Hoffman, B.J., & Baldwin, S.P. (2012, April). *Invariance analyses support a multifaceted interpretation of assessment centers.* Paper presented at the 27th Annual Society for Industrial and Organisational Psychology Conference, San Diego, CA.

Hoffman, B.J., & Meade, A. (2012). Alternate approaches to understanding the psychometric properties of assessment centres: An analysis of the structure and equivalence of exercise ratings. *International Journal of Selection and Assessment, 20*(1), 82–97. http://dx.doi.org/10.1111/j.1468-2389.2012.00581.x

Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods, 6*(1), 53–50.

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

Howcroft, B., Hewer, P., & Durkin, M. (2003). Banker-customer interactions in financial services. *Journal of Marketing Management, 19*(9), 1001–1020. http://dx.doi.org/10.1362/026725703770558295, http://dx.doi.org/10.1080/0267257X.2003.9728248

Huysamen, G.K. (1996). *Methodology for the social and behavioural sciences.* Halfway House, South Africa: Thomson.

Iles, P. (1992). Centres of excellence? Assessment and development centres, managerial competencies and human resource strategies. *British Journal of Management, 3*(2), 79–90. http://dx.doi.org/10.1111/j.1467-8551.1992.tb00037.x

Jackson, D.J.R., Atkins, S.G., & Stillman, J.A. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance, 18*(3), 213–241. http://dx.doi.org/10.1207/s15327043hup1803_2

Jackson, S.E., Barney, A.R., Stillman, J.A., & Kirkley, W. (2007). When traits are behaviours: The relationship between behavioural responses and trait-based overall assessment center ratings. *Human Performance, 20*(4), 415–432.

Kline, P. (1999). *The handbook of psychological testing*. (2nd ed.). London, UK: Routledge.

Kline, R.B. (2011). *Principles and practice of structural equation modelling*. (3rd ed). New York, NY: The Guilford Press.

Kock, R., Roodt, G., & Veldsman, T.H. (2002). The alignment between effective people management, business strategy and organisational performance in the banking and insurance sector. *SA Journal of Industrial Psychology, 28*(3), 83–91. http://dx.doi.org/10.4102/sajip.v28i3.66

Krause, D.E. (2010, March). *State of the art of assessment centre practices in South Africa: Survey results, challenges, and suggestions for improvement.* Paper presented at the 30th ACSG Conference, Stellenbosch, South Africa.

Krause, D.E., Kersting, M., Heggestad, E.D., & Thornton, C.G. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment, 14*(4), 360–371. http://dx.doi.org/10.1111/j.1468-2389.2006.00357.x

Lance, C.E. (2008a). Where have we been, how did we get there and where shall we go? *Industrial and Organisational Psychology: Perspectives on Science and Practice, 1*(1), 140–146.

Lance, C.E. (2008b). Why assessment centres do not work they are supposed to. *Industrial and Organisational Psychology: Perspective on Science and Practice 1*(1), 84–97. http://dx.doi.org/10.1111/j.1754-9434.2007.00028.x

Lance, C.E. (2012, April). *Discussant.* Paper presented at the 27th Annual Conference of the Society for Industrial and Organisational Psychology, San Diego, CA.

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 postexercise dimension ratings. *Journal of Applied Psychology, 89*(2), 377–385. http://dx.doi.org/10.1037/0021-9010.89.2.377, PMid:15065983

Lievens, F. (1998). Factors which improve the construct validity of Assessment Centres: A review. *International Journal of Selection and Assessment, 6*(3), 141–152. http://dx.doi.org/10.1111/1468-2389.00085

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, inter-rater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255–264. http://dx.doi.org/10.1037/0021-9010.86.2.255, PMid:11393438

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology, 87*(4), 675–686. http://dx.doi.org/10.1037/0021-9010.87.4.675, PMid:12184572

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises and dancing bears. *European Journal of Work and Organisational Psychology, 18(*1), 102–121. http://dx.doi.org/10.1080/13594320802058997

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 enter convergent and discriminant validity. *Journal of Applied Psychology, 91*(2), 267–258. http://dx.doi.org/10.1037/0021-9010.91.2.247, PMid:16551181

Lievens, F., & Christiansen, N.D. (2010). Core debates in assessment center research: Dimensions versus exercises. In D. Jackson, C. Lance, & B. Hoffman (Eds). *The psychology of assessment centers* (pp. 68–91). London, UK: Routledge.

Lievens, F., & Conway, J.M. (2001). Dimension and exercise variance in assessment centre scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*(6), 1202–1222. http://dx.doi.org/10.1037/0021-9010.86.6.1202, PMid:11768062

Lievens, F., & Schollaert, E. (2011). Adjusting exercise design in assessment centers: Theory, practice, and research. In N. Povah, & G.C. Thornton (Eds.), *Assessment centres and global talent management* (pp. 47–60). Surrey, UK: Gower Publishing Limited.

Lievens, F., Tett, R.O., & Schleicher, D.J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. *Research in Personnel and Human Resources Management, 28,* 99–152. http://dx.doi.org/10.1108/S0742-7301(2009)0000028006

Lowry, P.E. (1996). A survey of the assessment center process in the public sector. *Public Personnel Management, 25*(3), 307–321.

Maas, C.J.M., Lansvelt-Mulders, G.J.L.M., & Hox, J.J. (2009). A multilevel multitrait-multimethod analysis. *Methodology, 5*(3), 72–77.

Melchers, K.G., Wirz, A., & Kleinmann, M. (2012). *Dimensions and exercises: Theoretical background of mixed-model assessment centers*. Retrieved August 12, 2013, from http://www.psychologie.uzh.ch/fachrich

Meriac, J.P., Hoffman, B.J., Woehr, D.J., & Fleisher, M.S. (2008). Evidence of the validity of assessment center dimensions: Analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*(5), 1042–1052. http://dx.doi.org/10.1037/0021-9010.93.5.1042, PMid:18808224

Meyer, R.D., Dalal, R.S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *46*, 121–140. http://dx.doi.org/10.1177/0149206309349309

Mischel, W. (1968). *Personality and assessment.* Hoboken, NJ: John Wiley & Sons Inc.

Moyo, S. (2009). *A preliminary factor analytic investigation into the first-order factor structure of the fifteen factor questionnaire plus on a sample of black South African managers*. Unpublished master's dissertation, University of Stellenbosch, Stellenbosch, South Africa.

Murphy, K. (2010, March). *Psychometrics and Assessment Centers*. Paper presented at the 30th annual ACSG Conference, Stellenbosch, South Africa.

Neidig, R.D., & Neidig, P.J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69*(1), 182–186. http://dx.doi.org/10.1037/0021-9010.69.1.182

Oehley, A.M. (2007). *The development and evaluation of a partial talent management competency model.* Unpublished master's thesis, University of Stellenbosch, Stellenbosch, South Africa.

Oehley, A.M., & Theron, C.C. (2010). The development and evaluation of a partial talent management structural model. *Management Dynamics, 19*(3), 2–28.

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*(3), 262–274. http://dx.doi.org/10.1111/j.2044-8325.1987.tb00252.x

Rupp, D.E., Thornton, C.G., & Gibbons, A.M. (2008). The construct validity of the assessment centre method and usefulness of dimensions as focal constructs. *Industrial and Organisational Psychology: Perspectives on Science and Practice, 1*(1), 116–120. http://dx.doi.org/10.1111/j.1754-9434.2007.00021.x

Sackett, P.R., & Dreher, G.F. (1982). Constructs and assessment center dimensions: Some troubling findings. *Journal of Applied Psychology, 67*(4), 401–410. http://dx.doi.org/10.1037/0021-9010.67.4.401

Schmitt, N., Gooding, R.Z., Noe, R.A., & Kirsch, M. (1984). Meta analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 27*(3), 407–422. http://dx.doi.org/10.1111/j.1744-6570.1984.tb00519.x

Schneider, J.R., & Schmitt, N. (1992). An exercise design approach to understanding assessment centre dimension and exercise constructs. *Journal of Applied Psychology, 77*(1), 32–41. http://dx.doi.org/10.1037/0021-9010.77.1.32

Schollaert, E., & Lievens, F. (in press). The use of role-player prompts in assessment center exercises. *International Journal of Selection and Assessment.*

Spector, P.E., Schneider, J.R., Vance, C.A., & Hezlett, S. A. (2000). The relation of cognitive ability and personality traits to assessment centre performance. *Journal of Social Psychology, 30*(7), 1474–1491. http://dx.doi.org/10.1111/j.1559-1816.2000.tb02531.x

Tett, R.P., & Burnett, D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*, 500–517. http://dx.doi.org/10.1037/0021-9010.88.3.500, PMid:12814298

Tett, R.P., Guterman, H.A., Bleier, A., & Murphy, P.J. (2000). Development and content validation of 'hyperdimensional' taxonomy of managerial competence. *Human Performance, 13*(3), 205–251. http://dx.doi.org/10.1207/S15327043HUP1303_1

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

Thornton, G.C., & Gibbons, A.M. (2009). Validity of assessment centres for personnel selection. *Human Resource Management Review, 19*(3), 169–187. http://dx.doi.org/10.1016/j.hrmr.2009.02.002

Thornton, G.C., & Mueller-Hanson, R.A. (2004). *Developing Organisational Simulations: A Guide for Practitioners and Students*. New Jersey: Lawrence Erlbaum Associates.

Thornton, G.C. III, & Rupp, D.R. (2003). Simulations and assessment centers. In J. Thomas (Ed.), *Industrial and organisational assessment* (pp. 319–344). Hoboken, NJ: Wiley.

Thornton, G.C. III, & Rupp, D.R. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.

Van der Bank, F. (2007). *The development and validation of a partial competency model for branch managers in the clothing retail industry.* Unpublished Master's dissertation, University of Stellenbosch, Stellenbosch, South Africa.

Woehr, D.J., & Arthur, W. (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. http://dx.doi.org/10.1177/014920630302900206