Original Article

The Impact of Item Parceling Ratios and Strategies on the Internal Structure of Assessment Center Ratings

A Study Using Confirmatory Factor Analysis

Anne Buckett[1], Jürgen Reiner Becker[2], and Gert Roodt[3]

[1]Precision ACS and University of Johannesburg, South Africa
[2]Department of Industrial Psychology, University of the Western Cape, and University of Johannesburg, South Africa
[3]Department of Industrial Psychology and People Management, University of Johannesburg, South Africa

Abstract. The aim of the present study was to investigate whether using item parcels instead of single indicators would increase support for the factorial validity of assessment center (AC) ratings in factor analytic applications. Factor analytic analyses of AC ratings are often plagued by poor model fit as well as admissibility and termination problems. In the present study, three purposive item parceling strategies, in conjunction with three parceling approaches (specifying different ratios of indicators to dimensions), were investigated in relation to five confirmatory factor analysis specifications of AC ratings across two AC samples (Sample 1: *N* = 244; Sample 2: *N* = 320). The findings were equivocal across the two samples. Nonetheless, a three-parcel approach using a factorial allocation strategy performed better that a one-parcel approach (akin to the postexercise dimension rating).

Keywords: assessment centers, confirmatory factor analysis, factorial validity, internal structure, item parceling

Assessment centers (ACs) have become a mainstream method in personnel selection due to their positive correlation with job performance (Arthur et al., 2003; Gaugler et al., 1987; Hermelin et al., 2007). However, there is less consensus about the internal structure of AC ratings and how best to represent them in scoring, interpretation, and analysis (Howard, 1997; Jackson et al., 2016; Kuncel & Sackett, 2014; Lance, 2008; Putka & Hoffman, 2013). Historically, the AC has seemed to be a better measure of the exercises used to gather evidence about candidate behavior than of the dimensions it purports to measure (Jackson et al., 2016; Lance, 2008; Lievens et al., 2009; Sackett & Dreher, 1982).

A key finding from contemporary AC research is that the internal structure of AC ratings is likely to reflect candidate performance across a number of variables including (1) dimensions, (2) exercises, and (3) a number of other factors that are captured by a general performance factor (GPF; Hoffman et al., 2011; Merkulova et al., 2016). This contemporary structure suggests that a candidate's behavior, explained by dimensions, can also be examined in relation to their performance in different exercises and in relation to certain contextual factors such as individual differences and cognitive ability. According to this perspective, AC performance is cross-situationally specific due to the unique demands placed on the candidate in each simulation. However, despite the research findings enabling a mixed-method view of the internal structure of AC ratings, exercise variance still dominates candidate performance in the AC (Jackson et al., 2016; Siminovsky et al., 2015). This finding in and by itself is not problematic because exercise effects, in combination with dimensions, are valid sources of variance in AC ratings when a mixed-method design approach is followed. The problem is that most practitioners still ignore exercise variance when designing ACs and give feedback on candidate performance, which is often based exclusively on AC dimensions (i.e., a candidate's demonstrated behavior in relation to the dimensions measured in an exercise; Krause et al., 2011). If AC practitioners base their personnel decisions purely on dimensions, all exercise variance will be regarded as error variance rather than as a valid source of variance.

Notwithstanding, the conceptual challenges with using a dimension and exercise-based approach in AC design are the problems associated with using factor analytic approaches when investigating AC ratings (Howard, 2008; Jackson et al., 2016; Lievens, 2009; Thornton & Gibbons, 2009). Most confirmatory factor analysis (CFA) models used to investigate AC structures are plagued by out-of-bound parameters and model termination issues (Monahan et al., 2013). Furthermore, sample size, model complexity, and limits of traditional approaches to variance estimation are likely to lead to misspecified and confounded models (Jackson et al., 2016; Putka & Hoffman, 2013). This can be attributed in part to methodological problems associated with the specification of AC ratings in CFA models.

To this end, some research has focused on investigating alternative techniques, such as generalizability theory analysis (G theory; Bowler & Woehr, 2006; Cahoon et al., 2012; Jackson et al., 2016; Putka & Hoffman, 2013) and hybrid configurations of CFA (Hoffman et al., 2011; Merkulova et al., 2016; Monahan et al., 2013). Although these approaches have provided more insight into the proportion of dimension and exercise contributions in AC ratings, the findings still show strong support for exercises instead of dimensions (Jackson et al., 2016).

In the quest for finding more equitable sources of variance in AC ratings, the attention has shifted to possible methodological artifacts inherent in conventional factor analytic approaches used for investigating the internal structure of ACs. Resolving methodological issues inherent in AC research should reveal more consistent and accurate findings regarding the internal structure of AC ratings. It therefore remains important to investigate alternative procedures that bring the field closer to addressing the persistent disagreement regarding the appropriate analytical approach to assess the internal structure of AC ratings (Arthur et al., 2008; Thornton & Gibbons, 2009). Finding unbiased estimates of exercise and dimension variance largely determines the most effective strategies for integrating this information into the design and interpretation of human resource interventions.

Recently, Monahan et al. (2013) found evidence for the internal structure of AC ratings when increasing the indicator to dimension ratio prior to conducting CFA. When following this approach, Monahan et al. (2013) used a technique known as item parceling and were able to show increased model fit, model convergence, and admissibility across a wide variety of CFA configurations. Even so, while they found increased support for dimension effects, exercise variance still dominated across the CFA models. A second study replicating their approach found mostly similar results (Buckett et al., 2020). However, one of the key findings of the Monahan et al. (2013) research were the improved convergence and admissability rates of CFA models when increasing the indicator to dimension ratios. This has been a pervasive issue in AC research using CFA models. Accordingly, it can be argued that the difficulty to find equitable sources of dimension and exercise variance in AC ratings may be indicative of a statistical artifact inherent in CFA model specification. It thus seems likely that past CFA research, which predominantly makes use of postexercise dimension ratings (PEDRs), may not be equipped to find unbiased sources of variance due to misspecification of models.

The present study aims to contribute to the existing methodological debate by investigating different item parceling combinations (i.e., indicator to dimension ratios) in conjunction with selected item parceling allocation strategies prior to specifying CFA models, to determine whether such an approach leads to increased support for dimension effects, in addition to exercise effects, in AC ratings. This approach will be applied to five common CFA configurations typically used in research of AC ratings (see the paragraph preceding Research Question 3 below for an explanation of the five configurations).

Based on the above discussion, the current study makes four contributions to the literature. First, limited research has investigated parceling strategies in combination with a parceling approach when specifying CFA models of AC ratings. Despite parceling strategies being frequently used to specify CFA models in applied research (Bandalos, 2002; Little et al., 2002), this practice has not been prevalent in AC research. Given the recent proliferation of research focusing on the specification of AC ratings, we consider it important to investigate the impact of parcels on the internal structure of AC ratings. Second, the study aims to investigate whether using a specific parceling strategy (i.e., the way in which a researcher assigns items to parcels) will assist researchers to overcome some of the historical problems associated with specifying AC ratings with CFA models. Previous research on specifying CFA models indicates that researchers need to be mindful when assigning items to parcels, since the parceling strategy may have an impact on model termination, model fit, and out-of-bound estimates (Bandalos, 2002; Little et al., 2002, 2013; Nasser-Abu Alhija & Wisenbaker, 2006; Orcan, 2013). Third, the study aims to investigate whether there is an ideal combination of parceling strategy and the number of parcels to specify for AC ratings. This will serve to identify whether there is an interaction between the number of parcels and a specific parceling strategy. Finally, the study aims to investigate which CFA model configuration provides the best representation of the internal structure of AC ratings by taking into

consideration the parceling strategy and the number of parcels specified.

An Outline of the AC Method

Although the AC as a method of assessment is likely to be known to many readers of this article, its key features are outlined for context. The AC consists of job-related behavioral simulation exercises designed to measure specific dimensions relevant to job performance (Thornton et al., 2015). ACs are typically organized around dimensions and exercises. Dimensions are regarded as the focal constructs according to which candidate behavior is classified and evaluated. Focal constructs can also be conceived as behaviors in relation to specific job or task roles (International Task Force on Assessment Center Guidelines, 2015). A candidate's performance on the dimensions is typically used for feedback and decision making. Simulation exercises are the means by which information is collected about a candidate's dimensional performance. Exercises are based on job-related information and are designed to simulate a number of common workplace scenarios that employees must deal with in their daily jobs. They are designed to realistically recreate the work of an employee at a specific organizational level (Thornton et al., 2015).

Two additional components of the AC are the candidates and assessors. The AC usually includes several candidates who are assessed on the same dimensions and exercises during the AC in relation to a specific position. The AC also includes several trained assessors who observe candidates across a range of exercises (such as a role play exercise, group exercise, and in-basket exercise) and then evaluate their performance according to the targeted dimensions (International Task Force on Assessment Center Guidelines, 2015).

A typical AC process involves an assessor observing a candidate during a simulation exercise and recording the behavior of the candidate in relation to the simulations (International Task Force on Assessment Center Guidelines, 2015). Assessors rotate across exercises and candidates in order to observe and evaluate different candidates during the AC. At the end of each simulation exercise, one way to rate candidate performance is for the assessor to classify and evaluate the evidence provided by the candidate during the exercise using a dimension-based scoring sheet. This process is known as "*within-exercise*" scoring (Thornton et al., 2015). Assessors rate a number of behavioral indicators specified for each dimension and then calculate an overall dimension score known as a PEDR. The PEDR is most often used as the unit of analysis when investigating the internal structure of ACs when using a factor analytic approach. However, some scholars contend that PEDRs used in traditional factor analytic approaches are not optimal for investigating the internal structure of ACs as they represent single measures and may not be reliable (Howard, 2008; Kuncel & Sackett, 2014; Rupp et al., 2008).

The Internal Structure of AC Ratings

The frequent failure of research to find support for dimensions, in proportions similar to that of its exercises, as the focal constructs in AC ratings is problematic for practitioners and organizations because ACs claim to measure dimensions (in addition to exercises) and decisions are often made based on dimensional performance rather than exercise performance. When viewing the issue from this perspective, the lack of evidence supporting the internal structure of AC ratings would have serious implications for AC practitioners and organizations when it comes to making selection and development decisions.

Research conducted over 35 years has focused on resolving the "elusive" construct validity problem of AC ratings from two analytic perspectives, namely CFA and G theory. Three large-scale reviews have attempted to solve this problem using CFA approaches (cf. Bowler & Woehr, 2006; Lance et al., 2004; Lievens & Conway, 2001). Bowler and Woehr (2006) applied post hoc parameter constraints with the correlated-dimension correlated-exercise (CDCE) model and found that dimensions accounted for 22% of the variance while exercises accounted for 34% of the variance of PEDRs. Lance et al. (2004) used the same CFA model but arrived at a solution that included a general factor (27% variance) and multiple exercises (52% variance). Lievens and Conway (2001) used the correlated uniqueness CFA model and found that dimensions and exercises accounted for equal proportions of variance (34%), although they have been criticized for making inappropriate statistical assumptions. However, the dominance of exercises over dimensions was largely confirmed by the three reviews. Nevertheless, when reviewing these three studies it would appear that the biggest challenge facing CFA analysis pertains to issues of convergence and admissibility with model fit, which is exacerbated when PEDRs are used as the unit of measurement as they equate to single item measures (Woehr, Meriac, & Bowler, 2012).

On the other hand, research using G theory has focused on providing a greater understanding of the internal structure of AC ratings by specifying sources of reliable

and unreliable variance (Jackson et al., 2016; Putka & Hoffman, 2013). Putka and Hoffman (2013) first identified 15 AC-related effects and concluded that the level of aggregation had a direct bearing on the outcomes and resulting generalization practitioners could derive from AC ratings, especially in relation to dimension variance. Jackson et al. (2016) however found little to no impact on dimension variance, regardless of the level of aggregation. They credited this finding to the fact that they provided a more comprehensive list of sources of variance of AC ratings (29 in total) and still found that exercise variance was dominant compared to dimensions.

In finding substantially lower levels of dimension variance in AC ratings, several relevant reasons can be proposed, of which three are mentioned here. One explanation highlights the lack of a common taxonomy on AC dimensions (Arthur, 2012; Howard, 2008). Although attempts have been made to provide a standard taxonomy of dimensions (Arthur et al., 2003) and a framework for grouping dimensions into broader dimension factors (Hoffman et al., 2011; Meriac et al., 2014; Merkulova et al., 2016), these taxonomies have yet to be universally adopted by practitioners. The lack of consensus regarding the substantive definition of dimensions leads to expressions of diverse sets of behaviors by candidates on stimuli that should more or less measure the same dimensions. The consequence is that behavior that should be fairly consistent across the same dimension rarely manifests in this way. Another explanation points to the differences in candidate performance across different AC exercises. Research shows that these differences reflect systematic variance due to situational demands rather than bias in AC ratings (Gibbons & Rupp, 2009; Jackson et al., 2016; Lance, 2008; Lance et al., 2000; Putka & Hoffman, 2013). This view is analogous to contemporary views in the personality literature, showing that many candidates are likely to adapt their responses and behaviors depending on the type of situation they are dealing with (cf. cognitive-affective processing system [CAPS] theory, Mischel & Shoda, 1995; whole trait theory, Fleeson & Jayawickreme, 2015; and trait activation theory, Tett & Guterman, 2000). In the AC, this could see candidates demonstrating different behaviors for the same dimension being measured across exercises. For example, if the AC measures communication, then the candidate is likely to adapt the way in which they communicate in various settings. Therefore, the candidate may respond differently to a situation in writing, or when dealing with someone face-to-face, or when engaging with others in groups.

Yet a further possible explanation is the technique and level of analysis used to analyze AC ratings when conducting research on the internal structure of AC ratings (Howard, 2008; Lance, 2008; Thornton & Gibbons, 2009). Howard (2008) is of the view that PEDRs are not the appropriate level of input data that should be used to specify CFA models due to the unreliability of single indicators. Some authors argue that across exercise dimension ratings (AEDRs) should be used in multitrait–multimethod (MTMM) CFA configurations (Arthur et al., 2008; Rupp et al., 2008), although contemporary research shows limited construct-related validity evidence when AEDRs are used as the unit of measurement (Wirz et al., 2020).

These findings have key implications for the persistent debate in the AC literature: whether ACs measure the dimensions they are designed to measure (Lance, 2008; Sackett & Dreher, 1982). In addition, the value of CFA approaches to evaluate the internal structure of AC ratings has come into question as well as the appropriate level of aggregation (Arthur et al., 2000; Wirz et al., 2020). To this end, in response to the latter point, the current study aims to investigate whether using parcels contributes to evidence of improved model fit, interpretability, and admissibility of AC ratings using various CFA models. Specifically, the focus will be on how parcels are constructed prior to specifying the CFA models and by investigating the interaction between three parceling approaches and parceling strategies.

Item Parceling as a Way of Analyzing the Internal Structure of AC Ratings

Recent AC research supports mounting evidence that the misspecification of CFA models could lead to biased estimates of sources of variance in AC ratings (Kuncel & Sackett, 2014; Monahan et al., 2013). For example, in a recent study by Monahan et al. (2013), the authors found that using a higher ratio of indicators to dimensions to improve model termination and fit when using CFA models. This led the authors to conclude that "...it is important that research continues to analyze the psychometric properties of exercise and dimension variance in order to determine the most effective strategies to integrating this information into the design and interpretation of ACs" (Monahan et al., 2013, p. 1037).

In response to this call, we believe that parcels offer a potential remedy to some of the most pervasive problems associated with the specification and identification of AC ratings with CFA models. Simulation studies have found that using parcels instead of single indicators to specify CFA model leads to better model termination, admissibility, and fit (Bandalos, 2002). This finding can be ascribed to the improved measurement properties of parcels

when compared to single indicators (Little et al., 2013; Matsunaga, 2008). Although the use of item parceling as part of CFA studies is fairly common in applied psychology (Bandalos, 2002; Little et al., 2002), it is less often used in AC research. "Parceling involves summing or averaging item scores from two or more items from the same scale and using these parcel scores in place of the item scores in an . . . analysis" (Bandalos, 2008, p. 212). The parcels are then used, instead of the individual indicators, as manifest indicators of latent constructs (Little et al., 2013).

Recent research found that increasing the number of measurement points used during CFA eventually led to better support for the internal structure of AC ratings when considering the contribution of dimensions and exercises in AC ratings (Buckett et al., 2020; Monahan et al., 2013). Both studies found improved model–data fit and proper solutions for most of the models tested when using multiple indicators instead of single PEDRs, although to varying degrees. Nonetheless, when the ratio of indicators to dimensions increased, the net result was improved model fit and increased model convergence and admissibility. An important question still remained unanswered from the recent literature: whether the improvement in model fit was attributed to the higher indicator to dimension ratio or to the use of parcels.

Nonetheless, the use of parceling in the behavioral and social sciences has not been without criticism. One of the main points of contention pertains to the impact of unidimensionality on the outcomes of analyses using parcels. On the one hand, some researchers postulate that parceling improves unidimensionality (Little et al., 2013; Matsunaga, 2008) and consequently the normality of the distribution of scores (Little et al., 2013). These two aspects are important in that they are likely to increase model–data fit, improve reliability, and reduce model errors (Bandalos, 2002; Nasser-Abu Alhija & Wisenbaker, 2006; Orcan, 2013). On the other hand, and given how difficult unidimensionality is to achieve in behavioral and social sciences research, some researchers contend that when data are multidimensional, the use of parcels will not lead to an improvement in the unidimensionality of scores (Little et al., 2002). In fact, researchers argue that parceling in this instance will lead to misspecified and misrepresented models that obscure unmodeled factors, lead to biased parameter estimates, and confound the relative presence of unwanted sources of variance (Bandalos, 2002; Coffman & MacCallum, 2005; Little et al., 2002; Matsunaga, 2008).

Still, there are certain conditions where the considered use of parceling is appropriate (Coffman & MacCallum, 2005; Little et al., 2002; Nasser-Abu Alhija & Wisenbaker, 2006). For example, parceling may work well when data are not unidimensional, sample size is small, and the models being investigated are complex (Matsunaga, 2008). For these reasons, parcels may have a positive impact on CFA models used to specify the internal structure of AC ratings. At the very least, given how frequently parcels are used in behavioral and social sciences research, we believe it is timely that the impact of this practice is investigated in the context of the internal structure of AC ratings. In our own experience, anecdotal evidence suggests that practitioners may already make use of parcels when aggregating a large number of behavioral indicators together into more manageable scores for feedback and assessment purposes.

However, in order to specify parcels, multiple behavioral indicators need to be developed for each Dimension × Exercise combination. But practically speaking, there is probably a limit to the number of behavioral indicators that AC practitioners can develop from a cost and time perspective. More behavioral indicators may also call for the design of more comprehensive simulation exercises that activate finely grained trait-related behavior. Against the preceding background, the first research question is as follows:

> *Research Question 1*: Is there an ideal indicator to dimension and exercise ratio to find support for the internal structure of AC ratings?

A potential limitation of previous research is that these studies used a random allocation strategy to create parcels (Buckett et al., 2020; Monahan et al., 2013). In other words, behavioral indicators were randomly assigned to parcels prior to CFA. This strategy is completely permissible if the dimensions or exercises are unidimensional (Little et al., 2013; Matsunaga, 2008). However, given that AC ratings, by design, are multidimensional, it stands to reason that a random allocation strategy may not be the best solution in AC research when a parceling approach is employed. Therefore, investigating different parceling strategies prior to forming parcels may be warranted. At this point, however, little is known about the ideal parceling allocation strategy or the ideal number of parcels that should be used when investigating the internal structure of AC ratings.

In addition to investigating different numbers of parcels (i.e., one parcel, two parcels, and three parcels), this study will also investigate different parceling allocation strategies (i.e., random allocation, factorial allocation, and correlational allocation). A factorial allocation strategy uses exploratory factor analysis (EFA) to identify the factor loadings of items. Parcels are then created by grouping items together based on the size of the factor

loading (Matsunaga, 2008; Orcan, 2013). The aim of this process is to create parcels that are more unidimensional. A correlational allocation strategy creates parcels by grouping items together based on the size of the correlational coefficient (Matsunaga, 2008; Orcan, 2013; Williams & O'Boyle, 2008). The aim of this process is to create parcels that are more balanced and contain items that share specific reliable sources of variance. Both of these strategies should work well for data that are not unidimensional because they serve to create balanced parcels with less bias in parameter estimates (Little et al., 2013; Matsunaga, 2008). Hence, the second research question is formulated as follows:

Research Question 2: Is there an ideal parceling strategy to use in the analysis of AC ratings when parcels are used?

Furthermore, in order for us to contextualize our results to the extant literature (e.g., Bowler & Woehr, 2006; Lance et al., 2004; Lievens & Conway, 2001), three CFA models common across those studies were specified to test the impact of the number of parcels and the different parceling strategies. We expect that this combination of parceling approaches and allocation strategies may reveal differences in CFA model fit, model termination, and out-of-bound parameters. In addition, several contemporary studies suggest that a GPF explains significant portions of common variance in AC ratings and should be included in CFA model configurations (Bowler & Woehr, 2006; Jackson et al., 2016; Lance et al., 2004; Lance, Woehr, & Meade, 2007; Lievens & Conway, 2001; Merkulova et al., 2016). For this reason, two additional configurations that include GPF (CE + GPF and correlated dimensions [CD] + GPF) were added to our analysis strategy.

Nonetheless, this approach makes sense only to the extent that the same behavioral indicators are used to specify the item parcels across the different CFA configurations. In this regard, we deem it important to provide some clarity about our expectations. If there is no difference in the CFA model across parceling approaches and allocation strategies, we could arrive at two main conclusions. First, the parceling approach does not make a material impact on CFA results. In other words, the indicator to dimension or exercise ratio is not important. Second, the allocation strategy is not important, and we could probably make use of a simple random allocation strategy when employing a parceling approach. However, if significant differences are found between parceling approaches and allocation strategies, the research finding would suggest that practitioners should be more judicious in how parcels are constructed prior to specifying CFA models.

The five CFA models are briefly discussed. Model 1 is the correlated dimensions–correlated exercises (CDCE) model. This model assumes that dimensions and exercises are present in AC ratings (Bowler & Woehr, 2006; Lance, Foster et al., 2007). This model is similar to the MTMM approach used to establish discriminant and convergent validity (Campbell & Fiske, 1959). Model 2 represents the CD model and proposes no exercise factors. Model 3 represents the correlated exercises (CE) model and proposes no dimension factors. Model 4 and Model 5 include a GPF in addition to either dimensions or exercises. Thus, Model 4 is the CD plus a GPF (CD + GPF) model, and Model 5 is the CE plus a GPF (CE + GPF) model. Dimension factors were specified to be correlated with each other in the models that specified dimension factors and exercise factors to be correlated with each other in the models that specified exercise factors. Furthermore, exercise factors, dimension factors, and the GPF were specified to be uncorrelated with each other in Model 4 and Model 5. Thus, the third research question is as follows:

Research Question 3: Is there an ideal CFA model configuration that best represents the internal structure of AC ratings?

Method

Sampling and Data Collection Procedures

Sample 1 Participants

Data were collected from 244 participants completing a 1-day AC prior to attending a manager of other development programs. The participants were supervisors working for a chemical manufacturing and energy organization in South Africa. The ethnic composition of the sample was 46% Blacks, 36% Whites, 13% Indians, and 5% Coloreds.[1] The gender composition of the sample consisted of 68% males

[1] In South Africa, the four main ethnic groups are Black Africans, Whites, Indians, and Coloreds. These ethnic groups are used for statistical reporting in labor force reviews. "Colored" in this context indicates a person of mixed race with one parent who is White and one parent who is Black.

and 32% females. The mean age of participants was 39.86 years.

Sample 1 Data Collection and Procedure

The AC was designed to reflect the mixed-model approach. Using the competency framework provided by the organization, the AC designers identified dimensions that could be easily observed during the AC. As an additional measure, a desktop review was conducted of the most common jobs at this level in the organization to determine that the selected dimensions were appropriate to the context. This process resulted in five dimensions that were assessed during the AC: business acumen, communication, fostering relationships, leadership, and results driven. Furthermore, when designing the exercises, job-relevant situations were used, and these were aligned with dimensions that would be observable across the exercises. Three common simulation exercises were consequently designed specifically for the organization. The exercises included a role play exercise where the participant had to deal with a nonperforming staff member, a group exercise where participants had to work together to solve a mixture of five staff- and production-related problems, and an in-basket exercise where the participant had to respond to a number of items in writing. The exercises were designed so that each exercise measured between three and five dimensions. However, although the exercises were designed to tap into the same dimensions, the measurement of these dimensions was adjusted to each of the specific exercises. The dimensions measured within each exercise are listed in Table 1.

Behavioral indicators ranged from 5 to 16 per Dimension × Exercise combination. Rating forms were constructed as behavioral observation scales, and evidence was evaluated using a four-point scale where 1 = *Development area*, 2 = *Rounding off*, 3 = *On target*, and 4 = *Strength*. Assessors scored each behavioral indicator for a specific dimension within an exercise using the four-point scale. At the end of each exercise, assessors calculated a final PEDR for every dimension measured in the exercise.

Table 1. Dimensions measured within exercises for Sample 1

Dimensions	Role play exercise	Group exercise	In-basket exercise
Business acumen	X	X	X
Communication	X	X	
Fostering relationships	X	X	
Leadership	X	X	X
Results driven	X	X	X

Sample 2 Participants

Data were collected from 320 participants who were supervisors working for a government department. Participants were assessed for development purposes. The ethnic composition of the sample was 57% Blacks, 31% Whites, 7% Coloreds, and 5% Indians. The gender composition of the sample was 59% females and 41% males. Participant age was not captured.

Sample 2 Data Collection and Procedure

The AC consisted of two customized simulation exercises (i.e., a role play exercise and in-basket exercise) and a competency-based interview. Using the competency framework and a selection of supervisory job descriptions provided by the organization, seven dimensions were identified to be measured in the AC. The dimensions measured across the exercises included team management, customer service orientation, communication, interpersonal interaction, change orientation, planning and organizing, and problem solving and decision making. The dimensions measured within each exercise are listed in Table 2.

In this AC, each exercise measured five dimensions. However, in this sample, only five behavioral indicators per Dimension × Exercise were specified. A four-point rating scale was also used in this AC where 1 = *Major development need*, 2 = *Minor development need*, 3 = *Competent*, and 4 = *Strength*. The same scoring approach as described for Sample 1 was used, whereby assessors had to score each of the behavioral indicators on the four-point scale and then calculate the PEDR for each dimension for a specific exercise.

With regard to the internal structure of the AC models tested in the current study, it is important to note that the two ACs investigated in the current samples followed a traditional design approach by specifying dimensions nested in exercises. They were not designed to capture second-order factors, nor were they organized around empirically supported broad dimensions (cf. Hoffman

Table 2. Dimensions measured within exercises for Sample 2

Dimensions	Role play exercise	In-basket exercise
Change orientation		X
Communication		X
Customer service orientation	X	
Interpersonal interaction	X	X
Planning and organising	X	
Problem solving and decision making	X	X
Team management	X	X

et al., 2011; Merkulova et al., 2016). The competency framework was developed by looking at the inherent requirements of the targeted positions, and each of the dimensions and exercises were important from a development perspective. The data sets were however appropriate to investigate the research questions posed in the current study since a large number of behavioral indicators were developed to measure each of the Dimension × Exercise combinations.

Assessor Composition and Training

Assessors were experienced HR professionals and psychologists with experience in ACs. Assessors were ethnically diverse and representative of the participants completing the AC across both samples. In line with best practice guidelines assessors received one day of training for Sample 1 and 2 days of training for Sample 2, which included information about the organization, the purpose of the AC, and dimensions, and a practical component whereby each assessor evaluated and scored a simulated participant (International Task Force on Assessment Center Guidelines, 2015). In addition to frame-of-reference training, assessors were trained on the principles of behavioral observation, which included a focus on potential rater errors. Assessors were rotated across the exercises and participants, and the AC schedule was structured in such a way that assessors did not observe the same participant twice in the AC. However, only one assessor was responsible for scoring all dimensions within an exercise for one participant, and for this reason, it was not possible to calculate interrater reliability.

Analyses

Data were analyzed using Mplus 8 (Muthén & Muthén, 1998–2017). For meaningful statistical analysis and comparison, only dimensions that were measured across multiple exercises were included in the analysis. For Sample 1, all five dimensions were included in the analysis. For Sample 2, only three dimensions were retained for analysis since they were fully crossed. These were interpersonal interaction, problem solving and decision making, and team management.

Each of these models was tested across three parceling approaches and three parceling allocation strategies. The three parceling approaches were one parcel (note that this is the same as a PEDR whereby the overall dimension score per exercise is aggregated into a single score to create one parcel; (1P), two parcels (2P), and three parcels (3P). The three parceling allocation strategies included a random allocation strategy, factorial allocation strategy, and correlational allocation strategy. This resulted in 35 different combinations of parceling approaches and strategies for each of the two samples. Allocation of parceling strategies was only applied to 2P and 3P configurations because all behavioral indicators would have been included for the 1P CFA models. Note that the 1P approach is the baseline of comparison against the 2P and 3P approaches. This is because the PEDR is most often used to specify CFA models with AC data.

To create parcels using a factorial allocation strategy, EFA was conducted (Matsunaga, 2008; Orcan, 2013). Parcels were then created by allocating items to parcels on the rank-ordered factor loadings of the items. The factor loadings for the available behavioral indicators were ranked from highest to lowest. For example, to create two parcels, the item with the highest factor loading was allocated to Parcel 1 and the item with the second highest factor loading was allocated to Parcel 2. Thereafter, the order of allocated items was reversed such that the item with the third highest factor loading was allocated to Parcel 2 and the item with the fourth highest factor loading was allocated to Parcel 1. This alternating process was followed until all the items were allocated between the two parcels for every dimension × exercise combination.

To create parcels using a correlational allocation strategy, the correlation coefficients for each of the items from the interitem correlation matrix were used (Matsunaga, 2008; Orcan, 2013; Williams & O'Boyle, 2008). All items were included in the initial analysis that were obtained from Pearson's correlation coefficient matrix. Parcels were created by grouping pairs of items with the highest correlations into a parcel. To illustrate, the highest correlated pairs of items were allocated to Parcel 1 and the second highest pair was allocated to Parcel 2. Once each parcel had received a pair of items, the correlation analysis was re-run with the remaining items. The newly calculated interim parcel scores were used for the next grouping of items. In the new analysis, parcels were created by looking at the correlation between the interim parcel scores and the remaining items. Items were then assigned to Parcel 1 and Parcel 2. This procedure was followed until all the items had been assigned to a parcel.

The following fit indices were used to evaluate the models in this study: the standardized root mean squared residual (SRMR), the RMSEA, the Tucker–Lewis index (TLI), and the comparative fit index (CFI; Hair et al., 2010). The following cutoff scores were applied to indicate a good fit: SRMR < .08; RMSEA < .06; CFI and TLI ≥ .9 (Hair et al., 2010). In addition to model fit, models were evaluated according to convergence and admissibility.

Results

Each of the models tested contains parceling approach (1P, 2P, and 3P) and parceling allocation strategy (random, factorial, and correlational) factors.

For Sample 1, the best fitting solution across parceling approaches and allocation strategies for Model 1 (CDCE) was the 2P approach using a correlational strategy (SRMR = .05; RMSEA = .07; TLI = .95; CFI = .96). For Model 2 (CD), irrespective of the parceling approach and allocation strategy, all the CFA solutions reported poor model fit. For Model 3 (CE), the best fitting solution was our baseline model, the 1P approach (SRMR = .04; RMSEA = .06; TLI = .98; CFI = .99). For Model 4 (CD + GPF), the best fitting solution, albeit weak, was the 1P approach (SRMR = .10; RMSEA = .22; TLI = .73; CFI = .89). Although this still represents poor fit, in comparison to Model 2 (CD) for the same approach (SRMR = .17; RMSEA = .33; TLI = .39; CFI = .60), the fit indices were much better. For Model 5 (CE + GPF), the best fitting solution was our baseline model, the 1P approach (SRMR = .02; RMSEA = .03; TLI = .99; CFI = 1.00). Detailed model–data fit indices for the CFA models for Sample 1 can be found in Table E1 in the Electronic Supplementary Material (ESM 1).

For Sample 2, the best fitting solution for Model 1 (CDCE) was the 3P approach using a factorial strategy (SRMR = .02; RMSEA = .00; TLI = 1.01; CFI = 1.00). For Model 2 (CD), the same pattern of results as observed for Sample 1 emerged in that although all the tested configurations converged to a proper solutions, all the fit indices represent poor fit independent from the number of parcels and allocation strategies. For Model 3 (CE), the best fitting solution was our baseline model, the 1P approach (SRMR = .04; RMSEA = .13; TLI = .91; CFI = .95), although the RMSEA did not represent a good fit. For Model 4 (CD + GPF), the best fitting solution was the 3P approach using a factorial strategy (SRMR = .02; RMSEA = .01; TLI = 1.00; CFI = .99). Once again, the fit indices were better for this model when compared to the same configuration in Model 2 (CD; SRMR = .09; RMSEA = .14; TLI = .76; CFI = .81). For Model 5 (CE + G), the best fitting solution was the 3P approach using a factorial strategy (SRMR = .03; RMSEA = .04; TLI = .98; CFI = .99). Across the two samples, all the tested Model × Approach × Strategy combinations fit the data adequately, with the exception of Model 2 (CD), Model 3 (CE: specifically the RMSEA; Sample 2), and Model 4 (CD + GPF; Sample 1). Detailed model–data fit indices for the CFA models for Sample 2 can be found in Table E2 in ESM 1.

For model evaluations, variances and residual variances that were negative constituted out-of-bounds parameter estimates. In addition to model fit, models were further evaluated according to the number of out-of-bounds estimates to determine whether the model would be deemed admissible. Sample 1 data revealed two out-of-bounds estimates for Model 1 (CDCE: 1P; 2P + factorial strategy). These particular configurations also did not converge to an admissible solution. Sample 2 data revealed an out-of-bounds estimate for Model 4 (CD + GPF: 1P). The average factor loadings for Sample 1 were moderate for all models, except for Model 3, which produced higher factor loadings. The average factor loadings for Sample 2 were moderate for Models 1, 4, and 5, and higher for Models 2 and 3. Both the model fit indices and the out-of-bounds estimates across the two samples are comparable to, and partially support, findings from recent studies (Monahan et al., 2013) insofar as a random allocation strategy is used. A detailed summary of model parameters, including the number of out-of-bounds estimates and average standardized factor loadings across models for both samples, can be found in Table E3 (Sample 1) and Table E4 (Sample 2) in ESM 1.

In summary, in relation to Research Question 1 investigating an ideal indicator to dimension and exercise ratio that finds support for the internal structure of AC ratings, the 1P approach was the best fitting solution 40% of the time across both samples, while the 2P approach was the best fitting solution only 3% of the time and the 3P approach was the best fitting solution 17% of the time across both samples.[2] Additionally, when one considers that the 1P approach is only tested 10 times across the two samples and the 2P and 3P approaches were each tested 30 times across the two samples, the 1P approach clearly performed better than the 2P and 3P approaches. In response to Research Question 2 investigating the ideal parceling allocation strategy to use in the analysis of AC ratings, the random allocation strategy did not account for any of the best fitting models for either Sample 1 or Sample 2, while the correlational allocation strategy was the best fitting solution 10% of the time, and the factorial allocation strategy was the best fitting solution 20% of the time

[2] Since the 1P approach was constrained as our baseline approach across the five CFA model configurations, this approach only had 10 opportunities to emerge as the best fitting solution compared to 30 opportunities for the 2P and 3P approaches across the two samples. For this reason, the 1P approach reported a higher percentage of success since it was the best fitting approach for 4 of 10 models. The 2P approach was the best only once across 30 models, and 3P was the best model five times of 30 models.

across both samples. The random allocation strategy was not identified as the best fitting solution for any of the models tested across both samples, while the factorial allocation strategy was the best fitting solution in 4 of 20 models and the correlational allocation strategy was the best fitting solution in 2 of 20 models tested across both samples.

When considering the evidence collectively, the 3P factorial allocation strategy led to better fitting models across the five CFA model configurations. Furthermore, when comparing the random allocation strategy to the factorial allocation strategy and correlational allocation strategy, the random allocation strategy did not emerge as the best way to create parcels prior to the CFA of AC ratings.

For Research Question 3, which aims to identify the CFA model with the strongest statistical support for the internal structure of AC ratings (i.e., model fit, termination, and admissibility), collectively Model 5 (GE + GPF) and Model 1 (CDCE) performed the best in relation to the other models across both samples. In general, models containing exercises performed better than models containing dimensions (with the exception of Model 4 in Sample 2), but overall, models containing a GPF performed better than models without a GPF.

Discussion

The aim of this study was to investigate whether a specific parceling allocation strategy, in combination with a parceling approach, improved insights into the internal structure of AC ratings using a CFA approach. Earlier, in this article, we provided a basis for our research questions and outlined how the results would potentially impact our conclusions. Specifically, we indicated that should we find no significant differences in the CFA models tested across parceling approaches and allocation strategies, then we would conclude that neither the number of parcels nor the way in which we construct these parcels has any meaningful impact on CFA results. With this in mind, for Research Question 1, which concerns whether there is an ideal indicator to dimension and exercise ratio to investigate the internal structure of AC ratings, the overall pattern of results across the two samples demonstrated equivocal evidence in favor of one parceling configuration over another. Considering that the one parcel configuration was our baseline model that used no specific allocation strategy, it was the best fitting solution in 40% of models specified. This clearly lends support for using PEDRs across the two samples that we analyzed. In this respect, the answer to Research Question 1 is that a one parcel solution seems to work best across the five CFA models when the only criterion is the ratio of parcels to dimensions/exercises. This finding is however contrary to contemporary perspectives on what constitutes the most appropriate unit of measurement for specifying AC ratings in CFA models (Howard, 2008; Kuncel & Sackett, 2014; Monahan et al., 2013; Rupp et al., 2008). It also contradicts more recent AC research using parcels that generally found support for CFA models specified with higher indicator to factor ratios (Buckett et al., 2020; Monahan et al., 2013). However, a key limitation in both of these studies is that they did not investigate the impact of allocation strategies in conjunction with the use of parcels.

The results associated with Research Question 1 provided a nexus point to Research Question 2. More specifically, Research Question 2 investigated whether the parceling strategy has an impact on the CFA results, irrespective of the number of item parcels. Across both samples, the random allocation strategy did not once feature as part of the best fitting solutions. Stated differently, a dedicated parceling strategy seems to outperform a random allocation strategy when using multiple parcels to specify AC data with CFA models. In this case, the factorial allocation strategy was among the best fitting solution 20% of the time, while the correlational allocation strategy was among the best fitting solution 10% of the time. Thus, when considering all the evidence collectively, the factorial allocation strategy produced the best fitting models. Importantly, 6 of the 10 models performed better when using multiple parcels and a specific allocation strategy. Therefore, our finding indicates that the parceling strategy is important when assigning items to parcels.

However, when we look at the results specific to each sample and in relation to the number of parcels, the results are divergent. Without further research, it is difficult to draw definitive conclusions, but we posit that differences between the two ACs may provide answers to the divergent results. Differences in the design and construction of ACs have long been acknowledged as a potential limitation in AC research (Putka & Hoffman, 2013). We provide three examples of where the AC design might have impacted our findings. First, one potential explanation could be that the number of dimensions is a key variable when determining the best strategy. For example, Sample 1 used 5 dimensions for analysis, while Sample 2 used 3 dimensions for analysis. Second, in a similar vein, the number of behavioral indicators used during analysis could also be an important consideration. For instance, Sample 1 consisted of a larger number of behavioral indicators, including up to 16 indicators for certain Dimension × Exercise combinations. Sample 2, on the other hand, had no more than five behavioral indicators across the Dimension × Exercise

combinations. Therefore, analyses using a larger number of dimensions and/or behavioral indicators (as was the case in Sample 1) may favor the one parcel approach, while analyses using a smaller number of dimensions and/or behavioral indicators (as was the case in Sample 2) may favor the three-parcel (factorial allocation) approach. We further posit that when aggregating a large number of behavioral indicators, the correlated sources of common variance increasingly displace uncorrelated sources of variance, leading to more reliable input parcels (Kuncel & Sackett, 2014). This displacement of error variance is probably maximized when aggregating large numbers of items into a single parcel (as was the case with the one parcel models in Sample 1). Third, the small number of exercises used across both samples could also have produced problematic parameter estimates that tend to be endemic in factor analytic approaches (Jackson et al., 2016). Nonetheless, although this practice is fairly common in South Africa (Krause et al., 2011), we were able to mitigate this limitation by having item-level data available for analyses for each Dimension × Exercise combination.

Our findings once again highlight the challenges of using CFA in AC research to adequately understand the internal structure of AC ratings. To this end, G theory has been suggested as a suitable alternative to CFA (Jackson et al., 2016). There are three specific advantages of using G theory in AC research. First, it can be used to specify multiple sources of variance in AC ratings. Second, it can inform AC practitioners about the size of these sources of variance, which can then be applied to future AC design. Third, it allows researchers to have flexibility in how to treat variance, and therefore, it is possible to determine if, and in what circumstances, variance can be generalized (Alkharusi, 2012; Jackson et al., 2016; Shavelson & Webb, 1991). Therefore, G theory may indeed offer a more robust procedure for dealing with misspecification and model termination issues that tend to plague CFA. This, in turn, means that researchers can develop more finely grained ideas of the sources of common variance and error variance in AC ratings (Jackson et al., 2016). Given the benefits associated with G theory and the problems associated with AC CFA models, there have been calls to abandon the approach altogether (Arthur et al., 2000; Woehr, Putka et al., 2012).

However, we believe that abandoning CFA analyses of AC ratings may be premature. G theory has made strong contributions to the investigation of sources of variance in AC ratings, but the approach is not without limitations. First, G theory models are historically not concerned with testing the overall fit of models. The model fit approach is central to AC research as it allows the comparison of competing models and testing the nomological network of variables in which AC dimensions and exercises are embedded. As such, G theory aims to represent the average covariance structure among dimension–exercise units and does not use model fit indices to represent findings (Woehr & Arthur, 2003). Second, when viewed in relation to ease of use, G theory is not readily accessible to researchers and requires considerable computational resources (Alkharusi, 2012; Jackson et al., 2016). Third, when viewed in relation to graduate training, statistical education across tertiary institutions varies, and therefore, not all graduates may receive sufficient training in advanced analytical procedures such as G theory (Leech & Haug, 2015). Therefore, even if G theory is analytically more robust than CFA, some of the reasons listed here may impact how often G theory is used by researchers to investigate the internal structure of AC ratings.

However, the most promising development may be the integration of CFA and G theory of AC ratings to capitalize on the complementary strengths of both approaches to gain a better understanding of AC ratings. For example, G theory may be used to identify sources of variance, which can in turn be modeled with a CFA approach in relation to external criteria. This should be useful to interpret the systematic contribution of sources of variance in criterion ratings, thereby expanding the nomological network of exercise and dimension factors.

For Research Question 3, which concerns identifying the CFA model with the strongest statistical support for the internal structure of AC ratings, we considered the interaction between the parceling approach and the parceling strategy and collectively took into account the criteria of model fit, convergence, and admissibility. When applying these criteria, the exercises only + GPF (CE + GPF) model returned the best fitting solution across both samples, followed closely by CDCE. This finding is consistent irrespective of the number of parcels or the allocation strategy. Recent research confirms that CE + GPF is one of the best fitting models to describe the internal structure of AC ratings (Hoffman et al., 2011; Jackson et al., 2016; Merkulova et al., 2016), and so this finding is not unexpected in the current study. However, finding greater support for Model 1 (CDCE) comes as a surprise since these models are often plagued by issues of convergence and admissibility (Jackson et al., 2016; Monahan et al., 2013). Given that CDCE models specified with multiple behavioral indicators should result in greater convergence, this is not completely unexpected.

Although the debate regarding which source of variance is more important in AC ratings can now be regarded as settled, insofar as both exercise and dimension variance constitute true variance in AC ratings, it is still notable that CFA models that contain exercises outperform models with dimensions only. This finding is also consistent with the wider body of AC

research (Hoffman et al., 2011; Jackson et al., 2016) and is in line with situational performance explanations whereby the varying nature of participant behavior and performance across exercises is no longer regarded as random error but rather as an indication of substantive explanations of performance across various situations (e.g., CAPS theory, Mischel & Shoda, 1995; whole trait theory, Fleeson & Jayawickreme, 2015).

What is more important is that the CFA models that include both exercises and dimensions outperformed models with only exercises or dimensions. Another important finding is that when a GPF is added to CFA models that only contain dimensions or exercises, there was a noticeable improvement in model fit and model termination. These results indicate that a GPF explains sizable variance over and above exercises and dimensions. To date, the nomological network of a GPF has been tied to cognitive ability and personality (Hoffman et al., 2011; Merkulova et al., 2016), but it is likely that the actual components can be defined even more with further research. Since GPF is systematically linked to performance outcomes, we should also focus on unearthing the true nature of a GPF in AC ratings (Jackson et al., 2016) as it seems to be the component encapsulating many of the environmental and personal attributes that make a candidate successful in ACs (Thornton et al., 2015).

Implications and Recommendations for Practice

The present study aimed to contribute to the current debate on what is the most appropriate measurement input for specifying CFA models to investigate the internal structure of ACs. Recent studies indicated that most previous research may be bias due to model underrepresentation when using PEDRs as the indicators to CFA models (Kuncel & Sackett, 2014; Monahan et al., 2013). This can partially be explained by the persistent improper models when specifying exercises and dimensions in the same CFA models. Thus, the debate regarding the most appropriate level of AC ratings as well as the persistent disagreements between exercise or dimension-based interpretations of ACs may be largely obscured by improper CFA solutions. Monahan et al. (2013) argued that these findings can, in part, be explained by a statistical artifact of empirical underidentification of CFA models rather than any substantial theoretical explanation. For this reason, it may be premature to have a constructive discussion regarding the internal structure of ACs if the methodological problems surrounding the specification of AC ratings are not addressed conclusively. This study therefore aimed to investigate whether parceling approaches and allocation strategies offer a potential solution to the pervasive problems of weak model fit and high occurrences of improper solutions. In addition, the study aimed to see which CFA configuration represented the best fit to the data.

Our results will therefore be discussed under two broad themes. The first is whether there is an ideal parceling approach and strategy that leads to the best CFA solution in terms of model fit and admissibility. The overall findings seem to suggest that this is the three-parcel solution using a factorial allocation strategy. However, the difference between the aforementioned approach and the one parcel approach is negligible. Considering that most AC practitioners have historically used PEDRs is a convenient, albeit accidental, finding.

The second overarching theme is related to whether there is an ideal CFA model configuration irrespective of the parceling approach and strategy. Many of the problems relating to AC ratings in CFA models may be due to the misspecification of these models. Although G theory is a robust research tool to investigate the various sources of variance and has been recommended as a technique that may overcome many of the traditional issues inherent in CFA approaches, it has not necessarily completely revolutionized our thinking on the internal structure of AC ratings. For the most part, CFA and G theory results seem to converge. G theory may however arrive at a more nuanced taxonomy of exercise and dimension variance, but for the most part, the findings remain consistent, namely that exercise variance dominates dimension variance. For this reason, we believe that it is premature to dismiss CFA solutions that investigate AC ratings, even when considering the unresolved issues related to model termination and admissibility. Our analyses suggest that models containing exercises still outperform models that only contain dimensions. However, when adding a GPF, the fit, model termination, and admissibility seem to improve markedly. This leads us to speculate that a GPF is a stable and an important source of variance that should be modeled in AC data.

Practically, the findings seem to suggest that dimensions, exercises, and a GPF should be incorporated into AC design, feedback, and personal development plans. The primary problem with this recommendation is the conceptual definition of a GPF. However, a GPF is a multifaceted concept that probably encapsulates most of the environmental and personal trait interactions, which makes each individual unique (Thornton et al., 2015). The question that remains is how to improve on the GPF since it seems to be so multifaceted and may be quite difficult to segregate into clear sources of dimension and exercise variance. In part, the solution may be to interpret both sources of variance that we seem to understand relatively well, namely exercises and dimensions. This is in line with the mixed-method approach, which suggests that

candidate performance is best explained as the interaction across different situations in relation to dimensions nested in exercises (Hoffman, 2012).

We also see value in modeling broad dimensions, exercises, and GPF models in CFA AC ratings (cf. Hoffman et al., 2011; Merkulova et al., 2016). However, this only makes sense if the goal is to find the best fitting CFA model or when multiple behavioral indicators are not available to specify dimensions. When behavioral indicators are not available, dimensions with strong conceptual overlap can be grouped into broader dimensions that would provide more indicators to model dimensions and exercises, thus increasing the indicator to factor ratio. However, this approach may not be possible if the AC was developed to measure rather narrow and distinct dimensions that share very limited theoretical overlap. We also question the utility of this approach since the rich information gained from AC dimensions may be reduced to a limited set of very broad dimensions, which may see significant range restriction in scores. However, this approach may be useful when dealing with ACs that are very complex, based on a fragmented competency framework, or consist of too many dimensions that measure more or less the same thing. In this regard, there is much to be gained by unifying the conceptual framework and specifying broad models that fit empirical data better.

With regard to the internal structure of the AC models tested in the current study, the competency framework was developed specifically to operationalize the inherent requirements of the targeted positions and each of the dimensions and exercises were important from a development perspective. For this reason, it was not possible or useful to collapse distinct dimensions into broader dimensions. However, we think the approach has merits if it meets the criteria highlighted in the preceding section.

In summary, based on the extensive analysis conducted, it would seem that practitioners and scholars who are interested in investigating the internal structure of AC ratings using CFA would benefit from (1) specifying models that include dimensions and exercises, (2) including a GPF, (3) developing multiple indicators to measure dimensions, (4) and fitting CFA models that include at least three parcels[3]. However, G theory may be used to complement CFA approaches since it is more robust against some of the limitations of CFA. Using both approaches probably offers the most comprehensive view of the internal and external structure of ACs, and therefore, it makes sense to use them together where possible and practical. Furthermore, we believe that contemporary evidence suggests that a GPF represents an important feature of ACs that explain why individuals are effective in specific situations. However, it is also probably true that it is the most under-researched element of AC design (Jackson et al., 2016). Thus, a shift toward further exploring the nomological network and nature of a GPF is not only important but would further clarify how to incorporate GPF scores into feedback and personal development plans.

Limitations and Future Research

Several limitations and areas for future research are noted. First, this study was limited to two samples from South Africa. Future research is needed to see whether the results can be generalized to different settings. Second, only one assessor rated each candidate in the AC, and therefore, we were unable to calculate interrater reliability. This limitation was partly moderated in that assessor training was provided prior to the administration of each AC. It has been found that assessor training generally leads to quite high interrater reliability (Gorman & Rentsch, 2017). In addition, it has been found that sources of variance related to assessors often explain only a limited size of total or between-participant variance (cf. Jackson et al., 2016; Putka & Hoffman, 2013). Third, although we discuss research that represents the internal structure of AC ratings as exercises, broad dimensions, and a GPF, we did not explicitly test this model in the current study. This remains an important research topic especially when AC dimensions are specified broadly with overlapping content rather than when narrow nuanced dimensions are used. Additionally, it is important for future research to clarify the practical role of CFA models that contain broad dimensions, broad exercises, and GPF. We have offered some scenarios where these models may be useful in applied practice in the section Implications and Recommendations for Practice above. However, more research needs to be done in this regard.

Furthermore, examining the nomological network of exercise and dimension factors remains an important yet under-researched topic in the AC literature. We believe that G theory and factor analytic approaches can be used

[3] It is important to note that this suggestion is aimed at providing the researcher with analytical flexibility to specify various competing models based on the number of parcels. When multiple indicators or parcels are available to operationalize latent dimensions and exercises, the researcher can aggregate or disaggregate the specified indicators/parcels. However, when AEDRs and PEDRs are used, by default, there is no opportunity for the disaggregation of indicators or parcels. Therefore, this suggestion of 'more is better' instead of less, when it comes to the specification of items, should be seen as a practical rather than an empirical recommendation.

collectively to further explore the nomological network of AC dimensions and exercises. More work is needed in this regard, especially when considering the recent publications that identified large numbers of variance components in AC ratings (cf. Jackson et al., 2016). The position of these sources of variance in a nomological network is likely to determine their relative importance. It may be that the sources of variance could be reduced to smaller more manageable units. This type of theoretical unification has been hugely valuable in various fields of psychology including personality and values research (McCrae & Costa, 1999; Schwartz, 2012).

Fourth, to the best of our knowledge, no other studies have investigated parceling allocation strategies in combination with a parceling approach. Therefore, further research on parceling allocation strategies, specifically when there are a larger number of exercises included in the AC, may be warranted. Fifth, additional statistical considerations that were beyond the scope of the current study and that could have affected the results were not investigated. For example, some scholars have suggested hierarchical CFA as an alternative technique for AC factorial validity research (Lievens, 2009) or G theory (Jackson et al., 2016), and this line of enquiry warrants further attention. Specifically, G theory is recommended as a practical solution to dealing with model complexity and model misspecification (Jackson et al., 2016).

Sixth, this study made use of PEDRs (i.e., one parcel) and aggregates of behavioral indicators per Dimension × Exercise to create parcels. Some scholars have argued that AEDRs are more suited to research investigating the internal structure of AC ratings than PEDRs (Arthur et al., 2008; Kuncel & Sackett, 2014; Meriac et al., 2014; Rupp et al., 2008). Although some research shows that dimensions may not explain large amounts of variance independent of the level of aggregation (cf. Jackson et al., 2016; Wirz et al., 2020), further research could still yield important insights. Seventh, given our findings of strong exercise effects, we recommend that further research should investigate whether and how we can better understand exercises and their systematic differences and how we could use exercises to gain insights into participants' intraindividual variability or to better predict performance in organizations. One way to do so might be the recent trend of Multiple Speed Assessments (see Herde & Lievens, 2020, for an overview) that examine participants' behavior and performance across many different exercises. Finally, this study did not examine criterion-related validity. Scholars have suggested that this interaction is important in AC validation research (Lievens, 2009), and therefore, future research in this area should include a simultaneous investigation of the construct and criterion-related validity of AC ratings.

Conclusion

Decades-long research in the field of ACs has brought about a greater understanding of the internal structure of AC ratings. Various statistical techniques have slowly brought the field closer to finding a solution for dealing with complex models, such as ACs. A technique that has been recommended as a practical means to overcome some of the historical challenges found with CFA models has been to use a parceling approach prior to conducting CFA of AC ratings. In this study, a parceling approach in combination with an allocation strategy to create parcels was investigated. Whether an allocation strategy was used did not fundamentally change the conclusions already arrived at in relation to existing research on nonparceling approaches. Nonetheless, we advocate the use of multiple behavioral indicators to measure AC dimensions since it gives developers the flexibility to evaluate various aggregation strategies. Furthermore, an important finding of the study is that a GPF forms an important, yet currently overlooked, component of AC design and should be included in all configurations of the internal structure of AC ratings.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1866-5888/a000266

ESM 1. Detailed model–data fit indices for the CFA models and summaries of model parameters

References

Alkharusi, H. A. (2012). Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education*, *2*(1), 184–196. https://doi.org/10.5296/jse.v2i1.1227

Arthur, W. Jr. (2012). Dimension-based assessment centers: Theoretical perspectives. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 95–120). Routledge.

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *56*(1), 125–153. https://doi.org/10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternate 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.x

Arthur, W., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A

conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management*, *26*(4), 813–835. https://doi.org/10.1177/014920630002600410

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(1), 78–102. https://doi.org/10.1207/S15328007SEM0901_5

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(2), 211–240. https://doi.org/10.1080/10705510801922340

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

Buckett, A., Becker, J. R., Melchers, K. G., & Roodt, G. (2020). How different indicator-dimension ratios in assessment center ratings affect evidence for dimension factors. *Frontiers in Psychology*, *11*(4), 459. https://doi.org/10.3389/fpsyg.2020.00459

Cahoon, M. V., Bowler, M. C., & Bowler, J. L. (2012). A reevaluation of assessment center construct-related validity. *International Journal of Business and Management*, *7*(9), 3–19. https://doi.org/10.5539/ijbm.v7n9p3

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

Coffman, D. L., & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research*, *40*(2), 235–259. https://doi.org/10.1207/s15327906mbr4002_4

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*(1), 82–92. https://doi.org/10.1016/j.jrp.2014.10.009

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C. III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, *72*(3), 493–511. https://doi.org/10.1037/0021-9010.72.3.493

Gibbons, A. M., & Rupp, D. E. (2009). Dimension consistency as an individual difference: A new (old) perspective on the assessment center construct validity debate. *Journal of Management*, *35*(5), 1154–1180. https://doi.org/10.1177/0149206308328504

Gorman, C. A., & Rentsch, J. R. (2017). Retention of assessment center rater training. *Journal of Personnel Psychology*, *16*(1), 1–11. https://doi.org/10.1027/1866-5888/a000167

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Mulitvariate data analysis: A global perspective*. Pearson.

Herde, C. N., & Lievens, F. (2020). Multiple speed assessments. *European Journal of Psychological Assessment*, *36*(2), 237–249. https://doi.org/10.1027/1015-5759/a000512

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection*, *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 assessment center performance. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 281–306). Routledge.

Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). 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

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

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

International Taskforce on Assessment Center Guidelines (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

Jackson, D. J., Michaelides, G., Dewberry, C., & Kim, Y.-J. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Pyschology*, *101*(7), 976–994. https://doi.org/10.1037/apl0000102

Krause, D. E., Rossberger, R. J., Dowdeswell, K., Venter, N., & Joubert, T. (2011). Assessment center practices in South Africa. *International Journal of Selection and Assessment*, *19*(3), 262–275. https://doi.org/10.1111/j.1468-2389.2011.00555.x

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). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology*, *1*(1), 84–97. https://doi.org/10.1111/j.1754-9434.2007.00017.x

Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance*, *20*(4), 345–362. https://doi.org/10.1080/08959280701522031

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology*, *89*(2), 377–385, https://doi.org/10.1037/0021-9010.89.2.377

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance*, *13*(4), 323–353. https://doi.org/10.1207 /S15327043HUP1304

Lance, C. E., Woehr, D. J., & Meade, A. W. (2007). Case study. *Organizational Research Methods*, *10*(3), 430–448. https://doi.org/10.1177/1094428106289395

Leech, N., & A. Haug, C. (2015). Investigating graduate level research and statistics courses in schools of education. *International Journal of Doctoral Studies*, *10*, 93–110. https://ijds.org/Volume10/IJDSv10p093-110Leech0658.pdf

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

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285–300. https://doi.org/10.1037/a0033266

Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures*, *2*(4), 260–293. https://doi.org/10.1080/19312450802458935

McCrae, R. R., & Costa, P. T. Jr. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). Guilford.

Meriac, J. P., Hoffman, B. J., & Woehr, D. J. (2014). A conceptual and empirical review of the structure of assessment center dimensions. *Journal of Management*, *40*(5), 1269–1296. https://doi.org/10.1177/0149206314522299

Merkulova, N., Melchers, K. G., Kleinmann, M., Annen, H., & Szvircsev Tresch, T. (2016). A test of the generalizability of a recently suggested conceptual model for assessment center ratings. *Human Performance*, *29*(3), 226–250. https://doi.org/10.1080/08959285.2016.1160093

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. https://doi.org/10.1037/0033-295X.102.2.246

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. (1998–2017). *Mplus users' guide* (7th ed.). Muthén and Muthén.

Nasser-Abu Alhija, F., & Wisenbaker, J. (2006). A Monte Carlo study investigating the impact of item parceling strategies on parameter estimates and their standard errors in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 204–228. https://doi.org/10.1207/s15328007sem1302_3

Orcan, F. (2013). *Use of item parceling in structual equation modeling with missing data* (doctoral dissertation). Florida State University Libraries. https://diginole.lib.fsu.edu/islandora/object/fsu%3A185150

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology*, *98*(1), 114–133. https://doi.org/10.1037/a0030887

Rupp, D. E., Thornton, G. C. III, & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology*, *1*(1), 116–120. https://doi.org/10.1111/j.1754-9434.2007.00021.x

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

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, *25*, 1–65. https://doi.org/10.1016/S0065-2601(08)60281-6

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, *2*(1), 1–20. https://doi.org/10.9707/2307-0919.1116

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

Siminovsky, A. B., Hoffman, B. J., & Lance, C. E. (2015, April). *Revised estimates of general performance effects on AC ratings.* Paper presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology (SIOP), Philadelphia, PA, USA.

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, *34*(4), 397–423. https://doi.org/10.1006/jrpe.2000.2292

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

Thornton, G. C. III, Rupp, D. E., & Hoffman, B. J. (2015). *Assessment center perspectives for talent management strategies* (2nd ed.). Routledge.

Williams, L. J., & O'Boyle, E. H. (2008). Measurement models for linking latent variables and indicators: A review of human resource management research using parcels. *Human Resource Management Review*, *18*(4), 233–242. https://doi.org/1016/j.hrmr.2008.07.002

Wirz, A., Melchers, K. G., Kleinmann, M., Lievens, F., Annen, H., Blum, U., & Ingold, P. V. (2020). Do overall dimension ratings from assessment centres show external construct-related validity? *European Journal of Work and Organizational Psychology*, *29*(3), 405–420. https://doi.org/10.1080/1359432X.2020.1714593

Woehr, D. J., & Arthur, W. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management*, *29*(2), 231–258. https://doi.org/10.1177/014920630302900206

Woehr, D. J., Meriac, J. P., & Bowler, M. C. (2012). Methods and data analysis for assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 45–67). Routledge.

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multitrait-multimethod data. *Organizational Research Methods*, *15*(1), 134–161. https://doi.org/10.1177/1094428111408616

History
Received October 21, 2019
Revision received September 15, 2020
Accepted September 16, 2020
Published online January 27, 2021

Acknowledgments

This manuscript is part of the first author's doctorate, which is in an article-based thesis format. The degree was conferred by the University of Johannesburg on May 15, 2019.

Special thanks to a private energy producing organization that wishes to remain anonymous and to the Department of Rural Development and Land Reform in South Africa for the data sets.

ORCID
Anne Buckett
https://orcid.org/0000-0002-0853-3878

Anne Buckett
Precision ACS
PO Box 67815
Highveld, 0169
South Africa
anne@precisionacs.co.za

