Everything That You Have Ever Been Told about Assessment Center Ratings is Confounded

Duncan J. R. Jackson
Birkbeck, University of London and University of Johannesburg

George Michaelides and Chris Dewberry
Birkbeck, University of London

Young-Jae Kim
ASSESTA Co Ltd

Author Note

Duncan J. R. Jackson, Department of Organizational Psychology, Birkbeck, University of London and Faculty of Management, University of Johannesburg; George Michaelides and Chris Dewberry, Department of Organizational Psychology, Birkbeck, University of London; Young-Jae Kim, ASSESTA Co Ltd.

The authors would like to thank Kerr Inkson, Charles E. Lance, Jiyoung Seo, Myungjoon Kim, and Nigel Guenole for their contributions to this article.

Correspondence concerning this article should be addressed to Duncan J. R. Jackson, Department of Organizational Psychology, Birkbeck, University of London, Clore Management Centre, Torrington Square, London, WC1E 7JL.
E-mail: dj.jackson@bbk.ac.uk

Abstract

Despite a substantial research literature on the influence of dimensions and exercises in assessment centers (ACs), the relative impact of these two sources of variance continues to raise uncertainties because of confounding. With confounded effects, it is not possible to establish the degree to which any one effect, including those related to exercises and dimensions, influences AC ratings. In the current study (N = 698) we used Bayesian generalizability theory to unconfound all of the possible effects contributing to variance in AC ratings. Our results show that ≤ 1.11% of the variance in AC ratings was directly attributable to behavioral dimensions, suggesting that dimension-related effects have no practical impact on the reliability of ACs. Even when taking aggregation level into consideration, effects related to general performance and exercises accounted for almost all of the reliable variance in AC ratings. The implications of these findings for recent dimension- and exercise-based perspectives on ACs are discussed.

Keywords: generalizability theory, Bayesian analysis, assessment centers

Everything That You Have Ever Been Told about Assessment Center Ratings is Confounded

When, in the context of selection, appraisal, and development, behavioral criteria are used to evaluate individuals, it is essential that these criteria are measured reliably. Unsurprisingly, therefore, the measurement properties of assessment center (AC) ratings have come under close scrutiny in the applied psychology literature. In ACs, the behavior of job-holders or candidates is sampled across several work-related situations (exercises, e.g., a role play exercise, group discussion, presentation) and is typically assessed by trained assessors in terms of pre-defined behavioral dimensions (e.g., communication skills, teamwork, planning and organizing). As a result of their multifaceted measurement properties, incorporating dimensions, exercises, and assessors, ACs provide a rich source of information about the extent to which work-related behavioral criteria can be reliably measured in a job-relevant setting.

Historically, researchers have questioned the extent to which behavioral dimensions are measured reliably in ACs, and have implied that researchers should utilize an exercise-oriented approach to scoring ACs (Jackson, Stillman, & Atkins, 2005; Lance, 2008, 2012; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Sackett & Dreher, 1982; Sakoda, 1952; Turnage & Muchinsky, 1982). A few recent studies tend to support this view (B. J. Hoffman, Kennedy, LoPilato, Monahan, & Lance, 2015; Jansen et al., 2013; Speer, Christiansen, Goffin, & Goff, 2014). However, other recent research suggests that concerns surrounding behavioral dimensions are misplaced and that dimensions can, in fact, be measured reliably (Guenole, Chernyshenko, Stark, Cockerill, & Drasgow, 2013; Kuncel & Sackett, 2014; Meriac, Hoffman, & Woehr, 2014; Meriac, Hoffman, Woehr, & Fleisher, 2008; Putka & Hoffman, 2013). In the present article, we suggest that AC research, in general, has confounded both exercise- and

dimension-related effects with a multitude of other sources of variance. Such confounds threaten the interpretability of the true sources of reliable variance in AC ratings.

The main substantive contributions to theory of the present study are (a) to bring clarity to the interpretation of reliability in AC ratings by providing an unconfounded perspective based on the 29 possible sources of variance in AC ratings and (b) to use this unconfounded perspective to help reconcile dimension-based, exercise-based, and mixed theoretical perspectives on ACs. The main contribution of the study to practice is to provide an uncluttered perspective on the possible bases for reliability in AC ratings – a perspective that may offer guidance to applied psychologists in terms of the sources of measurement reliability in ACs and the design elements that may maximize AC reliability. In addition, recognizing that Bayesian approaches are well suited to complex models such as AC measurement models, and responding to calls in the organizational literature for Bayesian methods (Kruschke, Aguinis, & Joo, 2012; Zyphur, Oswald, & Rupp, 2015) and for applications of Bayesian generalizability theory (LoPilato, Carter, & Wang, 2015), we seek to contribute to the methodology used for the analysis of AC data and data from multifaceted measures generally.

The Debate about AC Ratings

ACs involve assessments of behavior in relation to work-relevant situations (Guenole et al., 2013; International Taskforce on Assessment Center Guidelines, 2015; Walter, Cole, van der Vegt, Rubin, & Bommer, 2012). They are configured in such a way that evidence for a given dimension is observed across multiple, different exercises, each of which simulates a work-related situation (Thornton & Mueller-Hanson, 2004). It is implied here that, in order for a dimension to represent a meaningful, homogenous behavioral category, observations relating to

the same dimension should agree, at least to some extent, across different exercises (Arthur, 2012; Kuncel & Sackett, 2014; Lance et al., 2004; Meriac et al., 2014).

The AC literature reflects confusion about how to create meaningful dimension categories (Howard, 1997), with views on this topic swinging between extremes. Although not the first to identify the issue (see Sakoda, 1952), Sackett and Dreher (1982) factor analyzed AC ratings and found that the resulting summary factors consistently reflected exercises and not dimensions; a finding that is referred to in the extant literature as the *exercise effect*. Sackett and Dreher's findings have since been repeated across different organizations, organizational levels, and diverse nations, including the USA, the United Kingdom, China, Singapore, Australia, and New Zealand (Lievens & Christiansen, 2012).

The lack of correspondence among same dimension observations across different AC exercises has often been perceived as a problem (e.g., see Lance, 2008). In the AC literature, attempts have been made to maximize the extent to which observations of the same dimension will correspond across different exercises. Some of these efforts have led to innovative intervention strategies. Examples include rating a single dimension across exercises (Robie, Osburn, Morris, Etchegaray, & Adams, 2000), reducing cognitive load through the application of behavioral checklists (Reilly, Henry, & Smither, 1990), rating dimensions after the completion of all exercises (referred to as post-consensus dimension ratings, see Silverman, Dalessio, Woods, & Johnson, 1986), and the use of video recordings (Ryan et al., 1995). Despite such efforts, the exercise effect has continued to manifest itself in operational ACs (Lance, 2008; Lievens & Christiansen, 2012; Sackett & Lievens, 2008).

Exercise- versus Dimension-Centric Perspectives

The current international guidelines on ACs permit alternative, exercise-based scoring approaches (International Taskforce on Assessment Center Guidelines, 2015). It therefore appears that AC designs that incorporate alternative, exercise-oriented scoring procedures are acceptable. But this does not mean that there has been a notable rise in research on exercise-specific scoring in ACs. In fact, there have been only a few recent studies focused on the role of exercise-related sources of variance: in contrast, there have been more dimension-based studies.

The few, relatively recent, exercise-oriented studies have provided insights into the nature of exercise-related effects. Speer et al. (2014) found that ACs employing exercises with differing behavioral demands tended to return higher criterion-related validity estimates than those with similar demands. This finding suggests that, notwithstanding the conventional aim of achieving concordance among dimension ratings observed across different exercises, cross-exercise variability is favorable to valid AC practice. Also, making reference to classic notions of *behavioral consistency*, Jansen et al. (2013) found that people who were better able to comprehend situational demands tended to score higher on behavioral measures used in selection (interviews and ACs) and on job performance ratings. Thus, it was suggested that individual differences with respect to situational appraisal explain the link between situationally-based assessment and outcome performance ratings. Moreover, B. J. Hoffman et al. (2015) found small-to-moderate criterion-related validities associated with individual AC exercises.

Generally, the findings from exercise-centric studies suggest that situational elements are important in ACs. However, none of these studies attempted to isolate exercise-related effects from other effects inherent in the AC process. Thus, while the findings above hint at the importance of situational influences, this conclusion can only be reached after other, potentially

confounded, effects (e.g., assessor-related effects, dimension-related effects) have been taken into consideration.

In contrast to the few, recent exercise-centric AC studies, several studies have reported findings suggesting that dimensions are meaningfully correlated with both work outcomes and with externally-measured constructs. Meriac et al. (2008) investigated the meta-analytic relationship between summative dimension scores over and above those associated with personality and cognitive ability and job performance. They found a multiple *R* of .40 between dimensions and job performance (which was close to previous estimates, see Arthur, Day, McNelly, & Edens, 2003) and that dimensions, over and above cognitive ability and personality, explained around 10% of the variance in job performance. In a different study, Meriac et al. (2014) meta-analyzed the structure of post-consensus dimension ratings. They found that a three-factor model based on dimensions (comprising administrative skills, relational skills, and drive) correlated with general mental ability and personality.

Other studies have looked at measurement characteristics internal to ACs that could help to reconcile earlier dimension-related criticisms. Kuncel and Sackett (2014) asserted that "the construct validity problem in assessment centers never existed" (p. 38) and titled their study "Resolving the assessment center construct validity problem (as we know it)" (p. 38). Part of their reasoning was that previous research had failed to acknowledge the effects of aggregation on AC ratings and, as a result, had misrepresented the magnitude of exercise-based relative to dimension-based variance in ACs. Also, using confirmatory factor analysis (CFA), Guenole et al. (2013) found that more variance in AC ratings was explained by dimensions (mean factor loading = .42) than by exercises (mean factor loading = .32) and concluded that their findings were due to up-to-date design approaches that had improved the measurement of dimensions.

Monahan, Hoffman, Lance, Jackson, and Foster (2013) addressed solution admissibility problems that often arise when dimensions are included in CFA models. They concluded that a sufficient number of dimension indicators improved the likelihood of solution admissibility and they also found evidence for moderate dimension effects.

Putka and Hoffman (2013) also looked at measurement characteristics internal to ACs, and covered a broad range of methodological factors that could have affected the expression of dimension- versus exercise-based variance. Like Kuncel and Sackett (2014), Putka and Hoffman acknowledged aggregation level, meaning that they recognized changes that might occur in variance decomposition as a result of aggregation (e.g., aggregating across exercises to arrive at dimension scores). They also criticized previous research (e.g., Arthur, Woehr, & Maldegan, 2000; Bowler & Woehr, 2009) for not having specified assessor effects appropriately (i.e., for not using unique identifiers for each assessor) and for confounding reliable (i.e., true score) with unreliable (i.e., error) sources of variance when estimating effects in ACs. Putka and Hoffman found two sources of reliable variance[1] in ACs relevant to dimensions. Firstly, a two-way person × dimension interaction, which only explained 1.1% of the variance in AC ratings and, secondly, a three-way person × dimension × exercise interaction, which explained a substantial 23.4% of variance[2]. Putka and Hoffman interpreted this three-way interaction as being "consistent with interactionist perspectives on dimensional performance and trait-related behavior" and that, as a result of its magnitude, researchers should not "discount the importance of dimensions to AC

[1] We adopt terminology from Putka and Hoffman (2013) here, where *reliable variance* is analogous to true score variance in classical test theory and *unreliable variance* is analogous to error. See Putka and Hoffman (2014) and Putka and Sackett (2010) for further clarification of this terminology as it relates to generalizability theory.

[2] We present non-aggregated results in this section so as to allow comparisons with previous results on ACs. Dimension-based variance is most likely to be expressed at the dimension-level of aggregation, at which the person × dimension interaction = 2.1% of variance in AC ratings and the person × exercise × dimension interaction = 15.2% of variance (in Putka & Hoffman, 2013).

functioning" (p. 127), i.e., as evidence in favor of the contribution of dimensions to reliable AC variance.

Do Confounds Limit the Interpretability of Research on AC Ratings?

ACs are multifaceted measures. This means that any aggregate score (e.g., based on dimensions) that is derived from an AC will reflect the constituent effects (i.e., sample dependencies, the general performance of the participants, dimensions, exercises, assessors, indicator items, and respective interaction terms) making up that score. If, when researchers attempt to model variance in AC ratings, any of these effects are not acknowledged, then that model is ultimately misspecified and, as a result, confounded (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963). Depending on the extent of the confounding, its presence generally renders hazardous the interpretation of results (Herold & Fields, 2004).

There are many possible reasons for model misspecification and thus confounding: these include omission, data unavailability, and design factors that prevent the isolation of particular effects (Brennan, 2001; Putka & Hoffman, 2013). For example, if the assessor-to-participant ratio in an AC is not >1:1, then it is impossible to isolate assessor-related effects. Another possible reason for misspecification is related to computer memory limitations, such that modeling every possible effect might be computationally impractical, at least with current technology and using traditional statistical estimation techniques (Searle, Casella, & McCulloch, 2006).

To varying degrees, all of the AC studies mentioned above suffer from confounds that potentially place limitations on the interpretability of their results. To illustrate, Speer et al. (2014) correlated summative exercise (and overall) ratings with job performance. Such

summative scores confound exercise-related variance with a multitude of other sources of variance (e.g., dimensions, assessors, items). There is therefore no definite way of knowing whether the effects observed by Speer et al. can be directly attributed to exercises. Likewise, Meriac et al. (2008) created summative dimension scores, thereby similarly introducing confounds. The correlations that Meriac et al. (2008) reported between dimensions and outcomes could therefore have resulted from any of the multiple variance sources that went into their summative "dimension" predictor scores (e.g. exercise effects, assessor effects, etc.). There is no way of knowing that the primary factor here was actually related to dimensions. Also, in a different study, Meriac et al. (2014) used post-consensus dimension ratings, which excluded potentially relevant, exercise-based sources of variance. Moreover, Meriac et al. (2014), like Speer et al., Guenole et al. (2013), Monahan et al. (2013), and Kuncel and Sackett (2014) did not model assessor-related variance. As a specific example, Kuncel and Sackett confounded participant × dimension effects with participant × dimension × assessor effects[3]. Thus, potentially *un*reliable sources of assessor-based variance in these studies were confounded with potentially *reliable* sources of variance. In such cases, effects held to be reliable might, in fact, be tainted with unreliability, so that confounding might potentially lead to misinterpretation.

In contrast, Putka and Hoffman (2013) made the best known effort so far to provide clarity on sources of variance in AC ratings: they isolated 15 separate AC-related effects. However, even this study introduced potentially nontrivial confounds in that the authors did not model item indicators (which have been found to influence variance in AC ratings, see Monahan et al., 2013) or sample effects (which have long represented an important consideration in organizational research, see Laczo, Sackett, Bobko, & Cortina, 2005). Had Putka and Hoffman also modeled indicator and sample effects, the number of distinct variance sources available to

[3] This example is also noted in Putka and Hoffman (2013).

them would have increased from 15, the number that they examined, to a total of 29 separate effects (see Table 2). Thus, whilst Putka and Hoffman's study was less affected by confounding than previous studies were, the authors only estimated around half of the total number of effects that it is possible to decompose from AC ratings.

One possibility is that the Putka and Hoffman (2013) study was limited by traditional approaches to variance estimation. In what has long been described in the education and statistics literature as the "Achilles heel" of generalizability theory (e.g., Brennan, 2001, p. 210; Shavelson, Webb, & Rowley, 1989, p. 927), traditional approaches to variance estimation (such as those based on analysis of variance, ANOVA) can result in problematic estimates (e.g., negative variances). Putka and Hoffman used restricted maximum likelihood (REML) estimators, which represent an advancement over ANOVA-based estimators (Marcoulides, 1990). However, whereas variances should, by definition, always return positive values (Searle et al., 2006), even REML estimators can result in problematic estimates, such as negative variances that are artificially set to (fenced at) zero. Putka and Hoffman's study included a total of four such fenced variances, all of which were associated with potentially important sources of unreliable assessor-related variance. Given that the variance estimates used in generalizability theory are interdependent, an additional concern is that non-admissible estimates might affect other estimates in a manner that is difficult or impossible to predict.

Moreover, depending on sample size and model complexity, the computational demands associated with REML estimators can be nontrivial or even insurmountable (Brennan, 2001). This consideration could have placed restrictive limitations on the number of effects that Putka and Hoffman (2013) were able to estimate. However, the literature still requires an unconfounded perspective on the contribution of all of the different effects involved in AC

ratings: if particular effects are left unmodeled, the net result is that confounds are introduced. This creates challenges when attempting to interpret the impact of dimensions, exercises, or other effects in published AC studies.

Based on the need for the isolation of unconfounded sources of variance in AC ratings, our research aim is straightforward. Our intention is to present a perspective on the decomposition of AC variance that minimizes confounding. In order to do so, we make use of recent advances in the literature with respect to the application of Bayesian estimation methods applied to generalizability theory (also, see LoPilato et al., 2015). Bayesian methods overcome the potentially problematic negative or fenced variance components that can occur with REML estimation, particularly when the number of levels specific to particular effects is small (e.g. when only 3 exercises are used in an AC, see Brennan, 2001). Through the use of weakly informative priors, Bayesian methods can even address the issue of estimates being too close to 0 (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013). Although Bayesian analysis typically requires substantial computational resources because of its use of Markov chain Monte Carlo (MCMC) estimation, it offers the advantage of being applicable to very complex models and thus renders practical the estimation of models with a large number of parameters (Gelman, Carlin, Stern, & Rubin, 2013). This offers a considerable advantage over REML estimation, in which model complexity can often lead to computational difficulties that cannot easily be addressed with additional computational power. Moreover, Bayesian generalizability theory allows for estimates of the posterior distribution of each variance component and associated reliability coefficients, thus providing a rich source of information on the distributional characteristics of these estimates (Gelman et al., 2013; Jackman, 2009; LoPilato et al., 2015).

In keeping with our research aims, our sole Research Question is as follows:

Research Question: When confounding is appropriately controlled, which sources (i.e., sources related to samples, participants, items, exercises, dimensions, or assessors) contribute to reliable variance in AC ratings?

In relation to our Research Question, if a dimension-based perspective (e.g., Arthur et al., 2003; Meriac et al., 2014) holds true, then dimension-based sources of variance should prevail in terms of variance explained in AC ratings. Such an outcome would also imply that because most of the variance in ratings would be attributable to dimension-related source of variance anyway, regardless of confounding, any confounding observed in recent dimension-centric research would not present an issue of practical import. Such an outcome would, however, suggest that factors other than those related to exercises were involved in the prediction of outcomes and would therefore present a concern for exercise-centric perspectives (e.g., Speer et al., 2014).

Alternatively, if an exercise-based perspective (e.g., Jackson, 2012; Jansen et al., 2013) holds true, then exercise-based sources of variance should prevail. This outcome would be favorable for the exercise-centric perspective. However, because it would suggest that the "real" reason for the structural or predictive properties of AC scores is attributable to exercise-related and not dimension-related variance, it would represent a concern for dimension-centric research. Moreover, if a combination, dimension-plus-exercise, perspective holds true (referred to in the literature as a mixed perspective, see B. J. Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011), then both dimension- and exercise-related variance sources should contribute substantially to variance in AC ratings. Such an outcome would imply that both exercise- and dimension-related variance sources are important in AC research and practice. There are also other possible

outcomes relating to alternative perspectives (e.g., a proportionately large influence related to items, assessors, and/or different samples), which we address in the present study. Due to the confounding apparent in extant studies of ACs, we urge that none of the above perspectives should be taken for granted.

Method

Sample Information

The sample in this study involved five separate administrations of an operational AC based in South East Asia. The purpose of the AC was to generate data that would be used to guide decisions around internal promotions from line management to senior management positions. Demographic information by sample is presented in Table 1. We analyzed data from all participants simultaneously (698 candidates) whilst formally modeling the fact that participants were nested in different administration subsamples.

AC Design and Development

The AC in the present study involved three exercises (an in-basket, a role play, and a case study, see Appendix, Table A1) that were designed to measure between three and five dimensions (see Appendix, Table A2). As is common in ACs, the configuration here was not fully crossed in that not every dimension was assessed in every exercise (see Appendix, Table A3). In accordance with design suggestions about minimizing cognitive load, the number of dimensions assessed within any given exercise was kept to a minimum (Chan, 1996; Lievens, 1998). Because this was a high-stakes evaluation, dimensions were not revealed to participants prior to the AC taking place.

Assessors were experienced senior managers from the participating organization. A total of 38 assessors participated in the first administration, 47 in the second, 77 in the third, 80 in the

fourth, and 80 in the fifth. The total number of assessors increased over time as the participant organization gained logistical insights. Assessors were matched to participants such that two assessors rated each participant on the in-basket and the role play exercises, and three assessors rated each participant on the case study. All assessors involved in this study were assigned an identification code and each unique assessor-participant combination was recorded for the purpose of analysis. Assessors were randomly assigned to candidates by means of a computerized algorithm.

The exercises in this study were based on job analyses of the focal position and interviews with subject matter experts in accordance with guidance from the AC literature (Howard, 2008; Schippmann, Hughes, & Prien, 1987; Thornton & Krause, 2009; Williams & Crafts, 1997). Specifically, in order to gain a balanced perspective on the position of interest, task lists were developed based on a review of existing documentation and in consultation with experienced senior managers and incumbents. In a workshop scenario, in order to determine tasks that were deemed to be important within the focal position, these task lists were reviewed and refined. The retention of tasks was also based on the practicability of including them in a simulation exercise. To determine task-relevance and to ensure against conceptual redundancy, task lists were reviewed by a team of experienced industrial-organizational (I-O) psychologists. In the final development stages, exercises were either designed to accommodate tasks identified by subject matter experts or, where this was not feasible, tasks were dropped from the procedure.

Application

Based on the task-lists developed at the job analysis phase, rating items were developed for each exercise (14 items for the in-basket, seven items for the role play, and 11 items for the case study) that provided behavioral descriptions for the dimensions of interest in accordance

with suggestions in the extant literature (Donahue, Truxillo, Cornwell, & Gerrity, 1997; Guenole et al., 2013; International Task Force on Assessment Center Guidelines, 2009; Lievens, 1998). Items were constructed such that they could be traced back to task lists and job analysis information. Each item was associated with a dimension, such that each within-exercise dimension observation was based on two or three item ratings. Assessors rated independently and there was no formal process of consensus discussions for each allocated set of ratings relevant to a given participant. Because of this, the ratings provided in this study are best considered as pre-consensus ratings.

Assessors were trained by experienced consultants with postgraduate degrees (Master's degrees or PhDs) in I-O psychology over an intensive 2-day course that covered a range of topics aimed at fostering content familiarization, assessors' awareness of errors, and rater skills. Content familiarization was provided for exercises, rating items, dimensions, and logistical issues. Assessors were also familiarized with potential rater errors (e.g., leniency biases and halo effects); however the majority of the training session was dedicated to rater skills training. In this respect, assessors were initially introduced to processes involved in the observation and recording of behavior and how such observations should be documented and summarized. In turn, assessors rated the performance of a mock candidate in their assigned exercise. In an effort to create a shared frame of reference regarding performance expectations (as guided by Gorman & Rentsch, 2009; Macan et al., 2011), they then discussed the assigned behavioral ratings. In order to help ensure that assessors were given ample practical experience in the assessment process, the mock candidate assessment was repeated three to four times, and ratings were checked for consistency after each practice run.

Data Analysis

In terms of measurement design, participants, dimensions, assessors, and exercises were specified as crossed random factors. In addition, participants were specified as being nested in subsamples and items were specified as being nested in exercises[4]. A grand total of 29 separate effects resulted from this design. Of these 29 effects, based on the literature reviewed previously, five effects were deemed to represent reliable sources of variance[5]. In the interests of brevity, we provide descriptions for only the five reliable variance sources (Appendix, Table A4). In addition, our study included 15 sources of unreliable variance, all of which were linked to assessor-based sources. A further nine "other" effects neither contributed to differences between assessor ratings nor to the rank ordering of participant scores and were therefore irrelevant to a consideration of reliability. All of these effects appear in the analyses that follow (see Table 2)[6].

Because assessors were not fully crossed with participants and dimensions were not fully crossed with exercises (i.e., it was an ill-structured design, see Putka, Le, McCloy, & Diaz, 2008), we dealt with data sparseness by defining our model as a hierarchical one. This enables the analysis of ill-structured designs using random effects models without the need to delete large portions of data.

To address concerns highlighted by Kuncel and Sackett (2014) around aggregation, we rescaled variance components resulting from our random effects models using the approach detailed in Brennan (2001, pp. 101-103) and in Putka and Hoffman (2013). We used these procedures in the Bayesian model so as to obtain full posterior distributions of the coefficients. We rescaled by taking levels of aggregation into account when representing the relative contribution of given effects to variance explained in the model. Aggregation to dimensions

[4] Note that the presence of a colon (:) denotes a level of nesting (e.g., i:e means that items [i] are nested in exercises[e])
[5] In fact the number of effects defined as "reliable" or "unreliable" depends on intended generalizations (see Tables 3 and 4). In the discussion above, we refer to generalization to different assessors only.
[6] We define reliability here in relative, rather than in absolute, terms (see Shavelson & Webb, 1991).

required us to average AC ratings (or post exercise dimension ratings, PEDRs) across exercises. Aggregation to exercises required us to average PEDRs across dimensions. Aggregation to overall scores required us to average PEDRs across both exercises and dimensions. To correct for the ill-structured nature of our measurement design with respect to assessors, we also rescaled any reliability-relevant variance components involving assessor-related effects using the *q*-multiplier approach detailed in Putka et al. (2008). Formulae relating to aggregation and the inclusion of the *q*-multiplier are shown in Tables 3 and 4.

The variance components resulting from our analysis were used to estimate reliability based on the ratio of reliable-to-observed variance for PEDRs and for scores aggregated to the dimension-, exercise-, and overall-level. Table 3 shows the formulae for the reliability of PEDRs and dimension scores for generalization to different assessors[7] and to a combination of different assessors and different exercises. We included only appropriate generalizations in our analyses[8]. In generalizability theory, when aiming to "generalize across" conditions of measurement, such conditions are treated as sources of *un*reliable variance (Brennan, 2001; Cronbach et al., 1972; Putka & Hoffman, 2014). In Table 3, for example, when scores aggregated to the dimension-level are generalized across both assessors and exercises, both assessor- and exercise-related sources of variance are specified as contributing to unreliable variance. Table 4 shows the formulae for the reliability of exercise scores and overall scores.

Bayesian Analysis and Model Specification

[7] The dangling modifiers "generalization to" and "generalizing to" are routinely applied in generalizability theory and are used to describe researcher intentions relating to measurements (e.g., "generalizing to different assessors" means that the researcher intends to generalize AC ratings to different assessor groups).

[8] Possible generalizations depend on how scores are aggregated (e.g., if the aim is to aggregate to exercise scores, then it is not possible to estimate generalization to different exercises). Also, and although such generalizations are possible, unlike Putka and Hoffman (2013), we did not attempt to generalize to different exercises at the PEDR, dimension, or overall levels of aggregation because doing so requires that assessor-related variance is considered as contributing to *reliable* variance. For many applied purposes, and particularly for ill-structured assessor configurations, such a representation would be considered as inappropriate.

For the analysis, we used R 3.2.2 (R Core Team, 2014), Stan 2.8.0 (Stan Development Team, 2015b), and Rstan 2.8.0 (Stan Development Team, 2015a). Stan is a probabilistic programming language for Bayesian analysis with MCMC estimation using Hamiltonian Monte Carlo (HMC) sampling. Specifically, Stan uses the No-U-Turn Sampler (M. D. Hoffman & Gelman, 2014) for automatic tuning of the Hamiltonian Monte Carlo sampling approach.

The model was defined as a hierarchical model with 28 random intercepts and one fixed intercept. For the fixed intercept we used a normal prior with a 0 mean and a standard deviation of 5. Considering the scale of our data, this is a fairly broad, weakly-informative prior which would easily converge towards the grand mean value of the data. The model was reparameterized using a non-centered parameterization (Papaspiliopoulos, Roberts, & Sköld, 2007). This parameterization requires that the random intercepts are sampled from unit normal distributions and then rescaled by multiplying them by the group-level standard deviation associated with each of the 28 random intercepts. The prior distributions for the standard deviation for each of the 28 group-level variance components as well as the residual was the half-Cauchy distribution, which is the recommended weakly-informative prior for variance components (Gelman, 2006). The scale of the half-Cauchy distributions of the 29 error terms (28 variance components and the residual term) was a hyper-prior sampled from a uniform distribution ranging from 0 to 5.

The analysis was conducted with four simulation chains using random starting values and 10,000 iterations. The first 5,000 iterations were essentially "warm-up" iterations and the remaining 5,000 were used for sampling. Iterations were thinned by a factor of 10, thus using one in every 10 samples. Although this may be considered a small number of iterations in the context of other sampling approaches (such as Gibbs sampling), HMC offers the advantage that

its samples are not so susceptible to autocorrelation and therefore a smaller number of iterations are often sufficient to properly explore the posterior distribution space. Convergence was evaluated using trace plots, density plots, and autocorrelation plots, which revealed acceptable mixing without any concerns about autocorrelation. Similarly, when we evaluated the potential scale reduction factor (Gelman & Rubin, 1992), all of the parameters were below the recommended $\hat{R}$< 1.05, which indicates convergence of the four chains and acceptable mixing. Effective sample size estimates and Monte Carlo standard errors indicated that the number of iterations used was sufficient.

Results

Table 2 shows all of the 29 effects that were estimated in this study, classified into reliable, unreliable, and reliability-unrelated groupings. Variance estimates are expressed in Table 2 as percentages of total variance explained in AC ratings, taking all 29 effects into account. Percentages of variance are also displayed for the effects that are relevant to between-participant comparisons (i.e., relevant to reliability). These between-participant percentages of variance are, in turn, presented with reference to levels of aggregation relating to dimension, exercise, and overall scores.

Our focus, from this point, is on percentages of variance indicating reliable between-participant effects. From Table 2, it is immediately clear that the effect pertaining directly to dimensions (p:sd) explained a very small percentage of variance in AC ratings, (ranging from 0.13% to 1.11 %), irrespective of the level of aggregation involved. Also, irrespective of level of aggregation, the analogues of general performance (p:s, ranging from 25.91% to 64.79% of variance) and exercise effects (p:se, ranging from 24.71% to 53.66%) explained the vast majority

of the reliable variance. In other words, general performance explained at least 23 times and exercise effects explained at least 22 times more variance than dimension effects.

The remaining dimension-related source of reliable variance was a three-way interaction involving participants, dimensions, and exercises (p:sde). The magnitude of this effect was highly dependent on aggregation level. At the non-aggregated level, p:sde explained around 12.75% of variance in PEDRs, which represented its strongest contribution. At the overall-level, p:sde explained only 1.77% of variance, which represented its weakest contribution. This variability is due to the fact that p:sde involves both dimension- and exercise-related effects, so that when aggregation takes place across both dimensions and exercises, its impact diminishes dramatically.

Tables 3 and 4 show reliability estimates for PEDRs and for dimension, exercise, and overall scores. Reliability estimates are provided (a) for generalization to different assessors and assessors/exercises for PEDRs, dimensions, and overall scores and (b) for generalization to different assessors and assessors/dimensions for exercise scores. From Tables 3 and 4 it is clear that, regardless of aggregation level, when assessor-related variance is considered as contributing to unreliable variance, reliability is high (with estimates $\geq .80$). However, when exercise-related sources of variance are considered unreliable at the PEDR, dimension, and overall-score levels, reliability is low (with estimates dropping to $\leq .65$). This suggests that exercise-based sources of variance should always be considered as contributing to reliable variance. Moreover, the results for exercise scores in Table 4 suggest that treating dimension-related sources as contributing to unreliable variance makes very little difference (only .04) to reliability outcomes.

Table 2 also shows credible intervals for variance estimates. In Bayesian analysis, parameter estimates are considered to be random (i.e. varying) and each possible value is

associated with a probability. The 95% credible interval represents the interval of the 95% most probable values that the parameter can take (Gelman et al., 2013). Figure 1 shows plots of credible intervals for each parameter estimate and, as can be seen, uncertainty is more prevalent when within-group level frequencies are low (i.e., the number of sub-samples and the number of exercises). However, even when taking the uncertainty level into consideration, the p:s and p:se effects remain higher than any other effect. Figure 2 shows credible intervals for reliability estimates based on posterior distributions. It is evident that regardless of desired generalization, uncertainty presents less of an issue for reliability based on exercise scores and more of an issue for reliability based on dimension and overall scores.

To frame our findings in relation to those of previous studies, Table 5 shows comparable (between-participant only) effects derived from previous studies that also applied random effects models. Table 5 shows that the number of *between-participant* effects being estimated in ACs and that precision has thus increased over the years (Arthur et al., 2000, five effects; Bowler & Woehr, 2009, eight effects; Putka & Hoffman, 2013, 12 effects, and the present study, 20 effects)[9]. It is also clear that, as precision-level has increased, the percentage of variance associated with the analogue of dimension effects (p:sd) has decreased, whereas, and except in Arthur et al. (2000), the percentage of variance associated with the analogue of exercise effects (p:se) has remained consistently high.

Our aim, in a similar vein to Putka and Hoffman (2013), was to question previous studies that had confounded effect estimates derived from PEDRs. However there are nontrivial and substantive differences between our findings and those of Putka and Hoffman, which could be attributed to the Putka and Hoffman study's confounding of sample- and item-related effects.

[9] Note that nine of the effects estimated in the present study did not involve participant- or assessor-related interactions and, therefore, are not regarded as between-participant effects.

Notably, our general performance (p:s) analogue was almost 9% larger than that estimated by Putka and Hoffman. Also our three-way interaction involving participants, dimensions, and exercises (p:sde) was almost 11% smaller than their analogue. When our findings are compared with those of Putka and Hoffman, these differences affect the rank ordering, by magnitude, of the modeled effects.

In addition, we found an extra, albeit small (4.75%), contributor to reliable variance in an effect involving participants and exercise-nested items (p:si:e). Early exercise-based perspectives on ACs (Goodge, 1988; Lowry, 1997) suggest that developmental feedback can be provided to AC participants on the basis of exercise-nested behavioral descriptors, which is akin to what the p:si:e effect represents. This effect is relevant at the non-aggregated PEDR level because it is at the PEDR level that feedback based on behavioral descriptors will be applied. This means that, despite being small in absolute terms, p:si:e still contributed almost nine times the reliable variance of the comparable dimension-related effect (p:sd) in our study.

Discussion

After over 60 years (see Sakoda, 1952), the literature on ACs still sways between a focus on dimension- and a focus on exercise-related sources as the major contributors to reliable variance in AC ratings. Scrutiny of this literature reveals confounding, which raises challenges to ascertaining which factors are associated with reliable AC variance. Because ACs are multifaceted measures incorporating numerous different effects that could potentially influence ratings, confounding is a threat to the interpretation of findings from studies involving AC data. Capitalizing on the advantages of Bayesian generalizability theory, ours is the first known study to decompose, and thus unconfound, all of the 29 sources of variance that could potentially contribute to variance in AC ratings. Of these 29 variance sources, two effects are relevant to

reliable dimension-based variance, three effects are relevant to reliable exercise-based variance, and one effect is akin to a general performance effect. Our proposition was that if dimension-based sources of variance contributed the majority of reliable variance in AC ratings, then the dimension perspective (e.g., Kuncel & Sackett, 2014; Meriac et al., 2014) would prevail. If, however, exercise-based sources of variance explained most of the reliable variance, then the exercise perspective would prevail (e.g., Jansen et al., 2013; Speer et al., 2014). If both dimension and exercises sources contributed meaningfully to reliable AC variance, then the mixed approach would prevail (e.g., B. J. Hoffman et al., 2011).

Dimension-Related Sources of Reliable Variance

The two reliable dimension-related effects that we decomposed comprise (a) the p:sd effect, which is analogous to the dimension effects that are typically estimated using CFA; and (b) the p:sde effect, for which there is no CFA analogue, but which essentially represents a three-way interaction involving participants, dimensions, and exercises (see Table 2). With the p:sd effect (and taking aggregation into consideration), the proportion of between-participant variance explained ranged between 0.13% and 1.11%, and was therefore too small to warrant further consideration.

The three-way p:sde effect, however, explained potentially more between-participant variance in AC ratings but ranged widely between 1.77% and 12.75%, depending on how ratings were aggregated (see Table 2). Putka and Hoffman (2013) estimated an analogue of this effect and found that it explained much more variance than was the case in our study (up to 23.4%). They stated that their findings were "consistent with interactionist perspectives on dimensional performance and trait-related behavior" and that, accordingly, researchers should not "discount the importance of dimensions to AC functioning" (p. 127). Putka and Hoffman's

findings, although less confounded than preceding studies, were, nonetheless, still confounded in that neither item- nor sample-related effects were modeled in their study. Our (unconfounded) results suggest that, at its strongest (i.e., at the non-aggregated, between-participant level), the p:sde effect was almost half the magnitude of the analogous estimate reported in Putka and Hoffman (see Table 5).

We also propose an interpretation of the p:sde effect which differs from the interpretation of Putka and Hoffman (2013). A three-way, p:sde interaction implies that participant-relevant dimension effects are dependent on exercises. Whilst we agree with Putka and Hoffman that this is consistent with interactionist perspectives, it is equally aligned with the view that any dimension effects of note in ACs are likely to be situation-specific. That is, the p:sde effect suggests that the ratings assigned to AC participants on the basis of dimensions depend on the exercises in which they are taking part. This ultimately implies that dimension scores in ACs cannot be meaningfully interpreted as reflecting cross-exercise consistent dimension-related behavior. Instead, at best, there is likely to be a contribution based on dimension-related behavior which is specific to particular exercises.

The Major Sources of Reliable AC Variance

If, as our findings above suggest, reliable variance in ACs does not emanate from dimension-related sources of variance, then from where does it emanate? Our results consistently suggest that there are two major sources of reliable variance in ACs. Those sources are represented by the analogue of a general performance factor (p:s) and the analogue of exercise effects (p:se). Depending on aggregation level, p:s explained between 25.91% and 64.79% of between-participant variance in AC ratings and p:se explained between 24.71% and 53.66%. These two variance sources essentially overwhelm any other source of variance in AC

ratings, reliable or unreliable, aggregated or not. According to our results, the reliable heart of the AC is concerned with its capacity to assess general performance and its capacity to identify variation in performance as a function of exercises. All other reliable sources of variance are either too small to have any noticeable effect on reliability (i.e., p:sd) or are likely to be manifestations of an exercise effect that concerns dimensions (i.e., p:sde). Figure 1 shows our results in graphical form along with credible intervals of each parameter estimate based on posterior distributions. As can be seen, because the number of sub-samples and the number of exercises was small, lower levels of certainty were associated with point estimates for p:s and p:se than for other reliable parameters. However, even when this uncertainty was taken into consideration, p:s and p:se were still clearly larger than any other estimated effect.

We also found evidence for an additional, small reliable effect, which has not been explored in previous research. The p:si:e effect, which explained 4.75% of non-aggregated between-participant variance, summarizes the interaction between participants and exercise-nested behavioral rating items. This effect aligns with a key design feature used in early exercise-based ACs. Specifically, Goodge (1988) and Lowry (1997) describe making use of exercise-nested rating items for the purposes of feedback and development, which makes the p:si:e effect pertinent to the non-aggregated level. The p:si:e effect bears relevance to such applied purposes and, despite explaining a small proportion of between participant variance, it was still almost nine times the magnitude of the comparable dimension effect at the non-aggregated level.

To provide another perspective on our findings, we estimated generalizability theory-based reliability coefficients for PEDRs and also for aggregation to dimension, exercise, and overall scores. This comparison, as yet absent from the literature across all possible aggregation

types in ACs, allows for an analysis of what happens to reliability, contingent on whether exercise- versus dimension-related sources of variance are considered as reliable versus unreliable. Tables 3 and 4 show that whenever exercise-related variance sources were treated as contributing to unreliable variance (i.e., when generalizing to assessors *and* exercises at the PEDR, dimension, and overall levels), the effects on reliability were notably unfavorable (falling from $\geq .89$ to $\leq .65$). This suggests that exercise-related sources of variance should, realistically, always be considered as contributing to reliable variance, regardless as to whether the researcher is interested in PEDRs, dimension scores, or overall scores. Furthermore, we looked at the outcome, when aggregating to exercise-level scores, of considering dimension-related variance as contributing to unreliable variance. Table 4 shows that for exercise scores, dimension-related variance sources had very little impact on reliability (the difference when dimension-related variance sources were treated as reliable versus unreliable was .04), suggesting that reliability in AC ratings ultimately has little to do with dimension-related sources of variance.

Figure 2 provides a Bayesian perspective on reliability in generalizability theory and displays credible intervals, based on posterior distributions, along with reliability parameter estimates (see Gelman et al., 2013). This provides another advantage over traditional approaches to generalizability theory, where levels of uncertainty around reliability estimates are often overlooked. Figure 2 shows that uncertainty was lowest for the reliability of exercise scores, irrespective of generalization type. Uncertainty was highest for generalization to assessors *and* exercises for PEDRs, dimension scores, and overall scores.

An Unconfounded Perspective on AC Ratings

Recent, exercise- and dimension-centric studies have presented an unclear perspective on the role of exercises and dimensions, respectively, because they have confounded the effects of

exercises and dimensions with other AC-related effects. For example, a correlation between a summative dimension score and job performance does not mean that a dimension-related effect is the primary contributor to this correlation. This is because any summative AC-based score, regardless as to how it was aggregated, will reflect the many effects that contribute to AC ratings. Before meaningful conclusions can be drawn about why correlations occur, specific effects need to be isolated, and this is particularly important when multifaceted measures like ACs are considered.

In contrast to previous studies, we aimed to isolate specific AC effects by capitalizing on advances in Bayesian statistics (Gelman, Carlin, Stern, & Rubin, 2004; Gelman & Hill, 2007) that enable researchers to unconfound the many effects that underlie AC ratings. Key differences were observed between our results and those of previous studies. Table 5 shows the results of preceding AC studies, and demonstrates that, when precision is increased (i.e., when a greater number of effects are estimated), the magnitude of dimension-related effects steadily diminishes. Our results also suggest that general performance-related effects play a more prominent role in ACs than previously thought and that exercise effects are almost always prominent (with the exception of those reported by Arthur et al., 2000). Note here that we did not invoke a direct comparison between our results and those of LoPilato et al. (2015). This is because Lopilato et al. did not estimate dimension-related effects; the estimation of which was necessary for cross-study comparisons relating to the aims of the present study.

Our findings generally suggest that confounding in AC ratings is more likely to present a challenge to the dimension-based perspective than to the exercise-based perspective. However, even those favoring the exercise-based perspective need to acknowledge the prominent role of general performance, which appears to emerge as an important influence that is separate from

exercise-related effects. We also find no evidence to favor a mixed perspective on reliable variance in AC ratings, i.e., one based on a combination of dimension- and exercise-based variance sources. Perhaps an alternative direction for the mixed perspective could be oriented towards the combination of exercise-related variance sources with general performance effects.

Expectations Surrounding AC Dimensions

Our take on the divergent perspectives relating to AC ratings is that the "problem", dating back to when exercise effects were first identified in ACs (Sackett & Dreher, 1982; Sakoda, 1952; Turnage & Muchinsky, 1982), is one of expectations and, particularly, expectations surrounding dimensions. To understand the background issues involved, it is helpful to consider early conceptualizations of behavioral dimensions in the Ohio State Leadership Studies. Here, seemingly conflicting ideas were presented which suggested that situationally-contextualized behavioral samples could be clustered into "meaningful categories" (Fleishman, 1953, p. 1). In ACs, this was taken to mean that observations relating to the same dimension could be expected to coalesce, at least to some degree, across exercises (Lance, 2008; Sackett & Dreher, 1982). (Lance, 2008; Sackett & Dreher, 1982). Because of the fact that, in practical scenarios, dimension observations are routinely aggregated across exercises (e.g., Hughes, 2013; Thornton & Krause, 2009), we argue that this belief still exists in AC practice. But the aggregation of dimension observations across different exercises presupposes that dimension observations rated in different exercises fit together meaningfully. In our study, we have isolated a direct analogue of this dimension-related expectation in the p:sd effect, which, according to our data, has very little impact on variance in AC ratings.

It is possible that, due to the Fleishman (1953) tradition that behavioral responses "should" pack neatly into meaningful dimension categories, dimensions have been conceptually confused

with traits. Also, the crossing or partial crossing of dimensions with exercises in ACs is, rightly or wrongly (see Lance, Baranik, Lau, & Scharlau, 2009), reminiscent of a multitrait-multimethod matrix (Campbell & Fiske, 1959), which serves to further reinforce the idea that dimensions equal traits. We consider this a flawed line of reasoning. Our results suggest that any impact on the basis of dimensions, however small, is likely to be one that is specific to exercises (i.e., situationally-specific, as manifested in p:sde interactions).

The investigation of nomological network relationships between summary scores from ACs and externally-measured traits (e.g., Meriac et al., 2008) has considerable merit. However, our results imply that the relationship is unlikely to be between an AC dimension and an externally-measured trait. Rather, it is much more likely to be between a *situationally*-driven behavioral outcome and an externally-measured trait. Only by considering AC ratings in this manner can future studies work to understand the true psychological basis for AC performance. That psychological basis is not, according to our results, manifest in AC dimensions.

Why is the Proportion of Reliable Dimension-Related Variance So Small?

We suggest that the proportion of dimension-related variance in ACs may be influenced by four factors: (a) the magnitude of between-person, dimension-related, variance in job performance, (b) the degree to which this between-person, dimension-related, variance is reproduced in the behaviors observed in ACs, (c) the accurate measurement of this reproduced variance by AC raters, and (d) the degree of theoretical congruence between dimensions and psychological phenomena.

The magnitude of dimension-related job performance variance. In ACs, assessors seek to measure the true (i.e., construct) levels of each candidate on each dimension. The theoretical justification for doing so rests on the assumption that variance in these true levels is

associated with differences in job performance. This raises the question of how much variance in job performance is uniquely dimension-related: an issue addressed by Viswesvaran et al. (2005). Based on a large-scale meta-analytic study, Viswesvaran and his colleagues concluded that about 60% of the construct-level variance in job performance is associated with a general performance factor *independent* of dimensions. If, as Viswesvaran et al.'s study suggests, more than half of the true variance in job performance is independent of dimensions, the amount of uniquely dimension-related variance available for measurement in ACs will surely be limited.

The reproduction of dimension-related job performance variance. ACs are often designed to replicate dimension-related variance in job performance. It is assumed that when candidates perform a series of AC exercises, the true dimension-related variance in their job performance will be reproduced or, at least, approximated. However, several factors are likely to limit the extent to which this can be achieved in practice. These include constraints on the number and range of situations in which behavior is sampled and on the amount of time available to sample behavior in each exercise. In addition, performance-related factors, such the extent to which or how candidates are motivated in the AC and their ability to anticipate and produce the behavior that assessors are seeking to observe in each exercise, may also constrain the extent to which their true levels on each dimension are replicated.

The accurate measurement of dimension-related variance. Dimension-related variance in job performance, as well as being replicated in ACs, must also be accurately measured by assessors. In AC practice, considerable attention is often given to the measurement of dimensions, including extensive assessor training and the use of multiple assessors for each candidate, (Krause & Gebert, 2003; Krause, Rossberger, Dowdeswell, Venter, & Joubert, 2011; Krause & Thornton, 2009; Spychalski, Quińones, Gaugler, & Pohley, 1997). Although such

steps are likely to reduce inaccuracies in the measurement of true levels of dimensions in job performance, it is unlikely that all sources of rater bias and error, including common rater variance (Kolk, Born, & van der Flier, 2002), rater mood (Fried, Levi, Ben-David, Tiegs, & Avital, 2000), halo effects (Viswesvaran, Schmidt, & Ones, 2005), and severity/leniency effects (Kane, Bernardin, Villanova, & Peyrefitte, 1995) are absent in ACs. Bearing these issues in mind, our results and those of Putka and Hoffman (2013) suggest, however, that assessor-related effects in ACs have a fairly trivial influence on AC ratings.

The degree of theoretical congruence between dimensions and psychological phenomena. AC ratings are, by their multifaceted, behavioral nature, made up of numerous influences. Some of these influences are likely to be psychological and, in terms of furthering AC theory, it is vital to develop an understanding of such psychological influences. One possibility is that the psychological phenomena that actually affect AC performance might not be theoretically aligned with or reflected by dimensions. AC dimensions are, perhaps, attempts to directly measure psychological phenomena. However, if most of the reliable variance in ACs is associated with general performance and exercise-related effects, then perhaps the psychological factors truly involved in ACs can be inferred only indirectly from relationships between external psychological measures (e.g., personality and cognitive ability measures) and AC scores.

Limitations and Future Directions

The AC used in this study was developed in a particular context, organization, and job-level. Our AC was, however, developed in keeping with international guidelines (International Task Force on Assessment Center Guidelines, 2009; International Taskforce on Assessment Center Guidelines, 2015) as well as with guidelines in the literature for the development of dimensions (Arthur et al., 2003; Guenole et al., 2013; Lievens, 1998), exercises (Thornton &

Mueller-Hanson, 2004), and on training for ACs (Gorman & Rentsch, 2009; Macan et al., 2011). Through these conditions, our intention was to develop an AC that would be comparable to those utilized in research subsequently reported in peer-reviewed journals.

Despite this intention, we also recognize that it is necessary to investigate possible cross-sample differences relating to the nature of the assessors, the rating processes, and specific AC design issues. With respect to assessors and rating processes, if any sample-specific issues were relevant to the present study, then these should have been manifested in the cross-sample assessor-related effects shown in Table 5. With reference to the four studies listed in Table 5, given that the Bowler and Woehr (2009) and Arthur et al. (2000) decomposed a relatively small number of assessor-related variance sources, the most logical comparison is with the estimates listed in the Putka and Hoffman (2013) study. Cross-study comparisons with respect to assessor-related sources of variance suggest similar data patterning, with the only difference of note relating to the pa:sd effect, which summarizes variance that might have resulted from assessors in different samples using dimensions differently. However, because Putka and Hoffman did not model variance according to different samples, a difference here is expected. Also, the analogous effect here in Putka and Hoffman was, in any case, only marginally different from our estimate (a difference of 3.58%). Because the assessor-related effects summarized in Table 5 account for issues concerning assessors, their rating behavior, and their use of rating instrumentation, our results suggest comparability at least with the Putka and Hoffman study, which involved an independent AC used in a different country with different assessors.

With respect to design elements, many of the design features in our AC, including the dimensions and the exercises, are at least theoretically comparable to examples in the AC literature (e.g., Guenole et al., 2013; Lievens & Conway, 2001; Putka & Hoffman, 2013). An

issue that has been raised in the AC literature is the possibility that there are often conceptual similarities among the different dimensions used in operational ACs (e.g., as implied in Bowler & Woehr, 2006) and this is likely to be true in the present study. If conceptual similarities between different dimensions are apparent, then this has the potential to inflate person main effects. However, our results support the position that any dimension-related effects were very small and, thus, unlikely to have a major influence on any of the other effects that we modeled. Nonetheless, future research could investigate whether there are differences observed in variance profiles for ACs that have conceptually similar versus conceptually different dimensions. In terms of specific design-related comparisons, we make reference to the ACs from the extant literature listed in Table 5, in contrast to which our AC contained fewer dimensions. The inclusion of a relatively low number of dimensions was a purposeful design feature that was oriented towards reducing assessor cognitive load (see Lievens & Conway, 2001) and, thus, was implemented with the intention of optimizing conditions for the assessment of dimensions. Also, the number of exercises that we used was only one less than the average number of exercises listed for the studies in Table 5.

To assuage concerns about the number of dimensions and exercises in our study relative to the other examples listed in Table 5, we applied a decision study from generalizability theory methodology, which is akin to applying the Spearman-Brown prediction formula for multifaceted measures (Brennan, 2001; Shavelson & Webb, 1991). Using a decision study, we extrapolated from our observed reliability outcomes based on three exercises and six dimensions to an alternative design based on five exercises and 13 dimensions: the largest number of exercises and dimensions used in the studies listed in Table 5 (i.e., in the Bowler & Woehr, 2009, study). Under these alternative conditions, the decision study revealed only minimal fluctuations

in reliability when generalizing to different assessors. In specific terms, the reliability of dimension, exercise, and overall scores changed by factors of only .01, <.01, and .01, respectively. Our estimates therefore suggest that the use of a larger number of exercises and dimensions would have been unlikely to have had much effect on the reliability outcomes in our study.

In terms of specifics concerning design content, Appendix Table A1 shows that our AC included two individual and one group exercise and Table A2 shows the definitions of the six dimensions that we used. Specific differences as well as similarities were observed in terms of the dimensions and exercises used across all of the studies listed in Table 5. With respect to dimensions, all studies included dimensions that concerned planning, teamwork, interpersonal influence, and communication skills. However, there were other dimensions that, at least in terms of theory, did not generalize neatly across studies. With respect to exercises, all of the studies listed in Table 5 used an in-basket and one or more goal-oriented group exercises. All studies also included a role play exercise, except for the Arthur et al. (2000) study. There were differences in exercise format, particularly in that Arthur et al. and Bowler and Woehr (2009) used written exercises. We raise cross-sample differences, vis-à-vis dimensions and exercises, as a potential limitation and, perhaps, one that is relevant to any study that invokes cross-AC comparisons.

Our sample was based in South East Asia, which might raise additional concerns about cross-sample generalization (however, see Highhouse & Gillespie, 2009, and the discussion above about assessor-related effects). Participants in the present study were at a managerial level and were attached to a well-established, large-scale organization. Also, previous findings suggest that patterns observed in AC ratings bear similarities when observed across Western and

Eastern nations (Lievens & Christiansen, 2012). Nonetheless, we recognize that, in order to establish more definitive conclusions, it is necessary to investigate the extent to which the unconfounded effects that we found generalize across other samples, organizations, contexts, and job-levels.

Our study introduced two sources of variance that have not yet been considered in a comprehensive decomposition of AC ratings – those concerned with samples and items. The introduction of these variance sources raises a consideration of nested design features because, in our study, participants were nested in samples and rating items were nested in exercises. This implies that each sub-sample in our study contained a unique group of participants and each exercise contained a unique group of items. A potential criticism here is that with nested design features, "it is not possible to estimate all variance components separately" (Shavelson & Webb, 1991, p. 55). While this inability arises from circumstances different from those apparent when effects are confounded (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb), it still results in some effects that cannot be considered in isolation (e.g., p cannot be separated from p:s). Nonetheless, we argue that the nesting involved in our study reflects characteristics that are intrinsic to the design of ACs. In high-stakes circumstances, participants are necessarily nested in different sub-samples. Also, rating items cannot be crossed with (i.e., repeated across) exercises because, if they were, then that would imply that the same exercise was being used repeatedly. Doing so would introduce redundancy into the AC design, which generally involves exercises that differ, content-wise, from one another (International Taskforce on Assessment Center Guidelines, 2015).

Given the presence of relatively large exercise-related effects, our results suggest that future research should explore exercise-based approaches to scoring ACs. Currently there is

very little work on this area and, whilst there are a few, not particularly visible, studies that explore the possibilities of exercise-based scoring approaches (also called task-based ACs, see Jackson & Englert, 2011; Lance, 2012; Lowry, 1997), there are currently no known published studies exploring the psychometric characteristics of ACs designed to be purely exercise-based. All known, published, exercise-based studies of ACs incorporate dimensions as a scoring basis in some way. Given the debates that have circulated in the AC literature (Lievens & Christiansen, 2012), the lack of studies focused on what happens when dimensions are removed from AC scoring approaches appears to be an oversight, and suggests a direction that would assist in informing the exercise-based AC literature.

Guidelines on developing task-based ACs have been in publication for around 20 years (see Lowry, 1995; Lowry, 1997) and have been expanded on in more recent years (Jackson, 2012; Jackson & Englert, 2011; Thoresen & Thoresen, 2012). A task-based AC is more or less identical to a regular AC, except that, in the former, dimensions and any dimension scoring across exercises do not form part of the assessment process (note that this is permissible under the current AC guidelines, see International Taskforce on Assessment Center Guidelines, 2015). Instead of a dimension-based scoring approach, a task-based AC utilizes a list of behavioral indicators that are specific to each exercise and that are based on job analysis data. Thus, output from a task-based AC includes (a) assessor responses to scaled behavioral indicators within each exercise, (b) a score per exercise that is based on exercise-specific behavioral indicators, and (c) an average exercise score. While there is some preliminary research evidence in support of this approach (see Lance, 2012), we know of no published studies that have investigated the psychometric properties of purely task-based ACs. This appears to be a key area for future research.

Our study is, to the best of our knowledge, the second in the organizational literature to employ the use of a Bayesian approach to generalizability theory after LoPilato et al. (2015). However, our application of Bayesian generalizability theory differs from that used by Lopilato et al. on the basis of three points. Firstly, in the Lopilato et al. study, neither item- nor dimension-related effects were modeled and, thus, their model contained fewer effects than did ours (Lopilato et al. = 7 effects, the present study = 29 effects). Our model therefore allowed us to estimate credible intervals for a broader range of effects and reliability coefficients (see Figure 2). Secondly, LoPilato et al. (2015) used uniform distributions to specify empirical, informative, and non-informative priors. A large number of effects were specified in our study and, because of this, we did not have enough prior information for all of our 29 effects. We therefore adopted a middle-ground approach and used weakly informative priors. Specifically, we employed half-Cauchy distributions for all 29 effects, which is the recommended approach for variance component models (Gelman, 2006). Thirdly, we used the No-U-Turn algorithm for HMC (M. D. Hoffman & Gelman, 2014), whereas Lopilato et al. used Gibbs sampling. Although the two algorithms should return the same results, Gibbs samples tend to be highly autocorrelated and therefore a larger number of samples are required to reach convergence. For example Lopilato, et al., used 100,000 iterations, whilst for our model, 10,000 iterations were sufficient to meet all convergence criteria. Conversely, Gibbs sampling can be more efficient per iteration than HMC. However, without a direct comparison using the same model and data, it is almost impossible to say which approach would be more efficient with respect to convergence.

Bayesian approaches present many possibilities for the progress of research, and we have only explored a few of these possibilities here. Bayesian approaches hold the potential to provide researchers with flexibility in terms of their application of statistical methods, and to

facilitate the examination of complex models. Such opportunities are worthy of exploration, not only as they pertain to the estimators used in generalizability theory, but also as they apply to a host of different analytical approaches.

Concluding Comments

Our research aimed to address the problem of confounding in studies of AC ratings. Our results reveal that when sources of variance in AC ratings are appropriately decomposed, and even when taking aggregation-level into consideration, dimension-based sources explain very little of the variance and have very little impact on the reliability of the ratings. In our unconfounded study, much more variance was explained by general performance and exercise-based sources. These findings call for further investigation into the primary reasons for correlations between summative scores based on AC ratings and outcomes. They suggest a challenge to the belief apparently espoused by proponents of the dimension approach, that such relationships are the result of the dimensions purportedly measured in ACs. In challenging this view, our findings also present a challenge to the mixed perspective that reliable variance results from a combination of dimension- and exercise-related variance. Our findings partly support the use of exercise- or task-based ACs; however, they also suggest that the role of general performance requires a greater emphasis and more thorough investigation.

References

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). New York: Routledge.

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*, 125-154. doi: 10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., Jr., Woehr, D. J., & Maldegan, 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*, 813-835.

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*, 1114-1124. doi: Doi 10.1037/0021-9010.91.5.1114

Bowler, M. C., & Woehr, D. J. (2009). Assessment center construct-related validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior, 75*, 173-182. doi: 10.1016/j.jvb.2009.03.008

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer Verlag.

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

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology, 69*, 167-181. doi: 10.1111/j.2044-8325.1996.tb00608.x

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika, 78*, 685-709. doi: 10.1007/s11336-013-9328-2

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*, 137-163. doi: 10.1111/j.2044-8317.1963.tb00206.x

Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behaviour and Personality, 12*, 85-108.

Fleishman, E. A. (1953). The description of supervisory behavior. *Journal of Applied Psychology, 37*, 1-6. doi: 10.1037/h0056314

Fried, Y., Levi, A. S., Ben-David, H. A., Tiegs, R. B., & Avital, N. (2000). Rater positive and negative mood predispositions as predictors of performance ratings of ratees in simulated and real organizational settings: Evidence from US and Israeli samples. *Journal of Occupational and Organizational Psychology, 73*, 373-378.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*, 515-533.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). New York: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Goodge, P. (1988). Task-based assessment. *Journal of European Industrial Training, 12*, 22-27.

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*, 1336-1344. doi: 10.1037/a0016476

Guenole, N., Chernyshenko, O. S., Stark, S., Cockerill, T., & Drasgow, F. (2013). More than a mirage: A large-scale assessment centre with more dimension variance than exercise variance. *Journal of Occupational and Organizational Psychology, 86*, 5-21. doi: 10.1111/j.2044-8325.2012.02063.x

Herold, D. M., & Fields, D. L. (2004). Making sense of subordinate feedback for leadership development: Confounding effects of job role and organizational rewards. *Group and Organization Management, 29*, 686-703. doi: 10.1177/1059601103257503

Highhouse, S., & Gillespie, J. Z. (2009). Do samples really matter that much? In R. J. Vandenberg & C. E. Lance (Eds.), *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences* (pp. 247-265). New York: Routledge.

Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology, 100*, 1143-1168. doi: 10.1037/a0038707

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*, 351-395. doi: 10.1111/j.1744-6570.2011.01213.x

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*, 1593-1623.

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

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

Hughes, D. (2013, November). *Evidence-based practices in assessment centres: Strengths, concerns, and challenges from a global survey.* Paper presented at the United Kingdom Assessment Centre Conference, Surrey, UK.

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*, 243-253. doi: 10.1111/ijsa.2009.17.issue-310.1111/j.1468-2389.2009.00467.x

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

Jackman, S. (2009). *Bayesian analysis for the social sciences*. West Sussex, England: Wiley.

Jackson, D. J. R. (2012). Task-based assessment centers: Theoretical perspectives. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers.* (pp. 173-189). New York, NY US: Routledge/Taylor & Francis Group.

Jackson, D. J. R., & Englert, P. (2011). Task-based assessment centre scores and their relationships with work outcomes. *New Zealand Journal of Psychology, 40*, 37-46.

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

Jansen, A., Melchers, K. G., Lievens, F., Kleinmann, M., Brändli, M., Fraefel, L., & König, C. J. (2013). Situation assessment as an ignored factor in the behavioral consistency paradigm underlying the validity of personnel selection procedures. *Journal of Applied Psychology, 98*, 326-341. doi: 10.1037/a0031257

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal, 38*, 1036-1051. doi: 10.2307/256619

Kolk, N. J., Born, M. P., & van der Flier, H. (2002). Impact of common rater variance on construct validity of assessment center dimension judgments. *Human Performance, 15*, 325-337.

Krause, D. E., & Gebert, D. (2003). A comparison of assessment center practices in organizations in German-speaking regions and the United States. *International Journal of Selection and Assessment, 11*, 297-312. doi: 10.1111/j.0965-075X.2003.00253.x

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*, 262-275. doi: DOI 10.1111/j.1468-2389.2011.00555.x

Krause, D. E., & Thornton, G. C., III. (2009). A cross-cultural look at assessment center practices: Survey results from Western Europe and North America. *Applied Psychology: An International Review, 58*, 557-585. doi: DOI 10.1111/j.1464-0597.2008.00371.x

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*, 722-752. doi: 10.1177/1094428112457829

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

Laczo, R. M., Sackett, P. R., Bobko, P., & Cortina, J. M. (2005). A comment on sampling error in the standardized mean difference with unequal sample sizes: Avoiding potential errors in meta-analytic and primary research. *Journal of Applied Psychology, 90*, 758-764. doi: 10.1037/0021-9010.90.4.758

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 84-97. doi: 10.1111/j.1754-9434.2007.00017.x

Lance, C. E. (2012). Research into task-based assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers.* (pp. 218-233). New York, NY US: Routledge/Taylor & Francis Group.

Lance, C. E., Baranik, L. E., Lau, A. R., & Scharlau, E. A. (2009). If it ain't trait it must be method: (Mis)application of the multitrait-multimethod design in organizational research. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and*

urban legends: Doctrine, verity and fable in the organizational and social sciences. (pp. 337-360). New York, NY US: Routledge/Taylor & Francis Group.

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*, 377-385. doi: 10.1037/0021.9010.89.2.377

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

Lievens, F., & Christiansen, N. D. (2012). Core debates in assessment center research: Dimensions 'versus' exercises. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 68-91). New York: Routledge.

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*, 1202-1222.

LoPilato, A. C., Carter, N. T., & Wang, M. (2015). Updating generalizability theory in management research: Bayesian estimation of variance components. *Journal of Management, 41*, 692-717. doi: 10.1177/0149206314554215

Lowry, P. E. (1995). The assessment center process: Assessing leadership in the public sector. *Public Personnel Management, 24*, 443-450.

Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior and Personality, 12*, 53-62.

Macan, T., Mehner, K., Havill, L., Meriac, J. P., Roberts, L., & Heft, L. (2011). Two for the price of one: Assessment center training to focus on behaviors can transfer to performance appraisals. *Human Performance, 24*, 443-457. doi: 10.1080/08959285.2011.614664

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports, 66*, 379-386. doi: 10.2466/PR0.66.2.379-386

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*, 1269-1296. doi: 10.1177/0149206314522299

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: a meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042-1052. doi: 10.1037/0021-9010.93.5.1042

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*, 1009-1047. doi: 10.1111/peps.12049

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science, 22*, 59-73. doi: 10.1214/088342307000000014

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*, 114-133. doi: 10.1037/a0030887

Putka, D. J., & Hoffman, B. J. (2014). "The" reliability of job performance ratings equals 0.52. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 247-275). New York: Taylor & Francis.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: implications for estimating interrater reliability. *Journal of Applied Psychology, 93*, 959-981. doi: 2008-12803-017 [pii] 10.1037/0021-9010.93.5.959

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of Employee Selection* (pp. 9-49). New York: Routledge.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*, 71-84.

Robie, C., Osburn, H. G., Morris, M. A., Etchegaray, J. M., & Adams, K. A. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance, 13*, 355-370.

Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., & McCormick, S. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology, 80*, 664-670.

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

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology, 59*, 419-450. doi: 10.1146/annurev.psych.59.103006.093716

Sakoda, J. M. (1952). Factor analysis of OSS situational tests. *Journal of Abnormal and Social Psychology, 47*, 843-852.

Schippmann, J. S., Hughes, G. L., & Prien, E. P. (1987). The use of structured multi-domain job analysis for the construction of assessment center methods and procedures. *Journal of Business and Psychology, 1*, 353-366.

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components*. New York: Wiley.

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

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932.

Silverman, W. H., Dalessio, A., Woods, S. B., & Johnson, R. L. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology, 39*, 565-578.

Speer, A. B., Christiansen, N. D., Goffin, R. D., & Goff, M. (2014). Situational bandwidth and the criterion-related validity of assessment center ratings: Is cross-exercise convergence always desirable? *Journal of Applied Psychology, 99*, 282-295. doi: 10.1037/a0035213

Spychalski, A. C., Quińones, M. A., Gaugler, B. B., & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology, 50*, 71-90.

Stan Development Team. (2015a). RStan: the R interface to Stan (Version 2.7.0).

Stan Development Team. (2015b). Stan: A C++ library for probability and sampling (Version 2.7.0).

Thoresen, C. J., & Thoresen, J. D. (2012). How to design and implement a task-based assessment center. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 190-217). New York: Routledge.

Thornton, G. C., III., & Krause, D. E. (2009). Selection versus development assessment centers: An international survey of design, execution, and evaluation. *International Journal of Human Resource Management, 20*, 478-498. doi: 10.1080/09585190802673536

Thornton, G. C., III., & Mueller-Hanson, R. A. (2004). *Developing organizational simulations: A guide for practitioners and students*. Mahwah, NJ: Routledge.

Turnage, J. J., & Muchinsky, P. M. (1982). Trans-situational variability in human performance with assessment centers. *Organizational Behavior and Human Performance, 30*, 174-200.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108-131. doi: 10.1037/0021-9010.90.1.108

Walter, F., Cole, M. S., van der Vegt, G. S., Rubin, R. S., & Bommer, W. H. (2012). Emotion recognition and emergent leadership: Unraveling mediating mechanisms and boundary conditions. *The Leadership Quarterly, 23*, 977-991. doi: 10.1016/j.leaqua.2012.06.007

Williams, K. M., & Crafts, J. L. (1997). Inductive job analysis: The job/task inventory method. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 51-88). Palo Alto, CA: Davies-Black Publishing.

Zyphur, M. J., Oswald, F. L., & Rupp, D. E. (2015). Rendezvous overdue: Bayes analysis meets organizational research. *Journal of Management, 41*, 387-389. doi: 10.1177/0149206314549252

Table 1
Demographic Characteristics by Sub-Sample

Characteristic	Sub-sample				
	1	2	3	4	5
Sex					
Male (frequency)	97.00	116.00	111.00	152.00	113.00
Female (frequency)	11.00	7.00	26.00	33.00	32.00
Total (frequency)	108.00	123.00	137.00	185.00	145.00
Age					
Mean (years)	52.84	54.36	53.05	53.64	52.36
SD (years)	4.30	3.82	4.32	3.65	4.55

Note. All participants were nationals based in South East Asia and employed in a line manager role. Combined sample $N = 698$.

Table 2
Variance Decomposition of Assessment Center Ratings

Source of variance	VE	CI (lo)	CI (hi)	Total (%)	B-W (%)	D-Score (%)	E-Score (%)	O-Score (%)
Reliable								
p:s	.2469	.2083	.2885	23.92	25.91	53.71	38.87	64.79
p:sd	.0051	<.0001	.0124	0.49	0.54	1.11	0.13	0.22
p:se	.3408	.3128	.3706	33.02	35.76	24.71	53.66	29.81
p:sde	.1215	.1121	.1307	11.77	12.75	8.81	3.19	1.77
p:si:e	.0453	.0425	.0483	4.39	4.75	1.00	0.73	0.37
Subtotal	-	-	-	73.59	79.71	89.35	96.58	96.97
Unreliable								
a*	.0038	.0013	.0073	0.05	0.05	0.11	0.08	0.13
p:si:eda + residual	.1424	.1394	.1453	13.80	14.94	3.15	0.38	0.19
p:sa	.0019	<.0001	.0063	0.18	0.20	0.41	0.30	0.50
pa:sd	.0240	.0106	.0349	2.33	2.52	5.22	0.63	1.05
pa:se	.0106	.0062	.0139	1.03	1.11	0.77	1.67	0.93
pa:sde	.0109	.0003	.0241	1.06	1.14	0.79	0.29	0.16
as*	.0005	<.0001	.0020	0.01	0.01	0.01	0.01	0.02
ad*	.0012	<.0001	.0029	0.02	0.02	0.04	<0.01	0.01
ae*	.0002	<.0001	.0010	<0.01	<0.01	<0.01	<0.01	<0.01
ade*	.0003	<.0001	.0012	<0.01	<0.01	<0.01	<0.01	<0.01
asd*	.0027	.0002	.0051	0.04	0.04	0.08	0.01	0.02
ase*	.0003	<.0001	.0012	<0.01	<0.01	<0.01	0.01	<0.01
asde*	.0006	<.0001	.0028	0.01	0.01	0.01	<0.01	<0.01
ai:e*	.0058	.0044	.0074	0.08	0.08	0.02	0.01	0.01
asi:e*	.0111	.0094	.0129	0.14	0.16	0.03	0.02	0.01
Subtotal	-	-	-	18.73	20.29	10.65	3.42	3.03
Reliability-unrelated								
d	.0286	.0004	.1059	2.77	-	-	-	-
e	.0075	<.0001	.0472	0.73	-	-	-	-
i:e	.0157	.0078	.0288	1.52	-	-	-	-
s	.0029	<.0001	.0163	0.28	-	-	-	-
si:e	.0037	.0023	.0057	0.36	-	-	-	-
se	.0033	<.0001	.0137	0.32	-	-	-	-
sd	.0016	<.0001	.0074	0.16	-	-	-	-
ed	.0060	<.0001	.0386	0.58	-	-	-	-
sde	.0099	.0046	.0172	0.96	-	-	-	-
Subtotal	-	-	-	7.67	-	-	-	-

Note. VE = variance estimate; CI (lo) = 2.50% credible interval; CI (hi) = 97.50% credible interval; p = participant; s = sample; e = exercise; i = item; d = dimension; a = assessor; o = overall; B-W = between-participant sources of variance; D-, E-, and O-Score = VE aggregated to dimensions, exercises, and overall scores, respectively. *These estimates have been rescaled using the *q*-multiplier, given the ill-structured nature of the measurement design herein (Putka et al., 2008).

Table 3
Composition of Variance and Generalization for Post Exercise Dimension Ratings and Dimension Scores

Level/G	Variance Composition		$E\rho^2$	Interpretation
	Reliable	Unreliable		
PEDRs				
a	p:s, p:sd, p:se, p:sde, p:si:e	a*, p:si:eda, p:sa, pa:sd, pa:se, pa:sde, as*, ad*, ae*, ade*, asd*, ase*, asde*, ai:e*, asi:e*	.80	Expected correlation between PEDRs for a given dimension-exercise combination rated by two different assessors
a,e	p:s, p:sd	p:se, p:sde, p:si:e, a*, p:si:eda, p:sa, pa:sd, pa:se, pa:sde, as*, ad*, ae*, ade*, asd*, ase*, asde*, ai:e*, asi:e*	.26	Expected correlation between PEDRs for a given dimension as rated in two different exercises by two different assessors
Dimensions				
a	p:s, p:sd, p:se/n_e, p:sde/n_e, p:si:e/$n_{i:e}$	a*, p:si:eda/$n_{i:e}$, p:sa, pa:sd, pa:se/n_e, pa:sde/n_e, as*, ad*, ae/n_e*, ade/n_e*, asd*, ase/n_e*, asde/n_e*, ai:e/$n_{i:e}$*, asi:e/$n_{i:e}$*	.89	Expected correlation between PEDRs averaged across n_e exercises for a given dimension as rated by two different assessors
a,e	p:s, p:sd	p:se/n_e, p:sde/n_e, p:si:e/$n_{i:e}$, a*, p:si:eda/$n_{i:e}$, p:sa, pa:sd, pa:se/n_e, pa:sde/n_e, as*, ad*, ae*/n_e, ade*/n_e, asd*, ase/n_e*, asde/n_e*, ai:e/$n_{i:e}$*, asi:e/$n_{i:e}$*	.55	Expected correlation between PEDRs averaged across n_e exercises for a given dimension as rated by one assessor and PEDRs averaged across a new set of n_e exercises as rated by a different assessor

Note. Level = post exercise dimension ratings (PEDRs) or aggregation to dimension scores; p = participants; s = samples; d = dimensions; e = exercises; i = response items; G = generalization to different assessors (a) or different assessors and exercises (a,e); $E\rho^2$= expected reliability, estimated as the proportion of reliable between-participant variance; n_e = number of exercises (in this study = 3); n_d = number of dimensions (in this study = 6); $n_{i:e}$ = number of items nested in exercises, which was estimated using the harmonic mean number of items per exercise, in keeping with the suggestions of Brennan (2001, in this study = 9.83). *These variance components were rescaled using the *q*-multiplier for ill-structured measurement designs (Putka et al., 2008).

Table 4
Composition of Variance and Generalization for Exercise and Overall Scores

Level/G	Variance Composition		$E\rho^2$	Interpretation
	Reliable	Unreliable		
Exercises				
a	p:s, p:sd/n_d, p:se, p:sde/n_d, p:si:e/$n_{i:e}$	a*, p:si:eda/$n_{i:e}n_d$, p:sa, pa:sd/n_d, pa:se, pa:sde/n_d, as*, ad/n_d*, ae*, ade/n_d*, asd/n_d*, ase*, asde/n_d*, ai:e/$n_{i:e}$*, asi:e/$n_{i:e}$*	.97	Expected correlation between PEDRs averaged across n_d dimensions for a given exercise as rated by two different assessors
a,d	p:s, p:se, p:si:e/$n_{i:e}$	p:sd/n_d, p:sde/n_d, a*, p:si:eda/$n_{i:e}n_d$, p:sa, pa:sd/n_d, pa:se, pa:sde/n_d, as*, ad/n_d *, ae*, ade/n_d*, asd/n_d*, ase*, asde/n_d*, ai:e/$n_{i:e}$*, asi:e/$n_{i:e}$*	.93	Expected correlation between PEDRs averaged across n_d dimensions for a given exercise measuring two different *sets* of dimensions and rated by two different assessors
Overall				
a	p:s, p:sd/n_d, p:se/n_e, p:sde/n_dn_e, p:si:e/n_i	a*, p:si:eda/n_in_d, p:sa, pa:sd/n_d, pa:se/n_e, pa:sde/n_dn_e, as*, ad/n_d*, ae/n_e*, ade/n_dn_e*, asd/n_d*, ase/n_e*, asde/n_dn_e*, ai:e/n_i*, asi:e/n_i*	.97	Expected correlation between PEDRs averaged across n_e exercises and n_d dimensions for an overall score as rated by two different assessors
a,e	p:s, p:sd	p:se/n_e, p:sde/n_dn_e, p:si:e/n_i, a*, p:si:eda/n_in_d, p:sa, pa:sd/n_d, pa:se/n_e, pa:sde/n_dn_e, as*, ad/n_d*, ae/n_e*, ade/n_dn_e*, asd/n_d*, ase/n_e*, asde/n_dn_e*, ai:e/n_i*, asi:e/n_i*	.65	Expected correlation between PEDRs averaged across n_d dimensions and n_e exercises for an overall score from two different *sets* of exercises and rated by two different assessors

Note. Level = aggregation to exercise or overall scores; p = participants; s = samples; d = dimensions; e = exercises; i = response items; G = generalization to different assessors (a) or different assessors and dimensions (a,d) or assessors and exercises (a,e); $E\rho^2$= expected reliability, estimated as the proportion of reliable between-participant variance; n_e = number of exercises (in this study = 3); n_d = number of dimensions (in this study = 6); $n_{i:e}$ = number of items nested in exercises, which was estimated using the harmonic mean number of items per exercise, in keeping with the suggestions of Brennan (2001, in this study = 9.83); n_i = total number of items across all exercises (in this study = 32). Note that for exercise-level scores, if p:si:e is treated as error, the effect on the reliability estimate is minimal (for generalization to a, $E\rho^2$ = .96, for generalization to a,d, $E\rho^2$= .93). *These variance components were rescaled using the *q*-multiplier for ill-structured measurement designs (Putka et al., 2008).

Table 5
Comparisons with Existing Random Effects Studies

Source of variance	Present Study	Putka & Hoffman (2013)	Bowler & Woehr (2009)	Arthur et al. (2000)
Reliable				
p:s	25.91	17.20	5.20	24.30
p:sd	0.54	1.10	18.00	27.00
p:se	35.76	35.20	32.00	6.80
p:sde	12.75	23.40	-	-
p:si:e	4.75	-	-	-
Subtotal	79.71	76.80	55.20	58.10
Unreliable				
a	0.05	0.50	0.00	8.10
p:si:eda + residual	14.94	11.80	44.80	33.80
p:sa	0.20	1.00	0.00	-
pa:sd	2.52	6.10	-	-
pa:se	1.11	2.10	-	-
pa:sde	1.14	-	-	-
as	0.01	-	-	-
ad	0.02	0.80	0.00	-
ae	<0.01	0.20	0.00	-
ade	<0.01	0.60	-	-
asd	0.04	-	-	-
ase	<0.01	-	-	-
asde	0.01	-	-	-
ai:e	0.08	-	-	-
asi:e	0.16	-	-	-
Subtotal	20.29	23.10	44.80	41.90
N(e)	3	4	5	4
N(d)	6	12	13	9

Note. Values represent percentages of between-participant variance in ratings accounted for by a given effect. VE = variance estimate; p = participant; s = sample; e = exercise; i = response item; d = dimension; a = assessor. Estimates from previous studies are based on Tables 4 and 6 from Putka and Hoffman (2013).

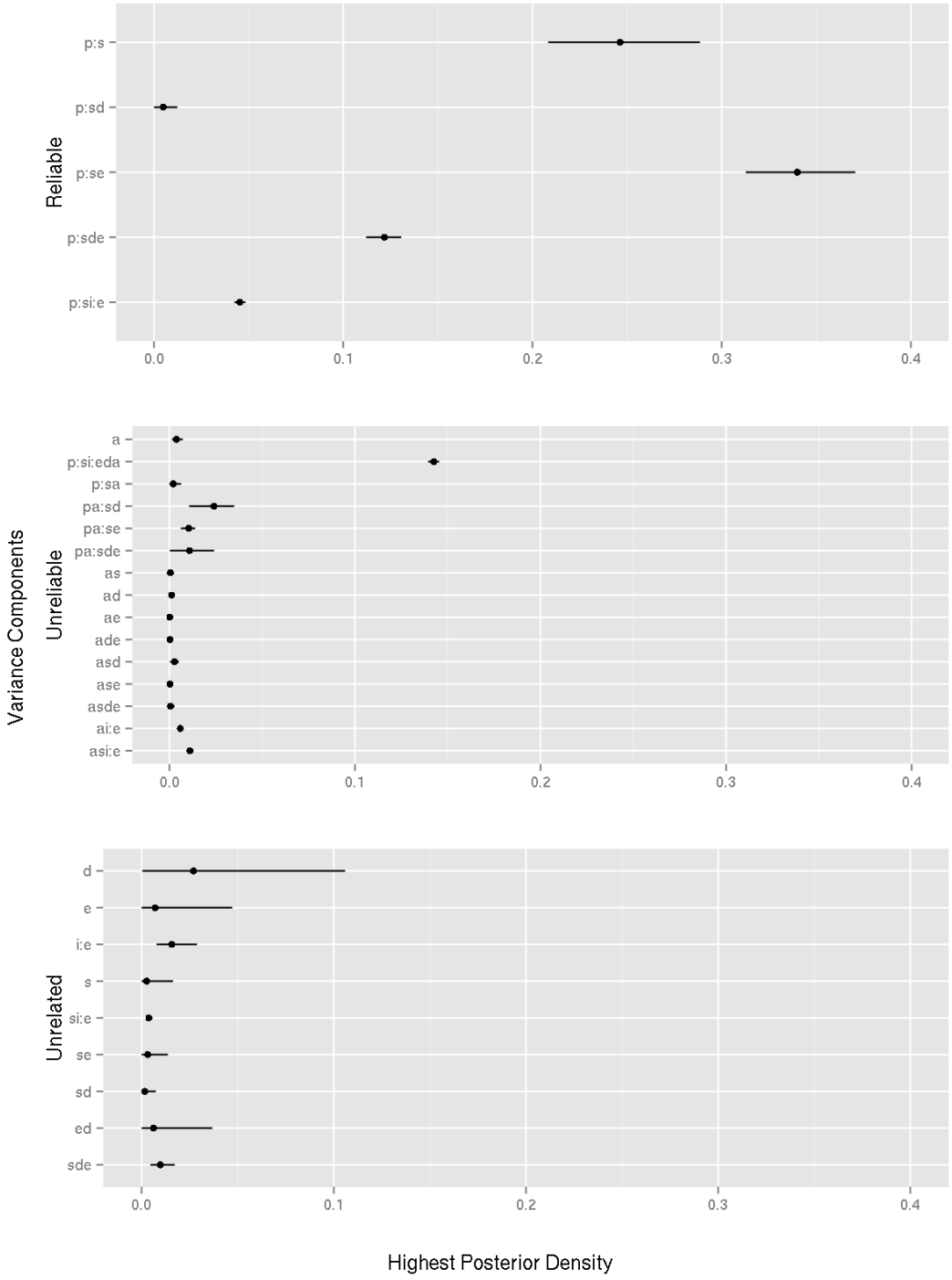

Figure 1. Variance estimates, grouped by reliable, unreliable, and reliability-unrelated status, plotted as a function of effect magnitude. Error bars show credible intervals, with wider intervals suggesting lower levels of certainty with respect to point estimates. p = participant, s = sample, d = dimension, e = exercise, i = response item.

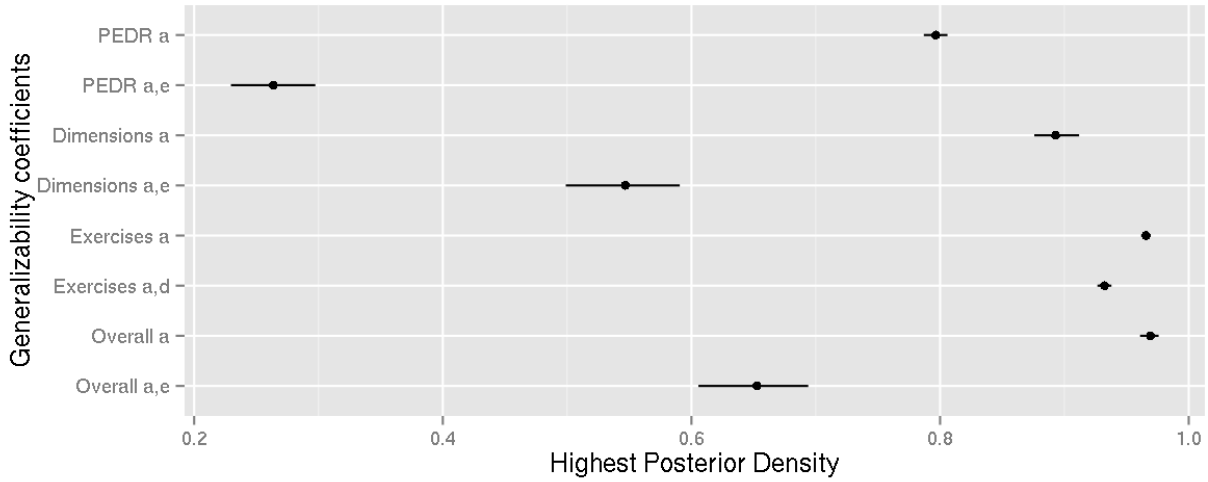

Figure 2. Generalizability coefficients for post exercise dimension ratings (PEDRs) and aggregate dimension scores (dimensions), exercise scores (exercises), and overall scores (overall) plotted as a function of magnitude. Each coefficient is shown as it relates to generalization to assessors (a), assessors and exercises (a,e), or assessors and dimensions (a,d), as appropriate. Error bars show credible intervals, with wider intervals suggesting lower levels of certainty with respect to point estimates.

Appendix

Table A1
Exercise Descriptions

Exercise	Description
In-Basket	Candidates are presented with a mixture of emails, memos, and phone messages that are typical of those experienced in the focal position. Candidates are expected to negotiate problems presented to them in a multimedia format and to achieve an effective outcome under a degree of time pressure.
Role Play	Candidates are expected to conduct a meeting with a line manager from a different department involving a set of challenging issues, including an employment challenge, a negotiation, and a problem around communication.
Case Study	Following an intensive briefing, candidates are requested to prepare a report with the aim of establishing plans, policies, and novel ideas to assist a mock organization to develop and to negotiate a challenging business environment.

Table A2
Dimension Definitions

Dimension	Definition
Policy Planning	Recognizing problems in pending issues, providing solutions for them, and establishing policies logically and systematically.
Teamwork Management	Building trust among team members and fostering a cooperative team climate. Leading the team by motivating and supporting members to achieve their goals.
Outcome Orientation	Establishing action plans for policy implementation, checking the level of task achievement, and developing outcomes that fit goals by proactively completing tasks.
Negotiation and Arbitration	Garnering opinions from interested parties (e.g., public customers, directors of other departments), which are related to coordination and consensus and providing balanced solutions based on evidence that can be adjusted to suit different interests.
Communication Skills	Understanding another party's perspective by carefully listening to them and logically conveying one's own opinion using appropriate communication approaches.
Change Management	Acknowledging (internal/external) organizational administrative climate changes and providing necessary improvements as well as actively fostering a participative organizational culture.

Table A3
Dimension by Exercise Matrix

Dimension	Exercise		
	In-Basket	Role Play	Case Study
Policy Planning	X		X
Teamwork Management	X		X
Outcome Orientation	X		X
Negotiation and Arbitration	X	X	
Communication Skills	X	X	
Change Management		X	X

Note. An X indicates exercises in which a given dimension was rated.

Table A4
Guide to Reliable Sources of Assessment Center Variance

Effect	Brief description	CFA analogue
p:s	Some participants (nested in samples) are generally rated higher than others, regardless of the dimension, exercise, assessor, or item involved.	General factors
p:sd	Some participants (nested in samples) are rated higher on some dimensions relative to others, regardless of the exercise, assessor, or item involved.	Dimension factors
p:se	Some participants (nested in samples) are rated higher on some exercises than others, regardless of the dimension, assessor, or item involved.	Exercise factors
p:sde	Some participants (nested in samples) are rated higher on dimension-exercise combinations than others, regardless of the items or assessors involved.	None
p:si:e[a]	Some participants (nested in samples) are rated higher on some sets of exercise-nested items than on other sets of exercise-nested items, regardless of the dimension or assessor involved.	None

Note. CFA = confirmatory factor analysis. p = participant, s = sample, d = dimension, e = exercise, i = response item. In this study, the CFA analogue of uniqueness is reflected in the highest-level effect (p:si:eda), which is confounded with residual error. [a]We define p:si:e as a reliable source of variance because exercise-nested items are used to guide developmental feedback in exercise-based assessment centers (Lowry, 1997).