SA Journal of Industrial Psychology
ISSN: (Online) 2071-0763, (Print) 0258-5200

# Investigating the construct validity of an electronic in-basket exercise using bias-corrected bootstrapping and Monte Carlo re-sampling techniques

**Authors:**
Jurgen Becker[1]
Deon Meiring[2]
Jan H. van der Westhuizen[3]

**Affiliations:**
[1]Department of Industrial Psychology, University of the Western Cape, Bellville, South Africa

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

[3]Experttech Group, Pretoria, South Africa

**Corresponding author:**
Jurgen Becker,
jbecker@uwc.ac.za

**Dates:**
Received: 23 Aug. 2018
Accepted: 09 July 2019
Published: 26 Sept. 2019

**How to cite this article:**
Becker, J., Meiring, D., & Van der Westhuizen, J.H. (2019). Investigating the construct validity of an electronic in-basket exercise using bias-corrected bootstrapping and Monte Carlo re-sampling techniques. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde 45*(0), a1582. https://doi.org/10.4102/sajip.v45i0.1582

**Orientation:** Technology-based simulation exercises are popular assessment measures for the selection and development of human resources.

**Research purpose:** The primary goal of this study was to investigate the construct validity of an electronic in-basket exercise using computer-based simulation technology. The secondary goal of the study was to investigate how re-sampling techniques can be used to recover model parameters using small samples.

**Motivation for the study:** Although computer-based simulations are becoming more popular in the applied context, relatively little is known about the construct validity of these measures.

**Research approach/design and method:** A quantitative *ex post facto* correlational design was used in the current study with a convenience sample ($N$ = 89). The internal structure of the simulation exercise was assessed using a confirmatory factor analytical approach. In addition, bias-corrected bootstrapping and Monte Carlo simulation strategies were used to assess the confidence intervals around model parameters.

**Main findings:** Support was not found for the entire model, but only for one of the dimensions, namely, the Interaction dimension. Multicollinearity was found between most of the dimensions that were problematic for factor analyses.

**Practical/managerial implications:** This study holds important implications for assessment practitioners who hope to develop unproctored simulation exercises.

**Contribution/value-add:** This study aims to contribute to the existing debate regarding the validity and utility of assessment centres (ACs), as well as to the literature concerning the use of technology-driven ACs. In addition, the study aims to make a methodological contribution by demonstrating how re-sampling techniques can be used in small AC samples.

**Keywords:** Assessment centres; electronic in-basket; Monte Carlo; bias-corrected bootstrapping; small sample analyses; computer-based simulations.

## Introduction

### Orientation

A number of selection instruments are available for the selection of personnel. They include personality questionnaires, targeted interviews, situational interviews, situational judgement tests, aptitude and ability tests, previous job roles and simulated exercises. Some of these instruments are more effective than others, while some have higher predictive validity than others (Sackett, Lievens, Van Iddekinge, & Kuncel, 2017; Schmidt & Oh, 2015). Assessment centres (ACs) have demonstrated superior criterion-related validity in comparison to other stand-alone measures (e.g. personality measures and interviews) (Arthur, Day, McNelly, & Edens, 2003; Meiring, Becker, Gericke, & Louw, 2015). In addition, there is also evidence that AC ratings can improve the prediction of job or training performance beyond other common predictors like cognitive ability and personality (Sackett, Shewach, & Keiser, 2017).

This may be the reason why organisations still employ ACs when selecting and developing their employees. Although literature supports the strong link between standardised tests of general mental ability and job performance, behaviour-based assessment provides a richer and more nuanced view of managerial potential (Lievens & Thornton, 2005). Notwithstanding the value of ACs, the relatively high cost of the method remains a deterrent for most organisations. In recent years, the development of workforce management software has enabled the large-scale automation of human resource functions such as benefit allocation, performance management, and skills development. Similarly, AC design and the delivery of exercises have been shaped by advances in information technology. Electronic in-baskets have probably been the most popular AC exercise to place within a technology-enabled platform or application because of the high degree of fidelity between traditional in-basket exercises and electronic in-basket applications (Meiring & Van der Westhuizen, 2011).

Despite the large-scale application of electronic in-baskets, applied research has not kept up with the prolific changes in industry. Relatively little is still known regarding the internal structure of electronic in-baskets compared to traditional in-baskets. More specifically, are the construct-related problems associated with traditional ACs still problematic within the technology-enabled simulations? The current study aims to answer these research questions through the investigation of a large-scale electronic in-basket used for development purposes. Finally, the study demonstrates how re-sampling techniques can be used to augment small samples that typically plague AC research. Bias-corrected bootstrapped confidence intervals and Monte Carlo re-sampling strategies were used to produce parameter estimates from empirically derived bootstrapped confidence intervals.

## Research purpose and objectives

The popularity of ACs in practice stems largely from the large body of literature that supports the link between dimension ratings and on-the-job performance (Arthur et al., 2003; Hermelin, Lievens, & Robertson, 2007). That is, AC ratings accurately predict which candidates have the potential to succeed in meeting the job objectives for an intended role. Research on the construct-related validity of ACs has, however, been less successful. Building on the seminal work of Sackett and Dreher (1982), research has consistently found strong correlations between dimensions measured by the same exercises rather than between the same dimensions measured across different exercises. Historically, dimensions have often been the main currency of ACs (Thornton & Rupp, 2012). In organisations, dimension ratings play an essential role in informing human resource practices such as selection, placement and development, thereby serving as a conventional and appropriate way to report on AC performance (Thornton & Gibbons, 2009).

The consistent finding that exercise effects dominate AC ratings has prompted numerous researchers to put forth plausible explanations for the findings. Lievens (2009) argues that the aberrant results may be because of poor AC designs, such as too many dimensions included in ACs, the absence of behavioural checklists to classify behaviour, the use of psychologists as raters and frame-of-reference training. Despite these design changes, ACs still predominantly display exercise effects (Brits, Meiring, & Becker, 2013; Lance, 2008).

In light of these persistent mixed findings, three notable large-scale reviews dealing with construct-related validity have aimed to find solutions for the statistical challenges facing AC research using Confirmatory Factor Analysis techniques (Bowler & Woehr, 2006; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lievens & Conway, 2001). Bowler and Woehr (2006) conducted meta-analytical analyses on the exercise-by-dimension matrices of various CFA model configurations, and found that dimensions accounted for 22% of the variance while exercises accounted for 34% of the variance of post-exercise dimension ratings (PEDRs). Next, Lievens and Conway (2001) conducted separate CFAs for published exercise × dimension correlation matrices, and found that dimensions and exercises accounted for more or less equal proportions of variance. Finally, Lance et al. (2004) analysed the same exercise × dimension correlation matrices used in the meta-analyses by Leivens and Conway (2001), however, avoided specifying correlated uniquenesses, which was a problematic statistical assumption in the Lievens and Conway (2001) study. They found stronger support for a general performance factor across exercises. Recent research suggests that increasing the number of behavioural indicators rated per dimension in an exercise leads to better model fit and stronger support for dimensions in AC data (Monahan, Hoffman, Lance, Jackson, & Foster, 2013).

Given the mixed findings it is difficult to anticipate which source of variance will be dominant in the current investigation. This may also not be the most important question. Rather, judgements regarding the validity of ACs are largely dependent on the design intention of the simulation. If the AC was developed to tap into a distinct aspect of that dimension's construct space across exercises, finding strong support for dimension effects will probably underscore the construct validity of the assessment. However, if the simulation was developed to tap into different elements of the dimensions across exercises, finding strong dimension factors at the expense of systematic dimension–exercise interaction effects would not be evidence of construct validity. Based on the fact that the current AC was designed to tap dimension-specific behavioural consistency across exercises, the construct validity of the assessment should rightly be investigated from the perspective of correlated dimensions.

Although the debate has recently moved beyond the exercise-versus-dimension debate (Lievens & Christiansen, 2010), we consider it important to investigate the internal structure of technology-delivered ACs because the medium of delivery can moderate the relationship between the stimuli and the behavioural response. Furthermore, in our experience very few practitioners structure interventions and provide individual feedback on exercise and dimensions. It probably remains true that most practitioners design their ACs based on a number of key competencies that are needed to be successful in a given position (Meiring & Buckett, 2016). If this is indeed the case, then it remains important to investigate the internal structure of ACs that are mainly designed to reflect dimensions.

An additional problem with AC research is that samples are typically small because of the cost of administration. Normally, ACs are administered to a small number of participants at the end of a multiple-hurdle assessment approach. The lack of construct validity, when defined as cross-exercise dimension-based behaviour congruence, may be explained in part by the lack of statistical power associated with the small sample sizes. Modern multivariate statistical techniques, especially confirmatory factor analytical approaches, require large and normally distributed data (Byrne, 2016). Thus, the secondary goal of the study was to present two re-sampling techniques that can be used by practitioners to improve the confidence in AC parameters.

The main research objective of this study was to design and implement computer-based simulation technology (CBST) as an electronic in-basket exercise (depicting the day-to-day activities of a supervisor) in an assessment development centre (ADC) for a major manufacturing enterprise in the United States. The goal of the ADC was to identify leadership potential – those individuals who may be ready to be promoted to higher levels in the organisation. This process incorporated group and individual online simulation exercises, which were used to measure behavioural and organisational competencies and performance areas on strategic and tactical levels. A secondary goal of the study was to investigate if re-sampling techniques can be used to assess model fit and to estimate confidence intervals around model parameters. Thus, the overarching research question can be described as follows: *Can a CBST as an in-basket exercise in ADCs be used to accurately measure behavioural dimensions?*

Based on the foregoing research question, the primary objectives of this study are to:

- examine the construct validity of the CBST in-basket exercise;
- demonstrate the use of re-sampling techniques in evaluating the quality of model parameters.

# Review of the literature

## The use of computer-based assessment centres in personnel development and selection

Recent research in employee selection has shifted the focus from traditional selection paradigms to more dynamic and flexible delivery methods. This is mainly driven by the higher fidelity of technology-enabled platforms and their associated cost savings. There is an increased interest in different selection methods such as situational judgement tests and the role of technology and the Internet in recruitment and selection. Social networking websites and video résumés have become part of selection procedures (Nikolaou, Anderson, & Salgado, 2012). A recent development in this regard relates to the use of technology in selection and assessment. The use of technology often takes the form of online simulations and web-based assessments. Simulations in selection and assessment are intended to closely replicate certain tasks, skills and abilities required for performance on the job (Schmitt, 2012).

This study made use of a computer-delivered in-basket exercise. By its very nature we can think of this assessment as a situational judgement test (SJT) rather than an AC because the assessment was made up of only a single exercise. Based on best practice guidelines for the use of the AC method in South Africa (Meiring & Buckett, 2016), ACs must consist of at least two simulated exercises. For this reason, we regard the electronic in-basket as a stand-alone competency-based simulation exercise rather than a full-fledged AC.

The electronic in-basket contained a computer-based in-basket exercise with multiple case studies, some of which had open-ended response formats while others made use of machine-driven scoring options. The scoring key was developed by an independent team of behavioural experts in collaboration with line managers and human resources in the given organisation. The response options reflect the desired behaviours of the supervisor in degrees of appropriateness (1 – least appropriate to 5 – most appropriate). Most of the responses were machine scored. There were some open-ended sections in the in-basket exercise that required direct input from the respondents. These responses were scored by a team of trained behavioural experts. The final overall assessment rating was a weighted combination of the scores achieved in the two sections of the same in-basket exercise. Thus, the in-basket exercise complies with the criteria for a traditional AC, at least as far as data integration is concerned, although only one simulation exercise was used.

## Validity issues of assessment centres

As the acceptance and widespread use of competency-based assessments have increased in the last two decades, various interest groups have published practice and research guidelines (International Task Force on Assessment Centre Guidelines, 2015). According to Arthur, Day and Woehr (2008), validity must be established at the start of test construction and prior to operational use. Lievens, Dilchert and Ones (2009) defined validity as the process of collecting evidence (AC results) to determine the meaning of such assessment ratings and the inferences based on these ratings. Validity is defined by the use and purpose of the AC and is crucial to the permissibility of inferences derived from AC measures.

The popularity of ACs is because of their many strengths, including that they have little adverse impact and predict a variety of performance criteria (Thornton & Rupp, 2006) with predictive validity correlations ranging from 0.28 to 0.52 (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hermelin et al., 2007). In addition, the method has been shown to have high criterion-related validity, as well as content validity (Gaugler et al., 1987; Lievens et al., 2009).

However, research evidence concerning the internal structure of ACs shows much less support for the construct validity of AC dimensions (Kleinmann et al., 2011; Tett & Burnett, 2003). This leads to the conclusion that ACs are perhaps not very successful at measuring the constructs (dimensions) that they aim to measure (Haaland & Christiansen, 2002). Specifically, evidence in support of the discriminant validity of dimensions in ACs is relatively sparse (Thornton, Mueller-Hanson, & Rupp, 2017).

In the context of ACs, the International Task Force (2015) describes validity as the extent to which an AC yields valuable, useful results. Validity relates to whether the method actually measures what it is designed to measure (Arthur et al., 2008). While ACs have demonstrated impressive evidence related to content and criterion validity, the approach has been criticised for not being able to prove that it does, in fact, measure a set of pre-determined dimensions.

Assessment centre research has largely used multitrait–multimethod (MTMM) analyses as a framework for the analysis of the internal structure of AC ratings (Campbell & Fiske, 1959). This framework is useful when multiple constructs are measured using multiple methods, as is the case with ACs. According to this framework, when the same construct is measured using different methods, the scores on the construct are expected to correlate strongly with one another or to converge. This is known as convergent validity. On the contrary, when different constructs are measured using the same method or when different constructs are measured using different methods, the scores should not overlap substantially. The discrimination between different constructs is known as discriminant validity.

Within the AC literature, Sackett and Dreher (1982) sent shockwaves through the AC community with their statement that differences in dimension scores could not be explained across different exercises. Sackett and Dreher (1982) studied the PEDRs in terms of MTMM, and argued that there is more constancy in the ratings of dimensions in a single exercise than in the ratings for a single dimension across multiple exercises. These findings and the conclusions they generated highlighted the shortcomings relating to the identification and use of dimensions within AC exercises (Jackson, Lance, & Hoffman, 2012).

Studies focusing on dimension-based ACs (DBACs) indicate that results relating to dimensions across exercises are most meaningful in decisions pertaining to candidates. The dimension-based focus is the most commonly used, most commonly researched and most commonly discussed AC perspective (Thornton & Rupp, 2012).

More recently, authors have argued that cross-exercise correlations of dimensions will remain elusive because behaviour is exercise-dependent (Hoffman, 2012). Although we agree with this thinking, our position remains that one would like to see some form of conceptual separation in dimension scores, especially if these dimensions are used as stand-alone criteria for decision-making. This will form the basis for any inferences regarding the discriminant validity of the AC. In contrast one would hope to find relatively high correlations between the same dimensions in different exercises. However, because competencies are measured with a single method, this assumption cannot be tested in the current study. However, one would at least expect similar competencies to be moderately correlated, which can be seen as a proxy of convergent validity.

## Statistical methods for construct validity and re-sampling techniques

Because the construct validity of ACs is largely concerned with the internal structure of the AC and is related to either a task-based orientation or a dimension-based orientation, Campbell and Fiske's (1959) MTMM framework is specifically useful when considering AC structures. This framework has been widely used to investigate the internal structure of AC ratings.

Despite the frequent use of MTMM matrices, Hoffman (2012) argues that MTMM approaches may not be well suited to analyse AC factor structures (Bowler & Woehr, 2006). These authors warn of a potential oversight of AC performance ratings, as theoretically the MTMM does not recognise variance in AC ratings stemming from the assessee, the assessor, the dimensions or exercises and the interactions between these sources of variance. Analytically, the MTMM matrices are plagued by non-admissible solutions and weak model termination (Lance, Woehr, & Meade, 2007). Lance et al. (2007) suggest using hybrid models when investigating ACs such as Correlated Dimensions Correlated Exercises (CDCE), One-Dimension Correlated Exercises (1DCE) and Unilateral Dimensions Correlated Exercises (UDCE + G), which may lead to more useful findings because these approaches overcome many of the shortcomings of MTMM.

More recently, approaches such as generalisability theory and other variance decomposition approaches have been used to investigate the relative contributions and interactions of assessor, dimension or exercise variance as sources of legitimate variance in AC ratings (Bowler & Woehr, 2008).

Nontheless, the CFA approach has dominated investigations of the internal construct validity of ACs and is ideal for large samples of data ($n > 200$). This may be explained, in part, by the fact that most dimensions included in ACs are based on the initial job analyses. Thus, CFAs are often used because the representative selection of the (task and contextual) demands, constraints and opportunities that constitute the job in question (and their potential interactions) are known prior to the development of the AC. For this reason, most validation studies are concerned with finding confirmation for the proposed structure rather than finding the appropriate structure. The foregoing line of reasoning suggests that an AC attempts to assess the latent behavioural dimensions comprising performance by eliciting observable behavioural denotations of latent performance dimensions or competencies via specific stimuli set in a variety of specific micro-contexts that differ in terms of situational characteristics. A conclusion that the instrument has construct validity (i.e. that inferences on the construct as constitutively defined may permissibly be inferred from measures of the instrument) will be strengthened if it can, in addition, be shown that a structural model reflecting the manner in which the construct or constructs of interest are embedded in a larger nomological network of constructs according to the constitutive definition fits the data.

However, ACs have generally been plagued by small samples (Lievens & Christiaansen, 2010). This is also true of the current study. Alternatives suggested by Bowler and Woehr (2008) include the use of re-sampling techniques that remedy some of the problems associated with MTMM.

This method was independently tested by Hoffman**,** Melchers, Blair, Kleinmann and Ladd (2011b), who found that the alternative re-sampling provided more insights than the traditional MTMM approach. Kuncel and Sackett (2014) used a different sampling technique based on the theory of composites to investigate the exercise–dimension variance. More recently, new approaches have been proposed to overcome the challenges related to CFA analyses using small samples.

The Monte Carlo family of re-sampling techniques may be fruitfully used to test the appropriateness of model parameters, standard errors, confidence intervals and even fit indices under various assumptions. Because of the low statistical power in small samples, standard errors may be overestimated, which may lead to significant effects being missed. In contrast, if standard errors are underestimated, significant effect may be overstated (Muthén & Muthén, 2002). The Monte Carlo Method for Assessing Mediation (MCMAM) was first described and evaluated by MacKinnon, Lockwood and Williams (2004) to assess small-sample performance. The Monte Carlo technique entails the extrapolation of data into thousands of simulated data frames, which model statistical characteristics similar to the original sample data (Muthén & Muthén, 2002). Monte Carlo in essence is an estimator or test statistic that has a true sampling distribution under a particular set of conditions; it assists in the true sampling distribution (Lance, Woehr, & Meade, 2005).

Monte Carlo features include saving parameter estimates from the analysis of real data to be used as population and/or coverage values for data generation in a Monte Carlo simulation study. Monte Carlo simulations involve identifying a mathematical model of the activity or process to be researched and defining the parameters such as mean and standard deviation for each factor in the model (Lance et al., 2005). It creates random data according to those parameters and simulates and analyses the output of the process. A typical Monte Carlo simulation involves the generation of independent datasets of interest and computing the numerical value of the data for each dataset. The larger the dataset, the closer the true sampling properties of the data (Davidian, 2005). When used in ACs, the Monte Carlo simulation addresses the extent to which the model fits the generated data (Lance et al., 2005). Lance et al. (2005) suggest that Monte Carlo simulation should be used to assess whether model fit should be used to compare competing CFA models.

An alternative re-sampling technique known as residual bootstrapping (Bollen & Stine, 1992) can also be used independently or in conjunction with Monte Carlo simulations. However, bootstrapping is aimed at investigating the standard errors of model parameters across volume bias-corrected draws. In an article by MacKinnon et al. (2004), bias-corrected bootstrap confidence intervals were found to be very accurate. Bootstrapping is also an option for smaller datasets and involves the re-sampling of an original dataset to a desired sample size (Efron, 1979).

Bootstrapping complements traditional confidence intervals by estimating standard errors of parameter estimates over a large number of hypothetical sample draws (Bollen & Stine, 1992; Hancock & Nevitt, 1999). One of the main reasons why Bollen–Stine bootstraps was included in the study is based on the fact that the technique provides a way to impose the covariance structure model on the sample data. This way the researcher is able to examine the bootstrapping performance of the fit statistics under the assumptions of the 'null hypothesis' that the model fits (Kim & Millsap, 2014). Furthermore, Monte Carlo simulated data may result in bias standard errors and parameter estimates if the data are generated under the assumption of normality when the data in the sample are actually non-normal (Muthén and Muthén, 2002). The Bollen–Stine bootstrap can therefore correct for standard error and fit statistical bias that occurs in structural equation modelling (SEM) applications because of non-normal data (Bollen & Stine, 1992). Bollen–Stine bootstraps (BSBS) are deployed whereby the original data are rotated to conform to the fitted structure. Bollen–Stine bootstraps take the empirical sample of size (*N*) and randomly draw repeated samples with replacement to the same size (*N*). The goal is to repeat the sampling size and form an integrated picture of the original sampling data (Efron, 1979). In this study, because of the relatively small sample size the BSBS standard errors and bias-corrected confidence intervals were used to get a sense of the width of confidence intervals around parameter estimates. Relatively narrow confidence intervals around parameter estimates suggest that standard errors are not abnormally high.

In the literature review, we focused on the historical debate regarding the construct validity and internal structure of ACs. Internet-delivered simulated exercises closely resemble the features of traditional sample-based assessment, yet the delivery and scoring platform differs significantly. However, relatively little is known about the internal structure of electronic simulations in general and in-baskets in particular. Thus, the overarching goal of the study is to assess the internal structure of an electronic in-basket. Furthermore, the potential benefits of re-sampling techniques were discussed in the context of ACs. The section concluded with a discussion of re-sampling techniques and the use of Monte Carlo and bootstrapping techniques.

# Research design

## Research approach

A non-experimental, quantitative research design was used in the current study to empirically test the main research objectives. More specifically, an *ex post facto* correlational design was used and implemented in a confirmatory factor analytical framework. Post-exercise dimension ratings were used as the level of measurement and served as manifest variables in the factor analytical models that were specified.

## Research strategy

Initially, the data were screened for multivariate outliers and out of range responses. Descriptive statistics were generated to investigate the distribution and central tendency of PEDR scores for each of the competency dimensions. Inferential statistics were generated by specifying a confirmatory factor analytical model. The internal structure of the electronic in-basket can be operationalised through the specification of fixed and freely estimated model parameters.

More specifically, the CBST measurement model can be defined in terms of a set of measurement equations, expressed in matrix algebra notation (see Equation 1):

$$X = \Lambda_X \xi + \delta \qquad \text{[Eqn 1]}$$

Where:

- $X$ is a 19 × 1 column vector of observable indicator variables (PEDR);
- $\Lambda_X$ is a 19 × 5 matrix of factor loadings;
- $\xi$ is a 1 × 5 column vector of latent competency dimensions;
- $\delta$ is a 19 × 1 column vector of measurement error.

In addition, all the off-diagonal elements of the phi covariance matrix, denoting the covariance between the five latent competencies, were freed up to be estimated. Model parameters of the CFA model were estimated using maximum likelihood with robust standard errors and fit indices because of the non-normality of the sample data. For identification purposes, each of the five latent competencies was standardised and all error variances were specified to be uncorrelated. Fit indices and model parameters were estimated using Mplus 7.2 (Muthén & Muthén, 2017. Multiple fit indices were used to evaluate the tenability of the CBST. These indices included the Satorra–Bentler $\chi^2$, the Comparative Fit Index (CFI; Bentler, 1990), the Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980) with accompanying confidence intervals, and the Standardised Root Mean Square Residual (SRMR; Joreskog & Sorbom, 1993). Comparative Fit Index values in excess of 0.90 (Bentler, 1990), RMSEA values lower than 0.08 (Browne & Cudeck, 1993) and SRMR values lower than 0.06 (Hu & Bentler, 1999) were regarded as satisfactory.

# Research method

## Research participants

A convenience sample of 89 supervisors were selected in a non-random fashion from a large multinational manufacturing organisation operating in the petroleum and rubber industry in North America. The sample was selected from incumbent supervisors, who were earmarked to partake in a larger leadership developmental programme in the organisation. The first step in the development programme was to complete the CBST to gain more insight into the strengths and development areas of each supervisor.

## Measuring instruments

Six broad competencies were identified by the client organisation for inclusion in the CBST based on their proposed link to job performance as identified through the job analysis process. An external consulting organisation was contracted to develop the behavioural indicators and scoring method for each competency. A summary of the six meta-competency clusters and sub-dimensions is presented in Table 1.

Because all the meta-competencies were operationalised within the online in-basket, there was only one simulation format. For this reason, the current research cannot be regarded as an AC because competencies were not measured within multiple exercises (Meiring & Buckett, 2016).

The electronic in-basket was scored using a combination of multiple-choice machine scoring and manual scoring by trained raters of the open-ended video vignettes. The scores

**TABLE 1:** Meta-competencies and sub-dimensions.

| Meta-competencies | Sub-dimensions | Abbreviation |
|---|---|---|
| Vision | Visionary thinking | VIS_VT |
| | Strategic orientation | VIS_SO |
| | Innovation | VIS_INN |
| | Leading change | VIS_LS |
| Drive | Initiative | DRI_INI |
| | Leading and steering | DRI_LS |
| | Self-determination | DRI_SD |
| | Passion and commitment | DRI_PAS |
| Execution | Problem-solving | EXE_PS |
| | Decision-making | EXE_DM |
| | Delivering results | EXE_DR |
| | Assertiveness | EXE_ASS |
| Entrepreneurship | Customer orientation | ENT_CUST |
| | Profit orientation | ENT_PROF |
| | Quality orientation | ENT_QUAL |
| | Integrity (in business) | ENT_INT |
| Learning | Building up business acumen | LRN_BA |
| | Self-reflection | LRN_SR |
| | Handling feedback | LRN_HF |
| | Coaching others | LRN_CO |
| Interaction | Clear and open communication | INT_COM |
| | Networking | INT_NET |
| | Fostering teamwork | INT_FT |
| | Motivating others | INT_MO |
| | Intercultural sensitivity | INT_IS |
| | Promoting diversity | INT_PD |

were integrated according to equal weighted averages for open-ended and multiple-choice response options. The open-ended responses were scored by a team of trained behavioural experts, who examined the responses in accordance with the conceptual definitions. The assessors attended frame-of-reference training to accurately observe, record, classify and assess the responses to the open-ended questions. During the training session, examples were provided with a range of responses, ranging from appropriate to less appropriate behavioural examples, and how to use the five point behaviourally anchored rating scale (BARS) to assess responses. As part of the training, all assessors had to complete the CBST.

In this regard, the simulated electronic in-basket complied with the criteria of traditional ACs insofar as each competency was observed and scored by multiple raters and integrated into an overall score. Because all the competencies were operationalised in a single simulation format, the in-basket cannot be regarded as a traditional AC. However, we believe that the results of the study hold important implications for sample-based assessment, and specifically for those simulations that are delivered on an electronic platform. Completion of the computer-based simulation in-basket exercise took around 40 min. All participants completed the task in the allocated time. For this reason, there were no missing values in the data.

## Research procedure and ethical considerations

Managers who participated in the AC were identified for future promotion because of strong performance and competence in their incumbent positions. Because the purpose of the AC was for development, all the managers who participated in the study consented to partake in the AC. All participants were informed that their data may be used for research purposes. The identity of all participants was kept anonymous by converting the raw data into an encrypted file that was shared with the researchers. Thus, the final dataset contained no personal information other than the race, age and gender of the participants.

## Statistical analysis

The internal structure of the CBST was assessed by specifying a confirmatory factor analytical model with Mplus 7.2 (Muthén & Muthén, 1998–2017). Because the researcher had a well-developed *a priori* conceptualisation of how the sub-dimension scores are related to the higher order latent dimensions, it was decided to conduct only CFA and not exploratory factor analysis (EFA). Thus, the goal of the CFA analysis was not to investigate the manner in which the PEDR scores are related to higher order latent dimensions, but rather to investigate the relative strength of relationships between PEDRs and latent dimensions in the conceptual model.

In addition, it was important to evaluate the overall fit of the proposed model to the observed data. If strong support was found for the model parameters and overall fit of the model to the data, it would be possible to conclude that the CBST has construct validity and may be used for diagnostic and selection purposes (Lievens & Christiansen, 2010). Because of the small sample size, the authors decided to use re-sampling techniques to assess the confidence intervals and bootstrapped standard errors of the factor loadings of parameter estimates. In small sample sizes, the statistical power available to reject the null hypothesis is limited when in reality (i.e. in the population) the linkages are statistically different from zero (Lievens & Christiansen, 2010). To overcome this fundamental methodological problem, the researchers employed Monte Carlo and bootstrapping techniques to assess the stability of standard errors and to construct confidence intervals around point estimates.

Monte Carlo simulations were used to extract 1000 simulated datasets with model statistical characteristics similar to the sample data. This approach used in the current study can be regarded as an external Monte Carlo study, insofar as parameters saved from the real data analyses are used for population values for the simulated data. Thus, a two-step approach is used to calculate the model parameters and then to use these values as input to generate data in step 2. The fact that the simulated data use the model parameters estimated from the real data may not be sufficient to capture the non-normality in the simulated data. However, when working with skew data, the robust maximum likelihood estimation (MLE) can be used.

This may be particularly important when examining the critical value chi-square fit statistic in the parent and simulated samples (Curran, West, & Finch, 1996; Wang, Fan, & Wilson, 1996). Kim and Millsap (2014) advocate the use of robust MLE for both the real and generated samples. Analyses by the Kim and Millsap (2014) indicated that the original assumptions about non-normality of data in simulated studies by Millsap (2012) may have been overly rigorous. Kim and Millsap (2014), however, advocate that small samples may lead to higher levels of discrepancy between fit indices in the simulated and original data. Given that the sample in the current investigation is very small, it is important to investigate the differences in model fit indices and parameter estimates reported in the parent and simulated data sets.

In addition, the Bollen–Stine bootstrap (residuals bootstrap) produces a correct bootstrapped sampling distribution for chi-square, and thus a correct bootstrapped *p*-value, without presuming a specific distribution of the data (Muthén & Muthén, 2002). Because the Bollen–Stine technique preserves the characteristics of the original data, the technique may be particularly useful when the source data are non-normal.

The Bollen–Stine bootstrap can be used to correct for standard error and fit statistical bias that occur in SEM applications because of non-normal data. Bollen and Stine (1992) perform bootstraps whereby the original data are rotated to conform to the fitted structure. By default, the Bollen–Stine technique re-estimates the model with rotated data and uses the estimates as starting values for each bootstrap iteration. It also rejects samples where convergence was not achieved (implemented through reject [e-converged = 0] option supplied to bootstrap) (Millsap, 2012)

It is also possible to use confidence intervals and bootstrapping to gain greater confidence in findings. This involves investigating 'real' sampling variability without assuming specific distribution for the data. Bollen–Stine bootstrap (residuals bootstrap) produces correct bootstrapped sampling distribution for chi-square, and thus correct bootstrapped *p*-values. Re-sampling variability used to get bootstrapped *p*-values are confidence intervals that can be used to aid in the interpretation of model parameters (Bollen & Stine, 1992; Millsap, 2012).

In each case, the results obtained from the original data were compared to the results generated with the Monte Carlo simulations and Bollen–Stine bias-corrected bootstrapping. The comparative results may contribute to the AC literature by demonstrating the utility of re-sampling techniques when working with relatively small sample sizes. It is important to emphasise that the re-sampling methods are not a `silver bullet' for small sample sizes, as any sampling error contained in the sample from which re-sampling is drawn will be included in the bootstrapped sample (Enders, 2005). Moreover, any errors are likely to be duplicated in re-sampled data sets when missing data analysis techniques are applied. However, we believe that the benefits of re-sampling techniques (improved statistical power, bias-corrected standard errors and confidence intervals around parameter estimates) outweigh the alternative, which is to do nothing.

### Ethical consideration

This article followed all ethical standards for a research without direct contact with human or animal subjects.

## Results

The primary objective of this study was to examine the validation of a CBST in-basket exercise within an ADC. This objective involves proving the behavioural validity of the workplace simulation. It further implies that if construct validity is intact, then the exercises comply with the principles of the ADC and may lead to valid development and selection decisions.

The results of the study are discussed according to the following structure:

- frequencies and descriptive statistics;
- screening the data;
- examining the appropriateness of the data for multivariate CFA;
- specification and estimation of CFA model;
- evaluating the model according to goodness-of-fit indices;
- evaluating the model according to model parameters;
- using Monte Carlo estimates;
- using bootstrap (BS) bias-corrected bootstraps.

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 assessed. 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 (RML) was specified as the estimation technique. Basic descriptive statistics were generated to assess the variability and central tendency of PEDRs. The means and standard deviations of PEDRs are presented in Table 2.

The results in Table 2 suggested that the range of scores was restricted as one would expect to find when assessing job incumbents. Next, we assessed the bivariate correlations between PEDRs prior to specifying the CFA model. In total there were 26 indicators, and there were thus 325 inter-correlations in the correlation matrix, of which 22 reported bivariate correlations greater than 0.90. This was problematic because it suggested that many of the PEDR scores lacked discriminant validity. Furthermore, variables that are perfect linear combinations of each other or are extremely highly

**TABLE 2:** Descriptive statistics.

| Meta-competencies | Abbreviation | *N* | Mean | Standard deviation |
|---|---|---|---|---|
| Interaction | INT_PD | 89 | 2.280 | 0.3792 |
| | INT_IS | 89 | 2.285 | 0.3859 |
| | INT_MO | 89 | 2.774 | 0.4575 |
| | INT_FT | 89 | 2.774 | 0.4575 |
| | INT_NET | 89 | 2.720 | 0.3720 |
| | INT_COM | 89 | 3.016 | 0.6144 |
| Learning | LRN_CO | 89 | 2.774 | 0.4575 |
| | LRN_HF | 89 | 2.774 | 0.4574 |
| | LRN _SR | 89 | 2.613 | 0.4179 |
| | LRN_BA | 89 | 3.484 | 0.6144 |
| Entrepreneurship | ENT_INT | 89 | 2.887 | 0.6085 |
| | ENT_QUAL | 89 | 2.608 | 0.4160 |
| | ENT_PROF | 89 | 2.382 | 0.5129 |
| | ENT_CUST | 89 | 2.608 | 0.4160 |
| Execution | EXE_ASS | 89 | 3.446 | 0.6188 |
| | EXE_DR | 89 | 3.484 | 0.6144 |
| | EXE_DM | 89 | 3.484 | 0.6144 |
| | EXE_PS | 89 | 2.849 | 0.4650 |
| Drive | DRI_PAS | 89 | 3.059 | 0.6463 |
| | DRI_SD | 89 | 3.059 | 0.6463 |
| | DRI_LS | 89 | 2.780 | 0.4512 |
| | DRI_INI | 89 | 2.565 | 0.4676 |
| Vision | VIS_LC | 89 | 2.882 | 0.6097 |
| | VIS_INN | 89 | 2.575 | 0.4420 |
| | VIS_SO | 89 | 2.790 | 0.6269 |
| | VIS_VT | 89 | 2.785 | 0.6273 |

INT_PD, Promoting Diversity; INT_IS, Intercultural Sensitivity; INT_MO, Motivating Others; INT_FT, Fostering Teamwork; INT_NET, Networking; INT_COM, Clear and Open Communication; LRN_CO, Coaching Others; LRN_HF, Handling Feedback; LRN_SR, Self-Reflection; LRN_BA, Building up Business Acumen; ENT_INT, Integrity (In Business); ENT_QUAL, Quality Orientation; ENT_PROF, Profit Orientation; ENT_CUST, Customer Orientation; EXE_ASS, Assertiveness; EXE_DR, Delivering Results; EXE_DM, Decision-Making; EXE_PS, Problem-Solving; DRI_PAS, Passion and Commitment; DRI_SD, Self-Direction; DRI_LS, Leading and Steering; DRI_INI, Initiative; VIS_LS, Leading Change; VIS_INN, Innovation; VIS_SO, Strategic Orientation; VIS_VT, Visionary Thinking.

correlated prevent the covariance matrices from being inverted (Tabachnick & Fidell, 2007). The 22 high correlations are presented in Appendix 1.

When the total CBST model was specified as CFA model, MPLUS issued a warning that the sample covariance matrix may be singular and that the model could not converge. Based on the singular covariance matrix, it was impossible to specify and assess the total CBST. One possible remedy would be to collapse highly correlated sub-dimensions into broader competencies. In previous studies, Hoffman et al. (2011a) found support for broad competencies in PEDR scores rather than smaller idiosyncratic competencies.

Collapsing dimensions into broader competencies may make sense from a theoretical and methodological perspective. From a methodological perspective, treating dimension scores (PEDRs) as indicators of broader dimensions will increase the indicator to dimension ratio. Monahan et al. (2013) found that greater indicator to dimension ratios leads to improved termination and admissibility in CFA models. Secondly, Hoffman et al. (2011b) argue that the dimensions-as-items approach has been used extensively to validate taxonomies of managerial performance, models of organisational citizenship behaviour, multisource performance ratings and measures of managerial skills. Howard (2008) advocates the use of broad facets where indicators form the basic input to factor analytic models. For this reason, it seems to make methodological and theoretical sense to group dimensions that are strongly correlated together in broader dimensions as long as they clearly share conceptual overlap.

In Appendix 1, five of the correlations with a correlation of 1.0 belonged to the Learning meta-competency. The same was true for the Executing meta-competencies (six correlations $r > 0.955$) and Directing (six correlations $r > 0.994$). That basically left the authors with the Entrepreneurship (four-dimension ratings) and Interaction (six-dimension ratings) meta-competencies to assess. The Vision meta-competency was not considered because only three sub-dimensions can be used as indicators in the CFA model and two of them were highly correlated. Against this background it was decided to focus on the Interaction meta-competency because this dimension had the most dimension ratings that could be used as indicators and seemed to report moderate positive correlations between the sub-dimensions. However, even within the Interaction meta-competency, multicollinearity seemed to be evident, although it was significantly lower than in the other meta-competencies. The bivariate correlations of the Interaction meta-competency are presented in Table 3.

Table 3 includes some problematic correlations. In particular, within the Interaction cluster, Promoting Diversity correlated highly with Intercultural Sensitivity and Motivating Others correlated highly with Fostering Teamwork. From a measurement theory perspective, these two pairs of items should be combined into two single factors because there seems to be very little distinction between them. Based on the content, it was deemed theoretically permissible to combine these item pairs. Thus, Motivating Others was combined with Fostering Teamwork to become INT_MOFT and Promoting Diversity was combined with Intercultural Sensitivity to become INT_PDIS. After combining these constructs, the CFA solution converged to an admissible solution with satisfactory model fit. In addition, the multicollinearity problem had at least been addressed. None of the remaining bivariate correlations within the Interaction cluster was greater than 0.90. The correlation matrix for the revised Interaction meta-competency is presented in Table 4.

Because non-normal data can lead to bias fit indices and standard errors in the simulated data when using Monte Carlo, the normality of the observed variables was assessed with SPSS (Version 25, IBM, 2017). Although visual inspection indicated that most of the observed variables were non-normal, the simple test of dividing the skewness score by its associated standard error indicated that most variables followed a normal distribution. If the results are greater than ±1.96, it suggests that the data are not normal with respect to this specific statistic (Rose, Spinks, & Canhoto, 2015). The results of the analyses are summarised in Table 5.

Results from Table 5 suggest that most of the variables are normally distributed with the exception of Interaction Motivating Others and Fostering Teamwork. For this reason, we decided to specify the RML estimator as suggested by

**TABLE 3:** Bivariate correlations of the Interaction meta-competency.

| Variable | INT_PD | INT_IS | INT_MOT | INT_FT | INT_NET | INT_COM |
|---|---|---|---|---|---|---|
| INT_PD | 1 | 0.954** | 0.603** | 0.603** | 0.502** | 0.715** |
| INT_IS | 0.954** | 1 | 0.584** | 0.584** | 0.504** | 0.634** |
| INT_MO | 0.603** | 0.584** | 1 | 1.000** | 0.663** | 0.632** |
| INT_FT | 0.603** | 0.584** | 1.000** | 1 | 0.663** | 0.632** |
| INT_NET | 0.502** | 0.504** | 0.663** | 0.663** | 1 | 0.626** |
| INT_COM | 0.715** | 0.634** | 0.632** | 0.632** | 0.626** | 1 |

INT_ PD, Interaction Promoting Diversity; INT_ COM, Interaction Clear and Open Communication; INT_ NET, Interaction Networking; INT_ FT, Interaction Fostering Teamwork; INT_ MOT, Interaction Motivating Others; INT_ IS, Intercultural Sensitivity.

**, Correlation is significant at the 0.01 level (2-tailed).

**TABLE 4:** Revised Interaction meta-competency bivariate correlations.

| Variable | INT NET | INT PDIS | INT MOFT | INT COM |
|---|---|---|---|---|
| INT_NET | 1 | 0.708** | 0.671** | 0.714** |
| INT_PDIS | 0.708** | 1 | 0.613** | 0.651** |
| INT_MOFT | 0.671** | 0.613** | 1 | 0.690** |
| INT_COM | 0.714** | 0.651** | 0.690** | 1 |

INT_NET, Interaction Networking; INT_PDIS, Interaction Promoting Diversity and Intercultural Sensitivity; INT_MOFT, Interaction Motivating Others and Fostering Teamwork; INT_COM, Interaction Communication.

**, Correlation is significant at the 0.01 level (2-tailed).

**TABLE 5:** Skewness of observed variables for the revised Interaction dimension.

| Variable | Skewness | Standard error | Statistic |
|---|---|---|---|
| INT_NET | 0.366 | 0.255 | 1.43 |
| INT_PDIS | -0.077 | 0.255 | 0.30 |
| INT_MOFT | 0.640 | 0.255 | 2.50 |
| INT_COM | -0.047 | 0.255 | 0.18 |

INT_NET, Interaction Networking; INT_PDIS, Interaction Promoting Diversity and Intercultural Sensitivity; INT_MOFT, Interaction Motivating Others and Fostering Teamwork; INT_COM, Interaction Communication.

Muthén and Muthén (1998–2017). According to the authors, the parameter estimates should be the same irrespective of whether RML or maximum likelihood is used. It is only the standard errors that will be adjusted when using RML. For this reason, Muthén and Muthén (1998–2017) recommend using RML as the default estimated in CFA analyses irrespective of the data distribution.

The correlations in Table 4 suggest that none of the remaining correlations was problematic. After two rounds of revision, we specified a CFA model with four indicators. The CFA solution converged to an admissible solution. Goodness-of-fit indices are reported in Table 6.

The overall model fit can be regarded as satisfactory based on the criteria and cut-off rules reported in the methodology section. The CFI and Tucker-Lewis Index (TLI) were in excess of 0.95, and the RMSEA and SRMR are close to the normative cut-off value of 0.05.

The unstandardised and standardised results demonstrated that most of the model parameters were indicative of good model fit. This provides further support for the revised Interaction measurement model. A summary of the model parameters is presented in Table 7.

The results in Table 7 suggest that the four broad dimension ratings are good indicators of the latent competency of Interaction. Monte Carlo simulations were once again conducted using the original input model configuration as a basis for the estimation. Tables 8 and 9 contain the mean and standard deviation of the chi-square and RMSEA fit indices over the 1000 replications of the Monte Carlo analyses.

Table 8 indicates the critical chi-square value given two degrees of freedom. Thus, the value in column 1 in Table 8 provides the probability that the chi-square value exceeds the critical percentile value of 5.991. Column 2 provides the proportion of replications for which the critical value is exceeded, 0.051, which is close to the expected value of 0.050. This suggests that the chi-square distribution is well approximated in the current investigation because the expected and observed percentiles and distributions are close to one another in absolute value.

**TABLE 6:** Goodness-of-fit indices for the revised interaction dimension.

| Variable | Category | Value |
|---|---|---|
| Chi-square test of model fit | Value | 2.553 |
| | Degrees of freedom | 2.000 |
| | *P* value | 0.279 |
| | Scaling correction factor for MLR | 0.982 |
| RMSEA | Estimate | 0.056 |
| | 90% CI | 0.000–0.226 |
| | Probability RMSEA £0.05 | 0.031 |
| | CFI | 0.997 |
| | TLI | 0.991 |
| SRMR | Value | 0.011 |

RMSEA, root mean square error of approximation; SRMR, standardised root mean square residual; CI, confidence interval; CFI, comparative fit index; TLI, Tucker-Lewis Index.

Similar results are displayed in Table 9, which indicates the probability that the RMSEA value exceeds the critical value.

The critical RMSEA value of 0.052 is exceeded in approximately 9.5% of the 1000 replications. Although the mean RMSEA value in the simulated data is indicative of good fit (0.014), the relatively large deviation between the expected and observed proportions containing the critical value raises concern regarding the approximate fit of the original CFA model, given the results of the Monte Carlo simulation.

The standard error, Monte Carlo-derived standard error, average standard deviation and average coverage values are presented in Table 10. The column labelled Population provides the parameter values in the sample data. The column Average provides the average model parameters across the 1000 replications. The difference between these two values can be regarded as the proportion bias in parameter estimates presented in the last column of Table 10. The column labelled Standard deviation provides the standard deviation of parameter estimates across the Monte Carlo replications.

**TABLE 7:** Unstandardised and standardised parameter estimates of the revised Interaction dimension.

| Variable | Model results | Estimate | SE | Est/SE | Two-tailed *p* |
|---|---|---|---|---|---|
| Two-tailed | **INT BY** | | | | |
| | INT_MOFT | 0.947 | 0.081 | 11.684 | 0.000 |
| | INT_NET | 0.467 | 0.043 | 10.853 | 0.000 |
| | INT_COM | 0.375 | 0.054 | 6.940 | 0.000 |
| | INT_PDIS | 0.913 | 0.098 | 9.294 | 0.000 |
| | **Intercepts** | | | | |
| | INT_NET | 2.742 | 0.051 | 53.817 | 0.000 |
| | INT_COM | 2.798 | 0.051 | 54.519 | 0.000 |
| | INT_PDIS | 5.461 | 0.111 | 49.065 | 0.000 |
| | INT_MOFT | 5.584 | 0.103 | 54.351 | 0.000 |
| | **Variances** | | | | |
| | INT | 1.000 | 0.000 | 999.000 | 999.000 |
| | **Residual variances** | | | | |
| | INT_NET | 0.013 | 0.004 | 3.256 | 0.001 |
| | INT_COM | 0.094 | 0.025 | 3.707 | 0.000 |
| | INT_PDIS | 0.268 | 0.090 | 2.963 | 0.003 |
| | INT_MOFT | 0.042 | 0.020 | 2.117 | 0.034 |
| Standardised model results: Standardisation | **INT BY** | | | | |
| | INT_MOFT | 0.977 | 0.011 | 88.631 | 0.000 |
| | INT_NET | 0.972 | 0.010 | 96.948 | 0.000 |
| | INT_COM | 0.775 | 0.068 | 11.356 | 0.000 |
| | INT_PDIS | 0.870 | 0.040 | 21.866 | 0.000 |
| | **Intercepts** | | | | |
| | INT_NET | 5.705 | 0.476 | 11.984 | 0.000 |
| | INT_COM | 5.779 | 0.489 | 11.823 | 0.000 |
| | INT_PDIS | 5.201 | 0.500 | 10.405 | 0.000 |
| | INT_MOFT | 5.761 | 0.455 | 12.665 | 0.000 |
| | **Variance** | | | | |
| | INT | 1.000 | 0.000 | 999.000 | 999.000 |
| | **Residual variances** | | | | |
| | INT_NET | 0.054 | 0.020 | 2.787 | 0.005 |
| | INT_COM | 0.400 | 0.106 | 3.783 | 0.000 |
| | INT_PDIS | 0.243 | 0.069 | 3.511 | 0.000 |
| | INT_MOFT | 0.045 | 0.022 | 2.064 | 0.039 |

SE, standard error; Est/SE, Estimate divided by standard error; INT, Interaction; COM, Clear and Open Communication; NET, Networking; FT, Fostering Teamwork; MOT, Motivating others; IS, Intercultural Sensitivity; PD, Promoting Diversity.

**, Correlation is significant at the 0.01 level (2-tailed).

**TABLE 8:** Mean, standard deviation, critical value of chi-square fit index across 1000 draws.

| Variable | Expected | Observed | Value |
|---|---|---|---|
| Proportions | 0.990 | 0.992 | - |
| | 0.980 | 0.984 | - |
| | 0.950 | 0.961 | - |
| | 0.900 | 0.913 | - |
| | 0.800 | 0.803 | - |
| | 0.700 | 0.711 | - |
| | 0.500 | 0.503 | - |
| | 0.300 | 0.300 | - |
| | 0.200 | 0.201 | - |
| | 0.100 | 0.099 | - |
| | 0.050† | 0.051† | - |
| | 0.020 | 0.024 | - |
| | 0.010 | 0.008 | - |
| Percentiles | 0.020 | 0.030 | - |
| | 0.040 | 0.054 | - |
| | 0.103 | 0.128 | - |
| | 0.211 | 0.242 | - |
| | 0.446 | 0.448 | - |
| | 0.713 | 0.765 | - |
| | 1.386 | 1.395 | - |
| | 2.408 | 2.391 | - |
| | 3.219 | 3.222 | - |
| | 4.605 | 4.565 | - |
| | 5.991† | 6.065† | - |
| | 7.824 | 8.273 | - |
| | 9.210 | 9.082 | - |
| Degrees of freedom | - | - | 2.000 |
| Mean | - | - | 2.008 |
| Standard deviation | - | - | 1.957 |
| Number of successful computations | - | - | 1000 |

†, The probability that the chi-square value exceeds the critical percentile value.

**TABLE 9:** Mean, standard deviation, critical value of Root Mean Square Error of Approximation fit index across 1000 draws.

| Variable | Expected | Observed | Value |
|---|---|---|---|
| Proportions | 0.990 | 1.000 | - |
| | 0.980 | 1.000 | - |
| | 0.950 | 1.000 | - |
| | 0.900 | 1.000 | - |
| | 0.800 | 1.000 | - |
| | 0.700 | 0.359 | - |
| | 0.500 | 0.328 | - |
| | 0.300 | 0.258 | - |
| | 0.200 | 0.208 | - |
| | 0.100 | 0.139 | - |
| | 0.50† | 0.095† | - |
| | 0.020 | 0.057 | - |
| | 0.010 | 0.040 | - |
| Percentiles | -0.039 | 0.000 | - |
| | -0.032 | 0.000 | - |
| | -0.023 | 0.000 | - |
| | -0.015 | 0.000 | - |
| | -0.002 | 0.000 | - |
| | -0.005 | 0.000 | - |
| | -0.015 | 0.000 | - |
| | -0.026 | 0.020 | - |
| | -0.034 | 0.035 | - |
| | -0.044 | 0.051 | - |
| | -0.052† | 0.064† | - |
| | -0.061 | 0.079 | - |
| | -0.068 | 0.084 | - |
| Mean | - | - | 0.014 |
| Standard deviation | - | - | 0.023 |
| Number of successful computations | - | - | 1000 |

†, The probability that the RMSEA value exceeds the critical value.

This value can also be regarded as the population standard error (Muthén & Muthén, 2002). The column labelled MSE gives the mean square error of each parameter, while the column labelled 95% Cover gives the proportion of replications for which the 95% confidence interval contains the population parameter value. In contrast to the original data, the 95% coverage was respectable, with most values exceeding 0.80. The coverage values reflect the proportion of replications for which the 95% confidence interval contains the true parameter value. A value of 0.80 would imply that in 80% of the replications the true parameter point estimate was present. Muthén and Muthén (2002) suggest that the coverage should remain between 0.91 and 0.98. Finally, the column labelled percentage significant coefficients provides an estimate of the proportion of replications for which the null hypothesis that a parameter is equal to zero is rejected at the 0.05 level. For parameters with population values equal to zero, this value is an estimate of power with respect to a single parameter, that is, the probability of rejecting the null hypothesis when it is false (type I error) (Muthén & Muthén, 1998–2017).

Against this background, the information in Table 10 provides a mixed picture of the Monte Carlo results with regard to the Interaction dimension. Although the 95% coverage indicates that a relatively large proportion of the 95% confidence interval contains the population parameter value across the 1000 replications, the power to reject the null hypothesis stating that the parameter is equal to zero in the sample is relatively large. However, the deviation between the average population parameter estimates and the sample estimates indicates that the parameters may be biased. This would indicate that the generated data did not estimate model parameters with a high degree of accuracy.

Next, we discuss the results from the Bollen–Stine residual bootstrapped standard errors and bias-corrected confidence intervals generated with regard to the Interaction sub-scale with 1000 bootstrap draws. The intention of this analysis is to provide valid inferences from the sample data to some large universe of potential data; in other words, to provide information about the population from statistics generated with random smaller samples. Because it would be virtually impossible to obtain access to random samples from populations that have the same characteristics as the larger population, statistical methods have been developed to determine the confidence with which such inferences can be drawn, given the characteristics of the available sample (Cohen, Cohen, West, & Aiken, 2003). The variability of model parameters as a function of the unreliability of scores can be inferred by means of confidence intervals.

Bootstrapping and other re-sampling techniques complement traditional confidence intervals by estimating standard errors of parameter estimates over a large number of hypothetical sample draws (Hancock & Nevitt, 1999). Results from the bias-corrected bootstrap procedure are delineated in Appendix 2.

The results of the bias-corrected bootstrapping indicate that the 95% confidence intervals between model parameters are

**TABLE 10:** Average population estimations with Mean Square Error, 95% coverage, proportion of replications equal to zero under H0, and parameter bias.

| Model results: Estimates | Population | Average | Standard deviation | SE average | MSE | 95% Coverage | Percentage significant coefficients | Bias |
|---|---|---|---|---|---|---|---|---|
| **INT BY** | | | | | | | | |
| INT_NET | 0.467 | 0.3498 | 0.3078 | 0.0156 | 0.1084 | 0.825 | 1.000 | 25% |
| INT_COM | 0.375 | 0.2814 | 0.2478 | 0.0181 | 0.0701 | 0.819 | 1.000 | 24% |
| INT_PDIS | 0.913 | 0.6850 | 0.6025 | 0.0369 | 0.4146 | 0.834 | 1.000 | 25% |
| INT_MOFT | 0.947 | 0.7092 | 0.6243 | 0.0312 | 0.4459 | 0.830 | 1.000 | 25% |
| **Intercepts** | | | | | | | | |
| INT_NET | 2.742 | 2.7414 | 0.0214 | 0.0214 | 0.0005 | 0.953 | 1.000 | 0.02% |
| INT_COM | 2.798 | 2.775 | 0.0209 | 0.0216 | 0.0004 | 0.958 | 1.000 | 0.80% |
| INT_PDIS | 5.461 | 5.4594 | 0.0458 | 0.0468 | 0.0021 | 0.955 | 1.000 | 0.02% |
| INT_MOFT | 5.584 | 5.5828 | 0.0421 | 0.0432 | 0.0018 | 0.946 | 1.000 | 0.02% |
| **Variances** | | | | | | | | |
| INT | 1.000 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 1.000 | 0.000 | |
| **Residual variances** | | | | | | | | |
| INT_NET | 0.013 | 0.0129 | 0.0018 | 0.0018 | 0.0000 | 0.942 | 1.000 | 0.76% |
| INT_COM | 0.094 | 0.0937 | 0.0061 | 0.0061 | 0.0000 | 0.943 | 1.000 | 0.31% |
| INT_PDIS | 0.268 | 0.2659 | 0.0185 | 0.0181 | 0.0003 | 0.942 | 1.000 | 0.78% |
| INT_MOFT | 0.042 | 0.0417 | 0.0068 | 0.0069 | 0.0000 | 0.957 | 1.000 | 0.71% |

INT_NET, Interaction Networking; INT_COM, Interaction Communication; INT_PDIS, Interaction Promoting Diversity and Intercultural Sensitivity; INT_MOFT, Interaction Motivating Others and Fostering Teamwork.

quite broad, which erodes confidence in the replication of specific point estimates in the population. In addition, the difference between the population parameter estimates and mean values recovered by Monte Carol draws, indicates substantial bias in the parameter estimates. The same conclusion can be reached with regard to the standard errors.

Considering all this information collectively, one would have to conclude that the construct validity evidence for the original CBST is limited. For example, the overall measurement model did not converge and was eventually abandoned because of high multicollinearity between dimension ratings. Consequently, only a small subsection of the measure was further investigated with additional analysis. Even these models of the Interaction dimension required extensive modification and manipulation before they showed acceptable fit to the data.

More supportive evidence for construct validity was found with regard to the revised Interaction meta-competency after the sub-dimensions of Motivating Others (INT_MO) and Fostering Teamwork (INT_FT), as well as the sub-dimensions of Promoting Diversity (INT_PD) and Intercultural Sensitivity (INT_IS) were combined. The newly combined sub-dimensions were labelled INT_PDIS (Promoting Diversity and Interpersonal Sensitivity) and INT_MOFT (Team Motivation). Theoretically it makes sense to group the dimension ratings of Promoting Diversity and Interpersonal Sensitivity, as well as Motivating Others and Fostering Teamwork.

# Discussion

## Outline of the results

The primary research objective was to examine the construct validity of an electronic in-basket using CBST technology. In the end, only a revised version of one of the six meta-competencies could be assessed. The results suggest that AC methodologies packaged in interactive software applications are not immune to the problems that face traditional sample-based assessments. Multicollinearity remains a particularly thorny issue, in part, because not enough consideration is awarded to the conceptual definition of competencies at the design stage. However, this problem does not seem to be unique to the current study. Hoffman et al. (2011a, 2011b) found that narrow dimensions grouped together in broader dimensions provide better model fit than models that contain many dimensions. Thus, when dimensions were modelled in a way that took the similarity between them into account, it was possible to find evidence for dimension factors in ACs. Similarly, Kuncel and Sackett (2014) found that greater levels of aggregation compound common variance and reduce error variance. This should improve convergence and fit in CFA models. An alternative explanation that cannot be ruled out is the impact of frame-of-reference training on assessor ratings. Often rigorous frame-of-reference training results in ratings that are very similar for multiple raters of the same participant in the AC. Although this practice may promote inter-rater reliability it may restrict the range of dimension ratings. In addition, the high correlations between dimensions further suggest that the scoring mechanism failed to indicate the relative differences between dimensions in a single simulation. However, this result is not unique. Previous AC research consistently found that behaviour is consistent across dimensions within exercises, rather than across exercises of the same dimensions.

The lack of correspondence among the same dimension observations across AC exercises has often been regarded as problematic, and several innovative interventions have been proposed to remedy the problem. However, proponents of the exercise-centric ideology will highlight the link between exercise effects and criterion scores. However, recent studies suggest that ACs have been misspecified and as a result the contribution of dimensions have historically been underestimated in AC ratings. This holds important implications for practice because most AC applications are probably still expressed in dimension-centric discourse.

In the second round of data analysis, we investigated the model parameters by way of two re-sampling techniques, namely, Monte Carlo simulations and bias-corrected bootstrapping. Confidence intervals were provided from the bias-corrected draws to assess the variability of model parameters because of the calculated population standard errors. In accordance with the original results, we found the bootstrap confidence intervals to be quite wide and coverage levels below the suggested level of 0.90. This provided further support that the results should be interpreted with caution because estimates may be biased. More specifically, these techniques may provide AC scholars and practitioners with another set of tools to assess the validity of ratings, especially when samples are relatively small. We may have arrived at a different conclusion regarding the validity of the revised Interaction dimension, albeit not the whole in-basket, if these two re-sampling techniques had not been employed. In general, the CFA results of the revised Interaction model showed satisfactory fit, low residuals and robust factor loadings. However, the re-sampling techniques indicate that the results may not be trustworthy and may be because of type I errors. These two approaches provide valuable tools for AC practitioners and researchers who often have to conduct research with very small sample sizes.

### Limitations of the study

Although the study provided a lot of useful findings, there are some conflicting results that need to be reported. One of the biggest limitations is that only one exercise type, namely, an in-basket, was used in the current study. This made the specification and estimation of method effects impossible. Typically, the size of the exercise effects provides important information regarding the functioning and internal structure of simulations. Based on best practice guidelines for the use of the AC method in South Africa, Meiring and Buckett (2016) stipulate that a single electronic in-basket exercise does not constitute an AC. This stipulation originates from the classic definition of an AC, which posits that multiple exercises and multiple observers are key differentiators between ACs and other evaluation methods. For this reason, results reported in the current study cannot be generalised to other traditional ACs.

Another limitation was the structure and design of the CBST. Ratings of competencies were reflected as PEDRs and not behavioural indicators. This greatly limited the number of data points to specify each of the meta-competencies. If meta-competencies were specified with behavioural indicators, the researcher could delete behavioural indicators that demonstrated collinearity, yet still measure the six meta-competencies. However, in the current study the researchers could only specify and evaluate the Interaction meta-competency because the other competencies had too few PEDR scores to combine into broader competencies.

An additional limitation of this study is that the performance ratings from managers, as well as the success of a follow-up supervisory development programme could not be investigated. As a result, the criterion validity of the six meta-competencies and job performance could not be investigated. It would have been interesting to see if differences on the meta-competencies translated into significant criterion-related differences.

### Practical implications

The research value and contribution of this study can be best described by discussing multiple perspectives. From a practical perspective, this application of CBST demonstrates that faster, more accurate solutions exist for conducting ACs for the purposes of selection and development. From a theoretical perspective, the research results and learning points from the CBST in-basket exercise depict the real-life events of the manager and may act as a workplace or business simulation, which adds to the incremental validity of selection or development strategy. From a corporate perspective, the accelerating rate of change and the increasing uncertainty in the outcomes of change are evident across the whole business arena. This is enhanced by the increased demand for experienced talent. From a research perspective, there is considerable bias in model parameters when using small samples. However, bias-corrected bootstrapping techniques and Monte Carlo simulations may be used productively to evaluate the bias in model parameters.

## Conclusion

This study set out to evaluate the construct validity of an electronic in-basket by investigating the internal structure of the exercise. The selection of competencies was based on job analyses and each of the six meta-competencies has a number of sub-dimensions. This design is similar to traditional AC exercises. The initial goal of the study was to assess the internal structure of the entire in-basket with a CFA methodology using MTMM matrices. However, initial statistical screening of the data suggested that a large number of dimensions were highly correlated and lacked discriminant validity. To remedy this problem, dimensions were collapsed whenever it made theoretical sense to do so. In the end, only one meta-competency, the Interaction dimension, could be evaluated with a CFA approach. The results showed that the proposed model fitted the sample data well.

However, results from the two re-sampling techniques suggested that the model parameters were contaminated by bias and may lead to invalid inferences. This study demonstrates how these two techniques can be used when using CFA approaches in small samples. Finally, the study demonstrates that for all the potential benefits associated with electronic- and Internet-delivered simulations, the *a priori* design and scoring mechanism should comply with best practice if one hopes to find support for construct validity in AC ratings.

# Acknowledgements

## 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

J.B. wrote the article and conducted the statistical analyses. D.M. conceptualised the literature review and J.H.v.d.W. collected the data and wrote the original master's thesis which forms the basis of the current article.

## Funding information

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

## Data availability statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

## Disclaimer

The views expressed in this article are of the authors' own, and do not represent the position of any related institutions or funding agencies. No funding was received by the authors during the process of completing the research contained in the current article.

# References

Arthur, W. Jr., 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*(1), 125–153. https://doi.org/10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., Jr., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternative view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology, 1*(1), 105–111. https://doi.org/10.1111/j.1754-9434.2007.00019

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research, 21*(2), 205–229. https://doi.org/10.1177/0049124192021002004

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

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

Brits, N., Meiring, D., & Becker, J. R. (2013). Investigating the construct validity of a development assessment centre. *South African Journal of Industrial Psychology, 39*(1), 1–13. https://doi.org/10.4102/sajip.v39i1.1092

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. London: Routledge.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105 Retrieved from psycnet.apa.org.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hamilton, NJ: Hamilton Printing Company.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16–29. https://doi.org/10.1037/1082-989X.1.1.16

Davidian, S. (2005). *Simulation studies in statistics. What is a Monte Carlo study?* Retrieved from http://www4.stat.ncsu.edu/~davidian/st810a/simulation_handout.pdf.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Enders, C. K. (2005). An SAS macro for implementing the modified Bollen-Stine bootstrap for missing data: Implementing the bootstrap using existing structural equation modeling software. *Structural Equation Modeling, 12*, 620–641. https://doi.org/10.1207/s15328007sem1204_6

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. https://doi.org/10.1037/0021-9010.72.3.493

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology, 55*(1), 137–163. https://doi.org/10.1111/j.1744-6570.2002.tb00106.x

Hancock, G. R., & Nevitt, J. (1999). Bootstrapping and the identification of exogenous latent variables within structural equation models. *Structural Equation Modeling, 6*(4), 394–399. https://doi.org/10.1080/10705519909540142

Hermelin, E., Lievens, F., & Robertson, I. (2007). The validity of assessment centre for the prediction of supervisory performance rankings: A meta-analysis. *International Journal of Selection and Assessment, 15*(4), 405–411. https://doi.org/10.1111/j.1468-2389.2007.00399.x

Hoffman, B. J. (2012). Exercises, dimensions, and the Battle of Lilliput: Evidence for a mixed-model interpretation of AC performance. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 281–306). New York: Routledge.

Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011a). Center validity exercises and dimensions are the currency of assessment centers. *Personnel Psychology, 64*(2), 351–395. https://doi.org/10.1111/j.1744-6570.2011.01213.x

Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011b). Exercises and dimensions are the currency of assessment centres [Electronic version]. *Personnel Psychology, 64*(2), 351–395. https://doi.org/10.1111/j.1744-6570.2011.01213.x

Howard, A. (2008). Making assessment centres work the way they are supposed to. *Industrial and Organizational Psychology, 1*, 98–104. https://doi.org/10.1111/j.1754-9434.2007.00018.x

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

IBM Corporation. (2017). *IBM SPSS statistics for windows*. New York: IBM Corporation.

International Task Force on Assessment Center Guidelines. (2015). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*(3), 243–253. https://doi.org/10.1111/j.1468-2389.2009.00467.x

Jackson, D., Lance, C. E., & Hoffman, B. (2012). *The psychology of assessment centres.* London: Routledge.

Joreskog, K. G., & Sorbom, D. (1993). *USREL vi: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago, IL: National Educational Resources.

Kim, H., & Millsap, R. (2014). Using the Bollen-Stine bootstrapping method for evaluating approximate fit indices. *Multivariate Behavioral Research, 49*(6), 581–596. https://doi.org/10.1080/00273171.2014.947352

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review, 1*(2), 128–146. https://doi.org/10.1177/2041386610387000

Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology, 99*(1), 38–47. https://doi.org/10.1037/a0034147

Lance, C. E. (2008). 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. https://doi.org/10.1111/j.1754-9434.2007.00028.x

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*(2), 377–385. https://doi.org/10.1037/0021-9010.89.2.377

Lance, C. E., Woehr, D. J., & Meade, A. W. (2005, April). *A Monte Carlo investigation of assessment centre construct validity models.* Published paper delivered at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Lance, C. E., Woehr, D. J., & Meade, A. W. (2007). Case study: A Monte Carlo investigation of assessment center exercise factors represent cross-situational specific, not method bias. *Human Performance, 13*(4), 323–353.

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

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. 68–94). New York: Routledge.

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

Lievens, F., Dilchert, S., & Ones, D. S. (2009). The importance of exercise and dimension factors in assessment centers: Simultaneous examinations of construct-related and criterion-related validity. *Human Performance, 22*(5), 375–390. https://doi.org/10.1080/08959280903248310

Lievens, F., & Thornton, G. C., III. (2005). Assessment centers: Recent developments in practice and research. In A. Evers, O. Smit-Voskuijl, & N. Anderson (Eds.), *Handbook of selection* (pp. 243–264). New Jersey: Blackwell Publishing.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*(1), 99–128. https://doi.org/10.1207/s15327906mbr3901_4

Meiring, D., Becker, R. J., Gericke, S., & Louw, N. (2015). Assessment centers: Latest developments on construct validity. In I. Nikolaou & J. K. Oostrom (Eds.), *Employee recruitment, selection, and assessment. Contemporary issues for theory and practice* (pp. 190–206). London: Psychological Press-Taylor & Francis.

Meiring, D., & Buckett, A. (2016). Best practice guidelines for the use of the assessment centre method in South Africa. *SA Journal of Industrial Psychology, 42*(1), 1–15. https://doi.org/10.4102/sajip.v42i1.1298

Meiring, D., & Van der Westhuizen, J. H. (2011). Computer-based simulation technology as part of the AC and DAC: A global South African review. In N. Povah & G. C Thornton (Eds.), *Assessment and development centres: Strategies for global talent management*. Surrey, UK: Gower Publishing Ltd.

Millsap, R. E. (2012). A simulation paradigm for evaluating model fit. In M. Edwards & R. MacCallum (Eds.), *Current issues in the theory and application of latent variable models* (pp. 165–182). New York: Routledge.

Monahan, E. L., Hoffman, B. J., Lance, C. E., Jackson, D. J. R., & Foster, M. R. (2013). Now you see them, now you do not: The influence of indicator-factor ratio on support for assessment center dimensions. *Personnel Psychology, 66*(4), 1009–1047. https://doi.org/10.1111/peps.12049

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (7th edn.). Los Angeles, CA: Muthén & Muthén.

Nikolaou, I., Anderson, N., & Salgado, J. (2012). Special issue on advances in selection and assessment in Europe. *International Journal of Selection & Assessment, 20*(4), 383–384. https://doi.org/10.1111/ijsa.12000

Rose, S., Spinks, N., & Canhoto, A. I. (2015). *Management research: Applying principles*. New York: Routledge Taylor and Francis.

Sackett, P. R., Lievens, F., Van Iddekinge, C., & Kuncel, N. (2017). Individual differences and their measurement: A review of 100 years of research. *Journal of Applied Psychology, 102*(3), 254–273. https://doi.org/10.1037/apl0000151

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

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology, 102*(10), 1435. https://doi.org/10.1037/apl0000236

Schmidt, F. L., & Oh, I. (2015). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research findings.* Working Paper. https://doi.org/10.13140/RG.2.2.18843.26400

Schmitt, N. (Ed.). (2012). *The Oxford handbook of personnel selection and assessment.* New York: Oxford University Press.

Steiger, J. H., & Lind, J. C. (1980). *Statistically-based tests for the number of common factors*. Paper presented at the Annual Meeting of the Psychometric Iowa, Taylor and Francis, Iowa City.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (7th edn.). New York: Pearson.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*(3), 500–517. https://doi.org/10.1037/0021-9010.88.3.500

Thornton, G. C., & Gibbons, A. M. (2009). Validity of assessment centres for personnel selection. *Human Resource Management Review, 19*(2), 169–187. https://doi.org/10.1016/j.hrmr.2009.02.002

Thornton, G. C., III, Mueller-Hanson, R. A., & Rupp, D. E. (2017). *Developing organizational simulations: A guide for practitioners, students, and researchers*. London: Routledge.

Thornton, G. C., III, & Rupp, D. E. (2012). Research into dimension-based assessment centers. In D. J. Jackson, B. J. Hoffman, & C. E. Lance (Eds.), *The psychology of assessment centers* (pp. 141–170). New York: Routledge.

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

Wang, L., Fan, X., & Wilson, V. L. (1996). Effects of non-normal data on parameter estimates and fit indices for a model with latent and manifest variables: An empirical study. *Structural Equation Modeling, 3*(3), 228–247. https://doi.org/10.1080/10705519609540042

Appendixs starts on the next page →

# Appendix 1

**TABLE 1-A1:** Bivariate correlations between assessment centres indicators.

| Assessment centres indicators | Bivariate correlations |
|---|---|
| INT-PD AND INT_IS | 0.954 |
| INT_FT AND INT_MO | 1.000 |
| LRN_CO AND INT_MO | 1.000 |
| LRN_CO AND INT_FT | 1.000 |
| LRN_HF AND INT_MO | 1.000 |
| LRN_HF AND INT_FT | 1.000 |
| LRN_HF AND LRN_CO | 1.000 |
| ENT_CUST AND ENT_QUAL | 1.000 |
| EXE_ASS AND LRN_BA | 0.995 |
| EXE_DR AND LRN_BA | 1.000 |
| EXE_DR AND EXE_ASS | 0.955 |
| EXE_DM AND LRN_BA | 1.000 |
| EXE_DM AND EXE_ASS | 0.955 |
| EXE_DM AND EXE_DR | 1.000 |
| DRI_SD AND DRI_PAS | 1.000 |
| DRI_LS AND INT_MO | 0.994 |
| DRI_LS AND INT_FT | 0.994 |
| DRI_LS AND INT_FT | 0.994 |
| DRI_LS AND LRN_CO | 0.994 |
| DRI_LS AND LRN_HF | 0.994 |
| VIS_LC AND ENT_INT | 0.996 |
| VIS_VT AND VIS_SO | 0.997 |

INT_PD, Promoting Diversity; INT_IS, Intercultural Sensitivity; INT_MO, Motivating Others; INT_FT, Fostering Teamwork; INT_NET, Networking; INT_COM, Clear and Open Communication; LRN_CO, Coaching Others; LRN_HF, Handling Feedback; LRN_SR, Self-Reflection; LRN_BA, Building up Business Acumen; ENT_INT, Integrity (In Business); ENT_QUAL, Quality Orientation; ENT_PROF, Profit Orientation; ENT_CUST, Customer Orientation; EXE_ASS, Assertiveness; EXE_DR, Delivering Results; EXE_DM, Decision Making; EXE_PS, Problem Solving; DRI_PAS, Passion and Commitment; DRI_SD, Self-Direction; DRI_LS, Leading and Steering; DRI_INI, Initiative; VIS_LS, Leading Change; VIS_INN, Innovation; VIS_SO, Strategic Orientation; VIS_VT, Visionary Thinking.

# Appendix 2

**TABLE 1-A2:** Bias-corrected bootstrap results.

| Confidence intervals of model results | Lower 0.5% | Lower 2.5% | Lower 5% | Estimate | Upper 0.5% | Upper 2.5% | Upper 5% |
|---|---|---|---|---|---|---|---|
| **INT BY** | | | | | | | |
| INT_MOFT | 0.740 | 0.789 | 0.817 | 0.947 | 1.094 | 1.117 | 1.168 |
| INT_NET | 0.333 | 0.367 | 0.386 | 0.467 | 0.552 | 0.565 | 0.590 |
| INT_COM | 0.232 | 0.272 | 0.290 | 0.375 | 0.462 | 0.478 | 0.506 |
| INTPDIS | 0.550 | 0.638 | 0.688 | 0.913 | 1.125 | 1.160 | 1.222 |
| **Intercepts** | | | | | | | |
| INT_NET | 2.606 | 2.639 | 2.657 | 2.742 | 2.823 | 2.837 | 2.869 |
| INT_COM | 2.664 | 2.695 | 2.710 | 2.798 | 2.881 | 2.897 | 2.927 |
| INT_PDIS | 5.157 | 5.221 | 5.261 | 5.461 | 5.630 | 5.661 | 5.725 |
| INT_MOFT | 5.313 | 5.380 | 5.412 | 5.584 | 5.749 | 5.780 | 5.838 |
| **Variances** | | | | | | | |
| INT | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **Residual variances** | | | | | | | |
| INT_NET | 0.001 | 0.004 | 0.005 | 0.013 | 0.022 | 0.024 | 0.029 |
| INT_COM | 0.043 | 0.053 | 0.059 | 0.094 | 0.150 | 0.162 | 0.183 |
| INTPDIS | 0.103 | 0.140 | 0.158 | 0.268 | 0.424 | 0.452 | 0.509 |
| INT_MOFT | 0.012 | 0.004 | 0.011 | 0.042 | 0.075 | 0.081 | 0.093 |

INT, Interaction; COM, Clear and Open Communication; NET, Networking; FT, Fostering Teamwork; MOT, Motivating others; IS, Intercultural Sensitivity; PD, Promoting Diversity.
**, Correlation is significant at the 0.01 level (2-tailed).