The Internal Structure of Situational Judgment Tests Reflects Candidate Main Effects: Not Dimensions or Situations

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

Alexander C. LoPilato
Georgia Institute of Technology

Dan Hughes
JCA Ltd

Nigel Guenole
Goldsmiths, University London

Ali Shalfrooshan
a&dc Ltd

Author Note

Duncan J. R. Jackson, Department of Organizational Psychology, Birkbeck, University of London; Faculty of Management, University of Johannesburg; Alexander C. LoPilato, School of Psychology, Georgia Institute of Technology; Dan Hughes, Product Development, JCA Ltd; Nigel Guenole, Department of Psychology, Goldsmiths, University of London; Ali Shalfrooshan, Research and Development, a&dc Ltd.

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 their popularity and capacity to predict performance, there is no clear consensus on the internal measurement characteristics of situational judgment tests (SJTs). Contemporary propositions in the literature focus on treating SJTs as methods, as measures of dimensions, or as measures of situational responses. However, empirical evidence relating to the internal structure of SJT scores is lacking. Using generalizability theory, we decomposed multiple sources of variance for three different SJTs used with different samples of job candidates (N_1 = 2,320; N_2 = 989; N_3 = 7,934). Results consistently indicated that (a) the vast majority of reliable observed score variance reflected SJT-specific candidate main effects, analogous to a general judgment factor and that (b) the contribution of dimensions and situations to reliable SJT variance was, in relative terms, negligible. These findings do not align neatly with any of the proposals in the contemporary literature; however they do suggest an internal structure for SJTs.

Practitioner Points

- To help optimize reliable variance, overall-level aggregation should be used when scoring SJTs.
- The majority of reliable variance in SJTs reflects a general performance factor, relative to variance pertaining to specific dimensions or situations.
- SJT developmental feedback should be delivered in terms of general SJT performance rather than on performance relating to specific dimensions or situations.
- Generalizability theory, although underutilised in multifaceted measurement, offers an approach to informing on the psychometric properties of SJTs that is well-suited to the complexities of SJT measurement designs.

The Internal Structure of Situational Judgment Tests Reflects Candidate Main Effects: Not Dimensions or Situations

Situational judgment tests (SJTs) comprise low-fidelity simulations often used in high-stakes circumstances in which respondents are required to indicate hypothetical responses to a range of situational dilemmas (Catano, Brochu, & Lamerson, 2012; Lievens, Buyse, & Sackett, 2005; Motowidlo, Crook, Kell, & Naemi, 2009). SJTs are scored by comparing responses to a predetermined scoring key defined by subject matter experts, empirical validation, and/or a theoretical model (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; Sternberg et al., 2000; Wagner & Sternberg, 1985). Researchers have consistently demonstrated that SJTs predict performance across a range of different organisational contexts (Christian, Edwards, & Bradley, 2010; McDaniel, Bruhn Finnegan, Morgeson, & Campion, 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007; Murphy & Shiarella, 1997; Rockstuhl, Ang, Ng, Lievens, & Van Dyne, 2015).

Although their capacity to predict performance is well established, there is no clear consensus in the literature about the internal measurement characteristics of SJTs. Historically, SJTs were thought to measure global constructs, such as tacit knowledge (Wagner & Sternberg, 1985), adaptability (Schmitt & Chan, 2006), or job knowledge (Schmidt & Hunter, 1993). Contemporary perspectives have tended to depart from the global constructs view and have, instead, addressed (a) an SJTs-as-methods perspective focused on correlations between SJT scores and externally-measured constructs, (b) discrete dimensions assigned specifically for measurement by SJTs, and (c) situationally-specific responses. In the present study, we capitalise on recent statistical advances with the aim of decomposing multiple sources of variance in SJTs in order to establish whether our evidence supports or refutes the SJTs-as-

methods, dimension, and/or situation perspectives from the literature. This aim is important because, without understanding the internal measurement properties of SJTs, theory relating to SJTs cannot be properly developed. Given the popularity of SJTs in high-stakes circumstances, fostering an understanding of their underlying measurement characteristics and the reasons for predictor-criterion relationships involving SJTs also has implications for employee selection decisions. Moreover, a failure to understand internal measurement characteristics means that the field will lack clarity on how to appropriately apply SJTs (e.g., for developmental feedback) or on how to improve them.

The SJTs-as-Methods Perspective

The SJTs-as-methods perspective views SJT internal construct evidence as arising by implication from correlations between SJT scores and externally-measured constructs. Responding to the idea that SJTs measure global constructs, McDaniel and Whetzel (2005, p. 523) stated that SJTs are "best viewed not as measures of a single construct, but as measurement methods that can and typically do assess the established constructs of *g*, conscientiousness, emotional stability, and agreeableness". The SJTs-as-methods approach involves correlating SJT scores with other, external measures and making inferences about what SJTs assess based on these correlations. Thus, if SJTs consistently correlate with *g* and personality, then it is inferred that SJTs measure *g* and personality.

In support of the SJTs-as-methods perspective, several meta-analyses have found that SJT scores indeed relate to general mental ability and personality variables (Arthur et al., 2014; McDaniel et al., 2001; McDaniel et al., 2007; McDaniel & Nguyen, 2001). However, under this perspective, the internal measurement structure specific to SJTs is essentially sidestepped and, as

a result, there is no way of knowing what it is about SJTs that is reliable and, thus, leads to observed correlations with externally-measured constructs.

The move to accept SJT scores as SJT-method scores is possibly influenced by challenges that have historically arisen when attempting to study item-specific variance in SJTs. While there are exceptions (see Sharma, Gangopadhyay, Austin, & Mandal, 2013), Schmitt and Chan's (2006, p. 140) review suggests that attempts at isolating an internal structure for SJTs based on exploratory factor analysis (EFA) have often resulted in "disappointing" outcomes. Furthermore, McDaniel and Whetzel (2005, p. 519) stated that the "construct heterogeneity of SJT items makes coherent factor analysis results difficult" and they went on to report mixed results from factor solutions that were mostly uninterpretable.

The SJTs-as-methods perspective provides compelling evidence that SJT scores share variance with well-established external construct measures. However, a potentially uncomfortable element to this perspective is that it essentially avoids an acknowledgement of what is going on, structurally, inside the SJT itself. Relationships between SJT scores and the likes of general mental ability and personality might be consistently apparent. However, what is it about SJTs that might lead to such relationships? Is it purported dimensions? Is it their situational elements? Or is it something else? At present, because there is no clear evidence regarding what SJTs measure internally, there is no consensus about where reliable variance stems from in SJTs and, thus, why SJTs correlate with externally-measured constructs.

The Dimension Perspective

In contrast to the SJTs-as-methods perspective, the dimension perspective encourages researchers to concentrate on constructs measured "directly" within the internal structure of SJTs. On this note, Christian et al. (2010, p. 87) recommended that SJTs should be developed to

"inherently tap certain constructs" in the form of discrete dimensions (e.g., *leadership*, *teamwork skills*). Dimensions are common in the SJT literature (see Christian et al., for a review). While proponents of the previously-presented SJTs-as-methods perspective essentially view SJTs as methods, conversely, Christian et al. urged "researchers to maintain the distinction between methods (e.g., SJTs) and constructs (e.g., leadership skills) by reporting information about the specific constructs measured by SJTs" (p. 107). Christian et al. further lamented that SJT "test developers and researchers often give little attention to the constructs measured by SJTs and instead tend to report results based on overall (or composite) SJT scores" (p. 84).

Similar to the SJTs-as-methods perspective, a concern with the dimension perspective is that EFA results from SJTs "seldom yield interpretable factors" (Whetzel & McDaniel, 2009, p. 190). Thus, evidence for substructures within SJTs that resemble discrete sets of meaningful and interpretable dimensions is limited (McDaniel & Whetzel, 2005; Schmitt & Chan, 2006). As we discuss later, some of the research on the role of dimensions in SJTs might be limited by the analytical strategies that have been applied to SJT data. It is possible that the SJT literature could stand to gain from "lessons learned" in other areas of the organisational literature with respect to dimensions and their measurement properties. In particular, the assessment centre (AC) literature has grappled with analogous dimensions and their contribution to AC ratings for over six decades (see Sakoda, 1952). While the AC context is different from that presented by SJTs, the AC literature has, nonetheless, utilised innovative analytic approaches to help inform on complex psychometric designs (e.g., Jackson, Michaelides, Dewberry, & Kim, 2016; Woehr, Meriac, & Bowler, 2012; Woehr, Putka, & Bowler, 2012).

While the psychometric structure of SJT-analogous dimensions has been extensively studied in the literature on ACs, this literature has, nonetheless, been steeped in controversy (see

Lance, 2008), since Sackett and Dreher (1982) found "virtually no support" for dimensions as "complex constructs such as leadership, decision making, or organizational acumen" (p. 409). Despite this view being contested, even by its own originators (see Kuncel & Sackett, 2014), recent estimates suggest that effects specifically concerned with dimensions explained only 2.1% of variance in AC scores *at best* (Putka & Hoffman, 2013)[1]. Much larger portions of variance in the Putka and Hoffman study were accounted for by effects analogous to general performance (33.7%) and effects specifically concerned with AC exercises (22.9%), akin to situational effects.

The Situation Perspective

In addition to dimensions, SJTs also include situational descriptors as part of their multifaceted measurement design. This aspect of the SJT design presents a key point of difference when comparing SJTs to other forms of psychometric evaluation in which situational characteristics are often not acknowledged (e.g., personality inventories). Only one known study has successfully partitioned situation- from dimension-related variance in SJTs. In this study, Westring et al. (2009, p. 45) developed an SJT such that it allowed the researchers to "partial response variance into trait and situational factors" using confirmatory factor analysis (CFA).

The intention in the Westring et al. (2009) study was to utilize an approach that represented "an improved attempt to model construct-relevant variance." Specifically, Westring et al. developed a measurement design that was amenable to analysis by CFA, in that it allowed response items to load onto both trait factors and situation factors. The educational context relevant to the Westring et al. study is different to the organisational context and the measurement design used by Westring et al. is possibly uncommon in the organisational literature. However, the Westring et al. design is innovative in that it lent itself to conventional

[1] By "at best" we refer to the person × dimension interaction when results were aggregated to the dimension-level in Putka and Hoffman (2013). The reader is also directed to Jackson et al. (2016) for a related discussion.

approaches to variance partitioning and is, thus, of methodological interest. The authors found that situation-related effects explained an average of three times more variance in SJT responses than did dimension-related effects. In specific terms, Westring et al. found that situations accounted for an average of approximately 43% of variance, whereas dimensions only accounted for an average of approximately 13%[2].

Since the findings of Westring et al. (2009) were published, little attention has been given to the role of situations in the measurement properties of SJTs, with the exception of Krumm et al. (2015), whose findings suggested that situational influences might actually have little impact on responses to SJTs. While the findings of Krumm et al. are seemingly at odds with those of Westring et al. (2009), the Westring et al. SJT was developed for an educational context, which might have led to findings that are specific to educational SJTs. However, the pervasive exercise effects observed as a matter of routine in the context of AC ratings (see Jackson et al., 2016) bear similarities to the findings reported by Westring et al. in the context of SJTs. Thus, it seems that further investigation is warranted into this key feature of the SJT design.

To summarize, the SJTs-as-methods perspective implies that any discernible structure internal to SJTs can be sidestepped in favour of a focus on correlations between SJT scores and externally-measured constructs. In contrast, the dimension perspective predicts that reliable variance in SJTs stems from dimensions that are "directly" measured by SJTs. Thus, under the dimension perspective, the majority of reliable variance internal to the structure of SJTs should be associated with dimension-related effects. Alternatively, the situation perspective implies that situations play a key role in the internal structure of SJTs. Thus, the situation perspective

[2] There were, however, very large differences in the proportions of variance explained by each of the three dimensions under scrutiny (between 1% and 23%). Nonetheless, average situational effects were still almost twice as large as the largest dimension effect found in Westring et al. (2009).

predicts that situation-related effects should explain the majority of reliable variance in the internal structure of SJTs.

An Alternative Perspective on Variance Decomposition in SJTs

A common thread relevant to both the SJTs-as-methods and dimension perspectives on the internal structure of SJTs is that results derived through EFA applied to SJTs are often found to be uninterpretable (McDaniel & Whetzel, 2005; Whetzel & McDaniel, 2009). Viewed from one perspective, this might suggest that internal SJT data are simply messy and difficult or impossible to analyse (see McDaniel, List, & Kepes, 2016). However, from another perspective, it could be the case that EFA is simply ill-suited to the analysis of SJT data. SJTs are multifaceted measures, and, as part of their measurement design, they require responses to items, which relate to situations, which, in turn, often relate to dimensions (Weekley, Ployhart, & Holtz, 2006). The ultimate purpose of EFA is to address the issue of shared variance among items (Fabrigar, Wegener, MacCallum, & Strahan, 1999). However, EFA is not equipped to handle any dependencies among items that might arise as a result of the presence of situations or dimensions within the measurement structure of SJTs (see Jackson, Putka, & Teoh, 2015) and, thus, EFA is likely to be ill-suited to SJT measurement designs.

The CFA approach taken by Westring et al. (2009) represented an advance over the EFA perspective because it allowed for separate situation and dimension factors to be specified. Also flexible in terms of the types of measurement designs that it can handle is generalizability theory (G theory, see Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; DeShon, 2002; Shavelson & Webb, 1991) which, to our knowledge, has not yet been applied to SJTs. In a single analysis, G theory can partition multiple sources of variance and, thus, can offer detailed insights into the internal measurement characteristics of SJTs. Putka and Hoffman (2013) and

Jackson et al. (2016) recently applied this approach to ACs and demonstrated that G theory can be used to summarize nuanced and informative components of reliable[3] (i.e., true score-relevant) and unreliable (i.e., true score-irrelevant) variance in a multifaceted measure.

Multiple Sources of Variance in SJTs

To introduce the sources of variance that can be decomposed in SJTs through G theory and that can potentially inform on the SJTs-as-methods, dimension, and situation perspectives on SJTs, we present an example that is of the same design used in the three samples presented later in this study. We introduce the design of the operational SJTs used in the present study here in order to facilitate an interpretation of our results presented later.

The measurement design of the SJTs in this study is configured such that all respondents (in this case job candidates, *c*) provide responses to all items (*i*). These items are contained within each of a number of situations (*s*), such that each situation is associated with its own distinct set of items. In this context, items reflect different response options for a specific scenario. Groups of situations are, in turn, categorized into specific dimensions (*d*), such that a subset of situations in the SJT are relevant to dimension 1, a different subset of situations are relevant to dimension 2, and so on (see Appendix Figure A1 for a graphical example of a dimension configured in this way). This design is one in which candidates are said to be crossed with (meaning that they complete all) items, which are nested in (meaning sub-grouped into) situations. Situations are, in turn, nested in dimensions (i.e., $c \times i{:}s{:}d$, where the multiplication symbol, ×, implies a crossed effect and the colon, :, implies a level of nesting). With a $c \times i{:}s{:}d$ design, it is possible to partition seven separate effects, four of which are relevant to observed

[3] We adopt terminology from Putka and Hoffman (2013) here, noting that this terminology is not widely applied in the more general literature on G theory (e.g., Cronbach & Shavelson, 2004; Shavelson & Webb, 2005).

scores in SJTs[4]. These four observed-score-relevant effects include candidate main effects, candidate-by-dimension interactions, candidate-by-situation (in dimension) interactions, and highest-order + residual error effects. All four of these effects are described below.

SJT-specific candidate main effects. In SJTs, candidate (or person) main effects (σ_c^2) imply that some candidates generally make "better" judgments than others regardless of the dimension, situation, or response item involved. However, this effect not only summarizes a unidimensional general judgment factor (e.g., Schmitt & Chan, 2006), but also reflects covariation between psychological constructs underlying any dimensions involved in the assessment as well as covariation between situational effects, if such effects have substantive psychological meaning (Meyer, Dalal, & Hermida, 2010; Putka & Hoffman, 2013; Woehr, Putka, et al., 2012). Candidate main effects, with respect to SJTs, are analogous to a general judgment factor, but should not be confused with general mental ability/*g* (e.g., Gonzalez-Mulé, Mount, & Oh, 2014) or with a dominant general factor as generated through principal components analysis (PCA, see Jackson et al., 2015; Lance & Jackson, 2015). Also, candidate main effects in the context of SJTs are different from candidate main effects identified in the AC literature (Lance, Foster, Nemeth, Gentry, & Drollinger, 2007; Putka & Hoffman, 2013). This is because, in SJTs, candidate main effects are concerned with judgments relating to hypothetical situations. Conversely, in ACs, such effects are concerned with behavioural responses.

Candidate-by-dimension interactions. Candidate-by-dimension interactions (σ_{cd}^2) imply that some candidates score higher on some dimensions than on others, regardless of the situation or item involved. From a covariance perspective, candidate-by-dimension interactions

[4] Seven separate effects can be partitioned with a *c* × *i*:*s*:*d* design , all of which are acknowledged in this study. However, we focus on the four effects that are relevant to observed (i.e., between-participant) scores because the remaining three effects are irrelevant to between-participant comparisons and are, therefore, irrelevant to many or most employment decisions.

reflect between-candidate variance that is specific to a given dimension and not variance shared with other dimensions or variance shared with general judgment (Putka & Hoffman, 2013). This variance component is analogous to the dimension factors typically estimated using CFA (Woehr, Putka, et al., 2012) and is, thus, analogous to the CFA-based dimension effects estimated by Westring et al. (2009). If the dimension perspective holds true, then relatively large candidate-by-dimension interactions should be evident in SJTs scores.

Candidate-by-situation (in dimension) interactions. Candidate-by-situation (in dimension) interactions ($\sigma^2_{cs:d}$) imply that some candidates score higher on some situations (nested in dimensions) than on others, regardless of the response items involved. From a covariance perspective, candidate-by-situation (in dimension) interactions reflect between-candidate variance that is specific to a given situation (nested in dimensions) and not variance shared with other situations (nested in dimensions) or variance shared with general judgment. This effect is analogous to CFA-based situation effects (e.g., Westring et al., 2009; Woehr, Putka, et al., 2012). If the situation perspective holds true, then relatively large candidate-by-situation (in dimension) interactions should be evident in SJT scores.

Highest-order + residual error effects. Highest-order effects ($\sigma^2_{ci:s:d,e}$) imply that some candidates score higher on some items nested in some situation-dimension combinations than on other situation-dimension combinations. The interpretation of $\sigma^2_{ci:s:d,e}$ is specific to a given item-dimension-situation combination and is similar to the uniqueness term estimated using CFA in that it confounds several different sources of systematic variance with random residual error. While the other three SJT-related effects described above could, potentially, constitute components of reliable variance, the highest-order effect here always constitutes unreliable variance because of its associated residual error.

Summary of SJT-Related Effects and Implications for SJT Variance Decomposition

If a relatively high proportion of observed SJT variance is due to candidate main effects, then this would imply the prevalence of an effect analogous to a general judgment factor. A relevant analogue (i.e., an SJT candidate main effect) has not yet been separated from other effects in SJTs. Doing so would help to clarify the role of candidate main effects in this context. If the dimension perspective holds true, then proportionately large candidate-by-dimension effects would be evident, indicating the analogue of dimension effects and highlighting the importance of specific dimensions in SJTs. Conversely, if the situation perspective holds true, then proportionately large candidate-by-situation (in dimension) effects would be evident, indicating the analogue of situational effects and highlighting the role of situations in SJTs.

Alternative Levels of Aggregation and Generalizability

Before interpreting the different effects involved in SJTs, it is first necessary to identify whether different approaches to aggregating SJT responses are worthy of interpretation with respect to reliability. Reliability also needs to be assessed against different types of generalization that are of interest to the researcher (Brennan, 2001; Cronbach et al., 1972). A consideration of generalizability determines which components of variance will be classified as contributing to reliable versus unreliable variance. For example, an SJT developer may wish to change the response items and situations in their test whilst retaining their existing dimensions. In this case, the developer would be interested in whether the reliability of the SJT is likely to generalize across different sets of items and situations. Under such circumstances, reliability would be estimated such that effects concerned with items and situations are treated as contributing to unreliable variance (see Appendix Table A3). In this study, when discussing

"reliability", we consider reliability with respect to generalizing across (a) different items or (b) different items and situations.

Equipped with information relating to the reliability of SJTs, it is possible to compare reliability outcomes as they pertain to different approaches to aggregating SJT responses. In theory, SJT responses could be aggregated to the level of summary scores relating to situations, dimensions, or across both situations *and* dimensions to an overall level. Despite aggregation being raised as an important consideration in the wider multifaceted measurement literature (Kuncel & Sackett, 2014), there is currently no known research on the impact of different aggregation levels on reliability outcomes for SJTs. This leads to our first research question:

> *Research Question 1:* Does aggregation to situations, dimensions, and/or to the overall-level lead to the most reliable outcomes for SJT scores?

Where Are the Source(s) of Reliable Variance in SJTs?

Research Question 1 is a necessary precursor to our second and main research question. Upon identifying aggregation level(s) fit for interpretation, we move to an analysis of reliable and unreliable sources of variance in SJTs, with the aim of contributing to an understanding of the internal measurement properties of SJTs. The rationale here is to produce a variance profile for SJTs that will inform on the SJTs-as-methods, dimension, and situation perspectives on SJTs. If the SJTs-as-methods perspective predicts no clear psychometric structure for SJTs, as suggested in previous factor analytic results, then the obtained variance profile should reveal no clear, interpretable pattern. This would imply that SJTs can only be treated as methods. If the dimension perspective holds true, then dimension-related effects should show prominence over

other effects. If the situation perspective holds true, then situation-related effects should prevail. Yet another possibility is that SJT-specific candidate main effects will prevail, for which there is no clearly-aligned proposition in the literature. This leads to our second research question, which focuses on potentially reliable sources of variance in SJTs:

> *Research Question 2:* Do candidate main effects, dimension-related effects, or situation-related effects contribute relatively more reliable variance to SJTs?

Method

Participants

Data were collected from three independent samples of participants. Each participant group provided responses to one of three operational SJTs, which were used as part of selection processes for three different types of job role. Sample 1 comprised 2,320 applicants for customer service positions in the leisure industry. Sample 2 comprised 989 applicants for graduate roles in a central government department. Sample 3 comprised 7,934 applicants for public service positions within local government agencies. Demographic characteristics by sample are provided in Table 1.

Materials and Procedure

The SJTs used in each sample differed by content and by test construction approach but all followed the same measurement configuration. This configuration followed a process whereby all candidates responded to all items. Items were nested in specific situations. Situations were, in turn, nested in dimensions (see Appendix Table A1 for the definition of each dimension used). Thus the SJTs used in all three samples followed a $c \times i{:}s{:}d$ configuration (see

Appendix Figure A1 for an example dimension). Each SJT was used as part of an online screening process and named candidates were invited to complete the SJT as a one-time assessment on an un-proctored basis. This conforms to the controlled mode of administration defined by the International Test Commission (2006). The hierarchical design used in this study was guided by the course of action set out in Weekley et al. (2006) and follows a general format relevant to that where item stems are associated with specific (i.e., nested) response options (for examples, see Guenole, Chernyshenko, Stark, & Drasgow, 2015; Stemig, Sackett, & Lievens, 2015).

Overview of SJT development. The SJTs in this study were developed with input from groups of subject matter experts (SMEs) who were line managers and/or high-performing job incumbents involved in the roles under scrutiny. Workshops or interviews were conducted with SMEs with the aim of generating critical incidents (see Motowidlo et al., 2009) relating to the dimensions identified in the Appendix (Table A1) and this process was repeated for each SJT. Response options (items) were generated and theoretically matched to each dimension listed in Table A1 by the psychologist and SME panel, whose decisions were guided by job-relevant information gleaned through critical incidents (akin to the course of action described in Motowidlo, Hanson, & Crafts, 1997). Analogous approaches are routinely used in the SJT literature (e.g., Christian et al., 2010; Krumm et al., 2015; Weekley et al., 2006) as they are in other applied contexts in the organisational literature (e.g., Bartram, 2005; International Taskforce on Assessment Center Guidelines, 2015). Based on SME input and, depending on the SJT in question (see below), a set of trial test situations were developed by a team of psychologists, which were, in turn, reviewed by SMEs. SME responses were then used to

establish a scoring key for the SJT. Incumbent responses were used to assess the difficulty level of each item-stem and related response options (items).

The SJTs used in Samples 1 and 3 consisted of 20 situations and four dimensions. The SJT used in Sample 2 consisted of 20 situations and five dimensions. In each SJT, situations provided a frame for an incident that a candidate could hypothetically face on the job (see Appendix Table A2 for example item-stems and response options). For each situation, candidates were required to rate the effectiveness of four possible response options (items) on a 5-point scale, where 1 = counter-productive, 2 = ineffective, 3 = slightly effective, 4 = effective, and 5 = very effective[5]. Each response option was scored using the *consensus weighting method* based on the approach specified in Chan and Schmitt (1997). Using this approach, scores are assigned to each point on a rating scale for a specific response based on expert consensus. A score of 2 is assigned if the rating point was endorsed by 50% or more experts, a score of 1 is assigned if between 25% and 49% of experts endorsed the rating point, and a score of 0 is assigned if less than 25% of experts endorsed the rating point. Use of the same scoring approach across samples permitted a degree of control over the influence of scoring type.

Data Analysis

We used the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015) to fit linear random effects models to data sets from Samples 1 through 3. Linear random effects models are similar in concept to the random effects analysis of variance (ANOVA) models traditionally used to estimate variance components in G theory (Brennan, 2001; Cronbach et al., 1972). However, they differ in that linear random effects models directly estimate variance components using a

[5] This rating scale did not provide an equal number of ineffective versus effective anchors. However, given the operational nature of this SJT, some gains in ecological validity were at the cost of experimental control. Also, the fact that multiple anchors were present here offered a potential advantage over 2-point scales used in some other studies (e.g., Motowidlo, Hooper, & Jackson, 2006) because, more generally, gains in reliability have been found when > 3 anchors are present (Li-Jen, 2004).

variety of different estimators including restricted maximum likelihood (REML) procedures (Searle, Casella, & McCulloch, 2006). In G theory, it is common to treat effects as random because conditions of measurement are often considered to be exchangeable with a wider universe of conditions that could be used for the same or a similar purpose (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991; Shavelson & Webb, 2005). For example, an alternative set of items to that used here could hypothetically be developed in order to achieve the same or a similar outcome.

REML-estimated variance components are preferred in the statistical literature to those generated through ANOVA models because they have all of the desirable properties of a maximum likelihood estimate (e.g. unbiased and small standard errors, see Bollen, 1989). Moreover, REML estimation is more practical in instances where a measurement design is unbalanced or where there are missing data. REML-based estimators, nonetheless, assume that the sampling distribution of a variance component can be approximated by a normal distribution (Fears & Benichou, 1996) and they can result in negative variances due to such factors as sampling error (Brennan, 2001). As a precaution, we repeated our analyses using a Bayesian estimation procedure, which relaxes the distributional assumptions associated with REML estimators (see LoPilato, Carter, & Wang, 2015). The results of the Bayesian analyses did not, however, alter any conclusions reached through the use of REML estimators[6] and are, therefore, not presented in this paper.

We also examined the effects that aggregating across situations and dimensions had upon the composition of reliable versus unreliable variance in each SJT (see Kuncel & Sackett, 2014; Putka & Hoffman, 2013). It is possible to aggregate SJT responses to the situation-, dimension-, or overall-levels (Chan & Schmitt, 2002; Weekley & Jones, 1997, 1999), all of which were

[6] Results of the Bayesian analyses are available from the second author upon request.

incorporated into our analyses. Based on formulae presented in the G theory-related literature (Brennan, 2001; Putka & Hoffman, 2013, 2014; Putka & Sackett, 2010; Shavelson & Webb, 1991), we used estimated variance components to compute Generalizability coefficients (G coefficients, which are reliability estimates) across samples for each level of aggregation. G coefficients were used here primarily to provide guidance on the most reliable level of aggregation (i.e., for Research Question 1). Two types of generalization were pertinent to the measurement configuration in this study, including generalization to different items (G_i) and different items and situations ($G_{i,s}$). Unlike coefficient alpha (or G_i for that matter), $G_{i,s}$ accounts for dependencies among items as a function of items being nested in situations, which makes it well suited to the SJT measurement design. Formulae for the G coefficients used in this study can be found in Appendix Table A3. Notwithstanding the mathematical differences between the two indices, G_i and $G_{i,s}$, returned similar results in this study[7].

Results

Table 2 displays summary statistics and correlations between dimensions for each sample. Across all three samples, correlations between dimensions (minimum average correlation = .18, maximum = .35) tended to be higher than correlations between situations (minimum average correlation = .05, maximum = .14). Coefficients alpha reflecting items within each dimension ranged from .19 to .58 and overall alphas based on all items ranged from .53 to .74.

Tables 3 through 5 summarize the results of the random effects models used in this study. All three of these tables are structured in the same manner with percentages of variance explained for random effects relating to item responses and to scores aggregated to situation-, dimension-, and overall-levels. From this point, our focus shifts to the interpretation of variance

[7] This is most likely due to the fact that situational influences were minor in all three studies.

components at the aggregate levels. Recent research on multifaceted measures has emphasised that substantively-relevant, alternative aggregation levels cannot be addressed at the item-level and that a focus on the item-level could lead to misinterpretation (e.g., Putka & Hoffman, 2013) because the item-level is potentially affected by "large amounts of specific variance and random error variance" (Kuncel & Sackett, 2014, p. 39). Aggregated levels also present a potentially more realistic picture of variance decomposition because operational SJTs are typically aggregated in some manner (e.g., Weekley et al., 2006). At the aggregate levels, only variance components relating to between-candidate variance are included in our analyses because only between-candidate sources of variance are relevant to comparisons among job applicants (Brennan, 2001; Putka & Hoffman, 2013). Reliability estimates are shown in Tables 3 through 5 for generalization to different (a) items and (b) items and situations.

It is clear from Tables 3 through 5 that, when comparing reliability estimates (G_i and $G_{i,s}$) across different aggregation levels, the overall-level of aggregation (with reliability estimates ranging from .54 in Sample 3 to .75 in Sample 1) was the only level that was worthy of consideration, regardless of the sample involved. Aggregation to the situation- and dimension-levels resulted in reliability outcomes that were too low to warrant further attention. Thus, with reference to Research Question 1, the only approach to aggregation worthy of consideration from a reliability perspective was that at the overall-level (i.e., across both situations *and* dimensions).

We turn now to our main Research Question 2, which focuses on the relative contribution of different sources of variance to reliable observed variance in SJTs. Given our results above with respect to aggregation, we focus solely on the overall-level. The overall-level columns of Tables 3 through 5 show that, across all three samples, SJT-specific candidate main effects (σ_c^2) clearly represented the strongest contributor to reliable SJT variance (Sample 1 = 67.35%;

Sample 2 = 63.15%; Sample 3 = 47.67%). The proportion of reliable observed variance attributable to candidate main effects (see the overall-level column in Tables 3 through 5) vastly overshadowed the relatively small contributions of dimension- (σ^2_{cd}, ranging between 0.29% in Sample 2 and 5.66% in Sample 1) and situation- ($\sigma^2_{cs:d}$, ranging between 1.91% in Sample 1 and 2.56% in Sample 2) related effects.

Discussion

Despite being frequently and successfully used to predict performance (Christian et al., 2010; McDaniel et al., 2001; McDaniel et al., 2007; Murphy & Shiarella, 1997), there is no agreement on what SJTs measure internally. Three different perspectives have emerged on the internal properties of SJTs in the contemporary literature. The first, SJTs-as-methods, perspective implies that SJTs are methods that do not lend themselves towards psychometric structure. Under this view, any SJT-measured constructs should be inferred from relationships between SJT scores and constructs (e.g., *g* and personality) as measured by external instruments (McDaniel & Whetzel, 2005; Whetzel & McDaniel, 2009). The second, dimension, perspective holds that SJT-measured constructs are manifest in the dimensions that are (or should be) assigned to SJTs by design (e.g., leadership and teamwork skills, see Christian et al., 2010). Under this view, dimension-related effects should prevail. The third, situation, perspective is informed by the results of Westring et al. (2009), who found a substantial portion of variance explained by situation-specific factors, implying that situations represent a major component of SJT-measured "constructs" of interest. Under this view, situation-related effects should prevail. Our results suggest that none of these three perspectives is likely to be (unconditionally) precise.

Across three large samples, we used G theory-based methods (Cronbach et al., 1972; DeShon, 2002) to decompose multiple sources of observed score variance in SJTs. Our initial

goal was to establish which level(s) of aggregation (i.e., situation-, dimension-, or overall-level) warranted consideration based on a comparison of their respective reliabilities (see Research Question 1). The issue regarding the level of aggregation that is most appropriate for the purposes of reliability estimation and interpretation has recently come to light in the broader multifaceted measurement literature. Alternative aggregation levels have been presented in this literature (e.g., Putka & Hoffman, 2013) as they are in the current paper. The item-level of analysis cannot address alternative levels of aggregation. Moreover, Kuncel and Sackett (2014, p. 39) stated that "individual items contain large amounts of specific variance and random error variance; that is why multiple items are aggregated into a scale" and suggested interpreting variance decomposition at aggregate levels, as have other researchers of multifaceted measures (e.g., Jackson et al., 2016; LoPilato et al., 2015; Putka & Hoffman, 2013, 2014). The same issues about aggregation are relevant to SJTs because, in practice, SJTs are typically aggregated in some manner (e.g., Weekley et al., 2006). Our results consistently suggested that the overall-level of aggregation was the only level that warranted consideration. Reliability estimates at the situation- and dimension-levels of aggregation were all unacceptably low (see Tables 3 through 5).

In light of the above findings, to address our second and main research question (see Research Question 2) we proceeded to interpret variance source profiles at the overall-level of aggregation. The effects of seven distinct sources of variance were decomposed from the SJT ratings. Of these seven, three sources were relevant to reliable between-candidate variance in SJTs: (a) SJT-specific candidate main effects (σ_c^2), which are analogous to a general judgment factor for SJTs, (b) candidate × dimension interactions (σ_{cd}^2), which are analogous to dimension-

related effects, and (c) candidate × situation (nested in dimension) interactions ($\sigma^2_{cs:d}$), which are analogous to situation-related effects.

Two findings relating to our second research question were consistently apparent across all three samples. Firstly, SJT-specific candidate main effects constituted by far the largest source of reliable SJT variance (explaining between 47.67% and 67.35% of variance). Secondly, in absolute terms, dimension-related effects (between 0.29% and 5.66%) and situation-related effects (between 1.91% and 2.56%) were consistently small and were also small relative to candidate main effects. To put this comparison into perspective, candidate main effects were at least 13 times larger than dimension-related effects and at least 19 times larger than situation-related effects.

These findings raise two questions. Firstly, what are SJT-specific candidate main effects and how do they fit in with psychological theory? Secondly, why do dimension- and situation-related effects explain such little variance in SJT responses? On the first question, candidate main effects in SJTs summarize a general judgment factor and the covariance between any underlying dimension and situation factors, if such factors hold psychological meaning (Woehr, Putka, et al., 2012). Thus, a candidate main effect should neither be confused with *g* in the Spearman tradition (e.g., Gonzalez-Mulé et al., 2014) nor with the first unrotated factor in a PCA (see Jackson et al., 2015; Lance & Jackson, 2015). Rather, SJT-specific main effects imply that, regardless of specific situations, dimensions, or response items; some people consistently score higher than others on judging an "appropriate" course of action when faced with a situational dilemma. Moreover, candidate main effects also subsume any psychologically meaningful covariance among dimensions and situations.

McDaniel and Whetzel (2005, p. 523) stated that SJTs are best thought of as "methods" that relate to externally-measured psychological constructs: a position driven, perhaps, by the finding that the "construct heterogeneity of SJT items makes coherent factor analysis results difficult" (McDaniel & Whetzel, 2005, p. 519). This view implies that any internal structure for SJTs can essentially be circumvented in favour of investigating correlations between SJT scores and externally-measured constructs. Our findings suggest that there is an intermediary step missing from this proposition, in that the structure of reliable SJT score variance appears to primarily reflect SJT-specific candidate main effects. The potential exists for candidate main effects to be isolated from other sources of variance (including unreliable sources) and then related to external measures. Thus, the relationship might not be between the SJT *method* and the likes of *g* and personality, but, rather, the relationship might be between *SJT-specific main effects* and *g* and personality. This distinction is important because the former proposal implies no discernible internal structure for SJTs. However, the latter proposal offers a, currently missing, psychometric structure for SJTs and a possibly more precise direction for future exploration.

In terms of why dimension- and situation-related effects explained relatively little variance in our study, we turn, initially, to SJT dimensions. Christian et al. (2010) urged researchers to pay more attention to dimensions in SJTs. However, our findings are at odds with this proposal because our dimension-related effects were trivial in absolute terms and, in relative terms, were dwarfed by SJT-specific candidate main effects. Relevant to this point, Arthur and Villado (2008) made the distinction between espoused and actual constructs. In the former, a label is ascribed to a set of expectations around a set of behavioural descriptors or responses. In the latter, there is empirical evidence to support the measurement of such constructs. We argue

further that in order to qualify as a measure of *actual* constructs, evidence needs to be provided relating to the *internal* measurement characteristics of an instrument. Dimensions might represent intuitive hypothetical categories for subsets of job-relevant behaviours. The inclusion of dimensions in SJTs is likely to be helpful from the perspective that dimensions promote a consideration of the job-relatedness of SJT content by way of their links to competency modelling processes (see Schippmann et al., 2000). However, evidence in support of their internal structure in SJTs or even in other contexts (e.g., in ACs, see Lance, 2008) presents a topic for debate. Counter to the proposition of Christian et al., our findings suggest that SJT research should be directed towards unravelling the multifaceted nature of SJT-specific candidate main effects, rather than dimensions.

Our findings, with respect to situation-related effects, do not align with those of Westring et al. (2009), who found that situational variance accounted for an average of 43% of variance in their SJT (in contrast to our findings for the contribution of situation-based variance at a mere ≤ 2.56%). A possible reason for this is that the Westring et al. SJT was developed to measure three traits (dimensions) in an educational context (mastery, performance approach, and performance avoid). These three dimensions might require different knowledge, skills, and abilities than those often applied in organisations (e.g., see Arthur, Day, McNelly, & Edens, 2003). Another possible reason is that Westring et al. also employed a design that is not typically used in organisational SJTs.

It is also possible that Westring et al. (2009) found more situational variance than was found in the present study because Westring et al used a less detailed variance partitioning approach than that which we used. In order to test this proposition, we re-analysed the Westring et al. dataset using a REML random effects model (see the Appendix, Table A4), consisting of

five between-candidate effects (as opposed to two effects in the original study). Regardless of aggregation level, we found that the analogue of dimension effects (σ^2_{cd}) and a three-way interaction involving candidate effects, dimensions, and situations (σ^2_{cds}) explained relatively large proportions of reliable variance. Counter to their original findings based on CFA, we found situational influences to be very small (<.01%), which aligns more closely to the finding of Krumm et al. (2015) as well as to our findings. Note that it has been argued elsewhere that σ^2_{cds}-analogous effects represent a type of situational effect (Jackson et al., 2016).

The Westring et al. design is unlike designs often used in the organisational SJT literature, which is unsurprising given its application in an educational setting. Therefore, it is difficult to generalise the results of our re-analysis of the Westring et al. data to organisational SJTs, including to our own SJTs. Nonetheless, in comparison to our results, this reanalysis of the Westring et al. data does suggest that different SJT designs can potentially result in very different internal structures. If the desire is to measure dimensions, then perhaps researchers could explore a design akin to that used by Westring et al. in an organisational setting. However, our results also suggest that the internal structure of a given SJT design should not be assumed and that G theory offers a flexible approach to exploring the internal structure of an SJT, whatever its design.

Limitations and Future Directions

The nested design in the present study restricted the number of individual variance components that we could estimate. To allow for a more comprehensive variance decomposition, future studies could design SJTs so as to allow for the estimation of a three-way candidate × dimension × situation (σ^2_{cds}) variance component as separate from a two-way candidate × situation (σ^2_{cs}) component. Similarly, a limitation of the G theory methods employed

here is that they do not explicitly model intercorrelations among dimensions or situations, as is often achieved using CFA approaches (Brennan, 2001; Shavelson & Webb, 1991). Analogues of the parameters estimated in G theory can also be estimated using CFA and CFA can be used to provide estimates from specific dimensions or situations, whereas G theory provides average estimates across all dimensions and/or situations (Le, Schmidt, & Putka, 2009; Le, Schmidt., Harter, & Lauver, 2010). However, CFA approaches can suffer from admissibility issues (e.g., Woehr, Putka, et al., 2012) and G theory offers a straightforward approach to handling aggregation, relative versus absolute error, and ill-structured measurement designs (DeShon, 2002; Putka & Hoffman, 2013, 2014; Putka, Le, McCloy, & Diaz, 2008). We suggest that both G theory and CFA offer important perspectives on the reliability of SJTs, but that these perspectives are often complementary (see Putka & Sackett, 2010) and, at present, the G theory perspective is underrepresented in the SJT literature.

The reader should be aware that the absolute magnitude of effects considered to be reliable at the overall level of aggregation might be dependent on model specifications specific to a given SJT. We ran supplementary decision studies to assess the extent to which this affected the results in the present study. Decision studies are essentially a version of the Spearman-Brown prediction formula that is applicable to multifaceted measures (see Shavelson & Webb, 1991). The results of these analyses suggested that (a) increasing dimensions (and nested situations and items) led to increases in the magnitude of candidate main effects (σ_c^2), but, in support of our main conclusions, (b) both σ_{cd}^2 and $\sigma_{cs:d}^2$ were, generally, consistently of a low magnitude, and (c) that σ^2 remained generally high relative to σ_{cd}^2 and $\sigma_{cs:d}^2$, regardless of model specifications, particularly when the number of dimensions involved was ≥ 2. In many or most

applied scenarios, it seems likely that the number of dimensions in a given SJT would exceed 2, given the relevant practices discussed in the SJT literature (e.g., Christian et al., 2010).

It is also possible to incorporate relevant substantive covariates into linear random effects models in order to examine how they relate to the different variance components (O'Neill, Goffin, & Gellatly, 2012; Putka, Ingerick, & McCloy, 2008). This approach is referred to as linear mixed effects modelling and allows researchers to examine how effects such as the candidate main effect (σ_c^2) are related to external correlates. For instance, future research could include measures of personality and cognitive ability and examine whether including these variables reduce (implying a relationship with) variance associated with the candidate main effect that was found to be prevalent in our study.

Although the present study included three separate samples, there remains a need to generalize our results over different occupations, different situations, and different dimensions. It is possible that a different, potentially less *g*-loaded, set of dimensions might result in a different variance profile. Moreover, it would also be interesting to see if the magnitudes of the variance components estimated in the present study generalize across different cultural contexts and geographical locations. In addition, a distinction is drawn in the SJT literature between knowledge (e.g., *what do you know about x?*) versus behavioural tendency (e.g., *what would you be likely to do if x occurred?*) response options (Ployhart & Ehrhart, 2003; Whetzel & McDaniel, 2009). The SJTs in the present study used a knowledge-based approach and, therefore, our findings are only likely to be relevant to SJTs incorporating the same type of response. Further research will be necessary for generalization to behavioural tendency response options.

Concluding Comments

Our findings suggest that, relative to situation- and dimension-related effects, the largest source of reliable SJT variance is represented by candidate main effects. Such effects are analogous to a general judgment factor in combination with covariance between situation and dimension factors, where such factors hold psychological meaning. In contrast to current proposals in the literature, we conclude that there is a discernible psychometric structure for reliable variance in organisational SJTs and that structure primarily pertains to candidate main effects. Moreover, we conclude that SJT scores should not be presumed to reflect structures that resemble discrete dimension- or situation-related constructs.

References

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-154. doi: 10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., Jr., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99*, 535-545. doi: 10.1037/a0035788

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435-442. doi: Doi 10.1037/0021-9010.93.2.435

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185-1203. doi: 10.1037/0021-9010.90.6.1185

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48. doi: 10.18637/jss.v067.i01

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14*, 223-235. doi: 10.1111/j.1468-2389.2006.00345.x

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

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

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment, 20*, 333-346. doi: 10.1111/j.1468-2389.2012.00604.x

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143-159. doi: 10.1037//0021-9010.82.1.143

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*, 233-254. doi: 10.1207/S15327043HUP1503_01

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83-117. doi: 10.1111/j.1744-6570.2009.01163.x

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., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391-218. doi: 10.1177/0013164404266386

DeShon, R. P. (2002). Generalizability theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 189-220). San Francisco, CA: Jossey-Bass.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299. doi: 10.1037/1082-989X.4.3.272

Fears, T. R., & Benichou, J. (1996). A reminder of the fallibility of the Wald statistic. *American Statistician, 50*, 226. doi: 10.1080/00031305.1996.10474384

Gonzalez-Mulé, E., Mount, M. K., & Oh, I. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology, 99*, 1222-1243. doi: 10.1037/a0037547

Guenole, N., Chernyshenko, O., Stark, S., & Drasgow, F. (2015). Are predictions based on situational judgement tests precise enough for feedback in leadership development? *European Journal of Work and Organizational Psychology, 24*, 433-443. doi: 10.1080/1359432X.2014.926890

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

Jackson, D. J. R., Michaelides, M., Dewberry, C., & Kim, Y. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology*. doi: Advance online publication. dx.doi.org/10.1037/apl0000102

Jackson, D. J. R., Putka, D. J., & Teoh, K. R. H. (2015). The first principal component of multifaceted variables: It's more than a g thing. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*, 446-452. doi: 10.1017/iop.2015.61

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology, 100*, 399-416.

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

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

Lance, C. E., & Jackson, D. J. R. (2015). Seek and ye shall find. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*, 452-463. doi: 10.1017/iop.2015.62

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods, 12*, 165-200. doi: Doi 10.1177/1094428107302900

Le, H., Schmidt., F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes, 112*, 112-125. doi: 10.1016/j.obhdp.2010.02.003

Li-Jen, W. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*, 986-972. doi: 10.1177/0013164404268674

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90*, 442-452. doi: Doi 10.1037/0021-9010.90.3.442

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

McDaniel, M. A., Bruhn Finnegan, E., Morgeson, F. P., & Campion, M. A. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730-740. doi: 10.1037//0021-9010.86.4.730

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63-91. doi: 10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., List, S. K., & Kepes, S. (2016). The "hot mess" of situational judgment test construct validity and other issues. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 9*, 47-51. doi: 10.1017/iop.2015.115

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103-113.

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33*, 515-525. doi: 10.1016/j.intell.2005.02.001

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management, 36*, 121-140. doi: 10.1177/0149206309349309

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology, 24*, 281-288. doi: 10.1007/s10869-009-9106-4

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 241-260). Palo Alto, CA: Consulting Psychologists Press.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A theoretical basis for situational judgment tests. In J. A. Weekley, R. E. Ployhart, J. A. Weekley, & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application.* (pp. 57-81). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology, 50*, 823-854. doi: 10.1111/j.1744-6570.1997.tb01484.x

O'Neill, T. A., Goffin, R. D., & Gellatly, I. R. (2012). The use of random coefficient modeling for understanding and predicting job performance ratings: An application with field data. *Organizational Research Methods, 15*, 436-462. doi: 10.1177/1094428112438699

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11*, 1-16. doi: 10.1111/1468-2389.00222

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., Ingerick, M., & McCloy, R. A. (2008). Integrating traditional perspectives on error in ratings: Capitalizing on advances in mixed-effects modeling. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 167-173. doi: 10.1111/j.1754-9434.2008.00032.x

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.

Rockstuhl, T., Ang, S., Ng, K. Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology, 100*, 464-480. doi: 10.1037/a0038098

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

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

Schippmann, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh, B., . . . Sanchez, J. I. (2000). The practice of competency modeling. *Personnel Psychology, 53*, 703-740.

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science, 2*, 8-9. doi: 10.1111/1467-8721.ep10770456

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement and application* (pp. 135-155). San Francisco: Jossey-Bass.

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

Sharma, S., Gangopadhyay, M., Austin, E., & Mandal, M. K. (2013). Development and validation of a situational judgment test of emotional intelligence. *International Journal of Selection and Assessment, 21*, 57-73. doi: 10.1111/ijsa.12017

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

Shavelson, R. J., & Webb, N. M. (2005). Generalizability theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Complementary methods for research in education* (3rd ed., pp. 599-612). Washington, DC: AERA.

Stemig, M. S., Sackett, P. R., & Lievens, F. (2015). Effects of organizationally endorsed coaching on performance and validity of situational judgment tests. *International Journal of Selection and Assessment, 23*, 174-181. doi: 10.1111/ijsa.12105

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., . . . Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. Cambridge: Cambridge University Press.

The International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*, 143-171. doi: 10.1207/s15327574ijt0602_4

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49*, 436-458. doi: 10.1037/0022-3514.49.2.436

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25-49. doi: 10.1111/j.1744-6570.1997.tb00899.x

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*, 679-700. doi: 10.1111/j.1744-6570.1999.tb00176.x

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement and application* (pp. 157-182). San Francisco: Jossey-Bass.

Westring, A. J. F., Oswald, F. L., Schmitt, N., Drzakowski, S., Imus, A., Kim, B., & Shivpuri, S. (2009). Estimating trait and situational variance in a situational judgment test. *Human Performance, 22*, 44-63. doi: 10.1080/08959280802540999

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*, 188-202. doi: 10.1016/j.hrmr.2009.03.007

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

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-Theory methods for modeling multitrait–multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods, 15*, 134-161. doi: 10.1177/1094428111408616

Table 1
Demographic Characteristics by Sample

Characteristic	Sample		
	1(%)	2(%)	3(%)
Gender			
Male	48.7	65.1	68.4
Female	49.7	34.9	30.4
Non-response	1.6	0.0	1.2
Ethnicity			
White British	80.6	56.2	91.8
Other White	13.1	8.4	2.3
Asian	2.2	18.9	1.2
Other	2.2	14.4	2.3
Non-response	2.1	2.1	2.4
Age-Band			
< 25	48.0	59.2	50.7
25-40	43.2	39.8	44.1
> 40	8.8	1.1	48.3
Non-response	0.1	0.0	1.1

Table 2
Summary Statistics and Correlations by Sample

Sample/Dimension	M	SD	1.	2.	3.	4.	5.
Sample 1 – Customer Service (N = 2,320)							
1. Convincing Others	20.24	3.33	**.47**				
2. Dealing with Challenging Customers	19.41	3.47	.39	**.58**			
3. Delivering Quality Service	21.30	4.47	.34	.40	**.51**		
4. Understanding Customer Needs	20.01	3.55	.31	.31	.32	**.45**	
Sample 2 – Central Government (N = 989)							
1. Achieving Results	17.78	3.17	**.33**				
2. Analytical Thinking	17.66	3.77	.14	**.35**			
3. Communicating & Influencing	16.54	3.40	.21	.17	**.31**		
4. Planning and Organising	14.69	3.47	.31	.22	.20	**.32**	
5. Relationship Building	18.42	4.27	.25	.11	.28	.23	**.42**
Sample 3 – Public Service (N = 7,934)							
1. Problem Solving and Decision Making	20.80	3.07	**.19**				
2. Leadership	17.41	3.48	.15	**.22**			
3. Planning and Organising	16.84	3.61	.13	.15	**.39**		
4. Strategic and Organisational Awareness	20.00	3.61	.16	.20	.31	**.27**	

Note. Mean correlation between situations in Sample 1 = .05 (*SD* = .04, overall coefficient alpha = .74); Sample 2 = .09 (*SD* = .03, overall coefficient alpha = .65); Sample 3 = .14 (*SD* = .09, overall coefficient alpha = .53). All correlations were significant at the $p < .05$ level. Coefficients alpha, estimated based on items within each dimension, appear bolded in the diagonal.

Table 3

Variance Estimates for Sample 1- Customer Service

	Item-Level				Situation-Level			Dimension-Level			Overall-Level		
VC	VE	% Total	% BC	G	Formula	% BC	G	Formula	% BC	G	Formula	% BC	G
BCSV													
σ_c^2	.01	2.25	3.15	-	σ_c^2	10.69	-	σ_c^2	34.02	-	σ_c^2	67.35	-
σ_{cd}^2	< .01	0.76	1.06	-	σ_{cd}^2	3.59	-	σ_{cd}^2	11.44	-	σ_{cd}^2/n_d	5.66	-
$\sigma_{cs:d}^2$	.01	1.28	1.79	-	$\sigma_{cs:d}^2$	6.07	-	$\sigma_{cs:d}^2/n_{s:d}$	3.87	-	$\sigma_{cs:d}^2/n_d n_{s:d}$	1.91	-
$\sigma_{ci:s:d,e}^2$	.35	67.08	93.99	-	$\sigma_{ci:s:d,e}^2/n_{i:s}$	79.64	-	$\sigma_{ci:s:d,e}^2/n_{i:s}n_{s:d}$	50.68	-	$\sigma_{ci:s:d,e}^2/n_i$	25.08	-
OSV													
σ_d^2	< .01	<0.01	-	-	-	-	-	-	-	-	-	-	-
$\sigma_{s:d}^2$	< .01	<0.01	-	-	-	-	-	-	-	-	-	-	-
$\sigma_{i:s:d}^2$	.15	28.63	-	-	-	-	-	-	-	-	-	-	-
G_i	-	-	-	.06	-	-	.20	-	-	.49	-	-	.75
$G_{i,s}$	-	-	-	.04	-	-	.14	-	-	.45	-	-	.73

Note. VC = variance component; VE = variance estimate; item-level = non-aggregated item responses; situation-, dimension- and overall-level refer to each, respective score aggregate; % Total = percent of total variance explained by each effect; % BC = percent of variance explained for each between-candidate effect; G = Generalizability coefficient; G_i, $G_{i,s}$ = G for generalization to different items, and items and situations, respectively; BCSV = between-candidate sources of variance; OSV = other sources of variance; *c* = candidate; *d* = dimension; *s*= situation; *i* = item. Number of dimensions (n_d) = 4; number of situations nested within a dimension ($n_{s:d}$) = 5; number of items nested within situations ($n_{i:s}$) = 4; total number of items (n_i) = 80.

Table 4

Variance Estimates for Sample 2 – Central Government

	Item-Level				Situation-Level			Dimension-Level			Overall-Level		
VC	VE	% Total	% BC	G	Formula	% BC	G	Formula	% BC	G	Formula	% BC	G
BCSV													
σ_c^2	.01	1.87	2.23	-	σ_c^2	7.93	-	σ_c^2	25.52	-	σ_c^2	63.15	-
σ_{cd}^2	< .01	0.04	0.05	-	σ_{cd}^2	0.18	-	σ_{cd}^2	0.58	-	σ_{cd}^2/n_d	0.29	-
$\sigma_{cs:d}^2$	.01	1.51	1.81	-	$\sigma_{cs:d}^2$	6.44	-	$\sigma_{cs:d}^2/n_{s:d}$	5.18	-	$\sigma_{cs:d}^2/n_d n_{s:d}$	2.56	-
$\sigma_{ci:s:d,e}^2$	.57	80.41	95.92	-	$\sigma_{ci:s:d,e}^2/n_{i:s}$	85.45	-	$\sigma_{ci:s:d,e}^2/n_{i:s}n_{s:d}$	68.72	-	$\sigma_{ci:s:d,e}^2/n_i$	34.00	-
OSV													
σ_d^2	< .01	0.06	-	-	-	-	-	-	-	-	-	-	-
$\sigma_{s:d}^2$	< .01	0.62	-	-	-	-	-	-	-	-	-	-	-
$\sigma_{i:s:d}^2$	.11	15.48	-	-	-	-	-	-	-	-	-	-	-
G_i	-	-	-	.04	-	-	.15	-	-	.31	-	-	.66
$G_{i,s}$	-	-	-	.02	-	-	.08	-	-	.26	-	-	.63

Note. VC = variance component; VE = variance estimate; item-level = non-aggregated item responses; situation-, dimension- and overall-level refer to each, respective score aggregate; % Total = percent of total variance explained by each effect; % BC = percent of variance explained for each between-candidate effect; G = Generalizability coefficient; G_i, $G_{i,s}$ = G for generalization to different items, and items and situations, respectively; BCSV = between-candidate sources of variance; OSV = other sources of variance; *c* = candidate; *d* = dimension; *s*= situation; *i* = item. Number of dimensions (n_d) = 5; number of situations nested within a dimension ($n_{s:d}$) = 4; number of items nested within situations ($n_{i:s}$) = 4; total number of items (n_i) = 80.

Table 5

Variance Estimates for Sample 3 – Public Service

	Item-Level				Situation-Level			Dimension-Level			Overall-Level		
VC	VE	% Total	% BC	G	Formula	% BC	G	Formula	% BC	G	Formula	% BC	G
BCSV													
σ_c^2	.01	0.93	1.25	-	σ_c^2	4.61	-	σ_c^2	18.55	-	σ_c^2	47.67	-
σ_{cd}^2	< .01	0.29	0.39	-	σ_{cd}^2	1.44	-	σ_{cd}^2	5.78	-	σ_{cd}^2/n_d	3.71	-
$\sigma_{cs:d}^2$	.01	0.96	1.30	-	$\sigma_{cs:d}^2$	4.78	-	$\sigma_{cs:d}^2/n_{s:d}$	3.85	-	$\sigma_{cs:d}^2/n_d n_{s:d}$	2.47	-
$\sigma_{ci:s:d,e}^2$	.43	72.03	97.06	-	$\sigma_{ci:s:d,e}^2/n_{i:s}$	89.18	-	$\sigma_{ci:s:d,e}^2/n_{i:s}n_{s:d}$	71.82	-	$\sigma_{ci:s:d,e}^2/n_i$	46.14	-
OSV													
σ_d^2	.00	0.00	-	-	-	-	-	-	-	-	-	-	-
$\sigma_{s:d}^2$	.04	6.73	-	-	-	-	-	-	-	-	-	-	-
$\sigma_{i:s:d}^2$	.11	19.06	-	-	-	-	-	-	-	-	-	-	-
G_i	-	-	-	.03	-	-	.12	-	-	.28	-	-	.54
$G_{i,s}$	-	-	-	.02	-	-	.06	-	-	.24	-	-	.51

Note. VC = variance component; VE = variance estimate; item-level = non-aggregated item responses; situation-, dimension- and overall-level refer to each, respective score aggregate; % Total = percent of total variance explained by each effect; % BC = percent of variance explained for each between-candidate effect; G = Generalizability coefficient; G_i, $G_{i,s}$ = G for generalization to different items, and items and situations, respectively: BCSV = between-candidate sources of variance; OSV = other sources of variance; *c* = candidate; *d* = dimension; *s*= situation; *i* = item. Number of dimensions (n_d) = 4; number of situations nested within a dimension ($n_{s:d}$) = 5; number of items nested within situations ($n_{i:s}$) = 4; total number of items (n_i) = 80.

Appendix

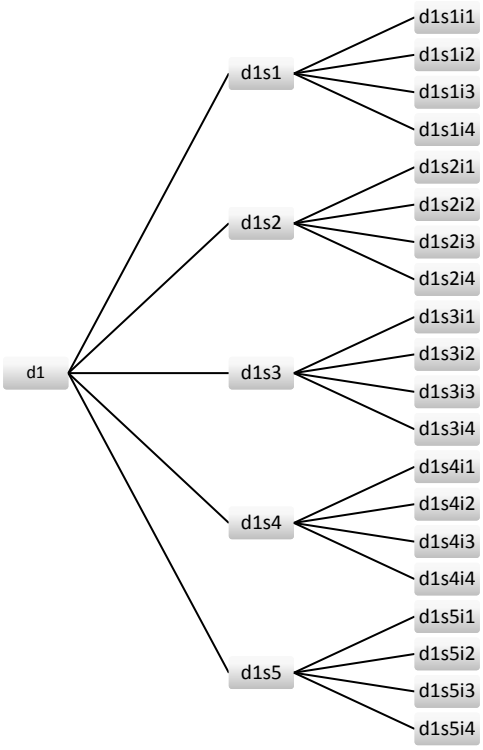

Figure A1. Diagrammatical representation of the nested structure used for the first dimension (d1) in the situational judgment test (SJT) in Sample 1. Four items (i) were nested in each of five situations (s), which were, in turn, nested in d1. Note that there were four dimensions in the Sample 1 SJT and, as such, this structure was repeated for each dimension.

Table A1
Dimension Definitions by Sample

Sample 1 – Customer Service	Definition
Convincing Others	Convincing customers of the value of a service or product.
Dealing with Challenging Customers	Dealing effectively with challenging customers, remaining calm under pressure and taking responsibility for customer complaints so that they are resolved promptly.
Delivering Quality Service	Delivering a high quality service to customers in spite of obstacles or challenges.
Understanding Customer Needs	Understanding the needs of the customer and seeking out information to provide tailored solutions.

Sample 2 – Central Government	Definition
Achieving Results	Overcoming obstacles and completing tasks to a high standard.
Analytical Thinking	Analysing data, making sound decisions and understanding the underlying cause of problems.
Communicating and Influencing	Communicating information, persuading others to own point of view or convincing them of a given course of action.
Planning and Organising	Prioritizing activities and managing time and resources to meet deadlines.
Relationship Building	Building effective working relationships with others, including dealing with sensitive issues and working as part of a team

Sample 3 – Public Service	Definition
Problem Solving and Decision Making	Analysing information rationally, evaluating alternative options and making clear, timely, and justifiable decisions based on available evidence.
Leadership	Putting self forward for more responsibility, taking control of situations and being confident in their own ability to adapt and cope with changing situations.
Planning and Organising	Taking a methodical approach, prioritizing activities, and planning their own time effectively.
Strategic and Organisational Awareness	Considering the bigger picture when making decisions and understanding how their own role fits into overall organisational objectives.

Table A2

Example Item-Stems and Response Options by Sample

Example for Sample 1 – Customer Service
Read the situation and rate each of the four actions. *You are working in a department store on a payment till. Your store is understaffed as several of your colleagues are ill. You are working as quickly as you can but a there is large queue of customers waiting to pay. A customer pushes her way to the front of the queue. She starts complaining loudly about how long she has waited and says that she will not wait any more. She is starting to annoy the other customers.* A. Tell the customer that you are sorry for the wait, explain that you are serving people as quickly as you can, and ask her to return to her place in the queue. B. Ask the customer to keep her voice down as it is annoying the other customers, and say that you will serve her in due course. C. Tell the customer that the more time she spends complaining, the longer it will take you to serve everyone. D. Serve the customer who is complaining next, so that she stops annoying the other customers in the queue.
Example for Sample 2 – Central Government
Read the situation and rate each of the four actions. *You are analysing numerical data relating to an organisational process that has been compiled from several departments. During your analysis, you find that the data from one department does not appear to match up with related data from some other departments. When you speak to the colleague who provided you with this data, he assures you that it is fully accurate and that it has been checked and double checked thoroughly.* A. Exclude the data from the analysis, as it appears that there may be some mistakes in it. B. Continue with the analysis using the data as it is, given that you have your colleague's assurance that his data is completely accurate. C. Identify possible reasons why the data might not match up and investigate these in turn to see if you can identify the problem. D. Show your colleague the specific areas where his data does not appear to match up with the rest and ask him for his opinion about what the possible reasons might be.
Example for Sample 3 – Public Service
Read the situation and rate the effectiveness of each of the four actions. *You are working in a particularly challenging environment with a number of requirements and priorities that change daily. You have noticed that one of your colleagues is not dealing particularly well with these volatile circumstances and s/he is struggling to keep up with everything that is going on. S/he is being critical of the organisation and you feel their actions are starting to have a negative impact on others in your team.* A. Ignore your colleague's comments and focus on getting your own work done. B. Speak to your colleague and ask her/him to be more aware of her/his impact on others, as s/he is having a negative impact on the team. C. Privately request that your colleague is transferred to another team as s/he clearly doesn't fit in with everyone else. D. Talk to your colleague and try to understand why s/he is struggling to deal with the changes and look for some ways to help her/him.

Table A3
Formulae for Generalizability Coefficients

Level/Generalization to…	Formula
Item-level responses	
Items	$\left[\sigma_c^2 + \sigma_{cd}^2 + \sigma_{cs:d}^2\right] / \left[\sigma_c^2 + \sigma_{cd}^2 + \sigma_{cs:d}^2 + \sigma_{ci:s:d,e}^2\right]$
Items and situations	$\left[\sigma_c^2 + \sigma_{cd}^2\right] / \left[\sigma_c^2 + \sigma_{cd}^2 + \sigma_{cs:d}^2 + \sigma_{ci:s:d,e}^2\right]$
Situation-level aggregation	
Items	$\left[\sigma_c^2 + \sigma_{cd}^2 + \sigma_{cs:d}^2\right] / \left[\sigma_c^2 + \sigma_{cd}^2 + \sigma_{cs:d}^2 + \left(\sigma_{ci:s:d,e}^2 / n_{i:s}\right)\right]$
Items and situations	$\left[\sigma_c^2 + \sigma_{cd}^2\right] / \left[\sigma_c^2 + \sigma_{cd}^2 + \sigma_{cs:d}^2 + \left(\sigma_{ci:s:d,e}^2 / n_{i:s}\right)\right]$
Dimension-level aggregation	
Items	$\left[\sigma_c^2 + \sigma_{cd}^2 + \left(\sigma_{cs:d}^2 / n_{s:d}\right)\right] / \left[\sigma_c^2 + \sigma_{cd}^2 + \left(\sigma_{cs:d}^2 / n_{s:d}\right) + \left(\sigma_{ci:s:d,e}^2 / n_{i:s} n_{s:d}\right)\right]$
Items and situations	$\left[\sigma_c^2 + \sigma_{cd}^2\right] / \left[\sigma_c^2 + \sigma_{cd}^2 + \left(\sigma_{cs:d}^2 / n_{s:d}\right) + \left(\sigma_{ci:s:d,e}^2 / n_{i:s} n_{s:d}\right)\right]$
Overall-level aggregation	
Items	$\left[\sigma_c^2 + \left(\sigma_{cd}^2 / n_d\right) + \left(\sigma_{cs:d}^2 / n_d n_{s:d}\right)\right] / \left[\sigma_c^2 + \left(\sigma_{cd}^2 / n_d\right) + \left(\sigma_{cs:d}^2 / n_d n_{s:d}\right) + \left(\sigma_{ci:s:d,e}^2 / n_i\right)\right]$
Items and situations	$\left[\sigma_c^2 + \left(\sigma_{cd}^2 / n_d\right)\right] / \left[\sigma_c^2 + \left(\sigma_{cd}^2 / n_d\right) + \left(\sigma_{cs:d}^2 / n_d n_{s:d}\right) + \left(\sigma_{ci:s:d,e}^2 / n_i\right)\right]$

Note. Item-level = non-aggregated item responses; situation-, dimension- and overall-level aggregation refer to each, respective score aggregate; *c* = candidate; *d* = dimension; *s*= situation; *i* = item.

Table A4

Between-Candidate Percentage of Variance Explained for Westring et al. (2009)

VC	Item-Level	Situation-Level		Dimension-Level		Overall-Level	
	%	Formula	%	Formula	%	Formula	%
σ_c^2	< .01	σ_c^2	< .01	σ_c^2	< .01	σ_c^2	< .01
σ_{cd}^2	31.28	σ_{cd}^2/hn_d	47.90	σ_{cd}^2	84.36	σ_{cd}^2/n_d	78.54
σ_{cs}^2	< .01	σ_{cs}^2	< .01	σ_{cs}^2/ hn_s	< .01	σ_{cs}^2/n_s	< .01
σ_{cds}^2	29.48	σ_{cds}^2/ hn_d	45.14	σ_{cds}^2/ hn_s	11.04	$\sigma_{cds}^2/n_{d\times s}$	10.09
$\sigma_{ci:ds,e}^2$	39.24	$\sigma_{ci:ds,e}^2/ hn_d hn_{i:d}$	6.95	$\sigma_{ci:ds,e}^2/hn_s hn_{i:s}$	4.59	$\sigma_{ci:ds,e}^2/n_i$	11.37

Note. VC = variance component; c = candidate, d = dimension, s = situation. Harmonic mean of dimensions (hn_d) = 2.67; harmonic mean of situations (hn_s) = 7.20; harmonic mean of items nested within dimensions ($hn_{i:d}$) = 8.64; harmonic mean of items nested within situations ($hn_{i:s}$) = 3.2; number of dimensions (n_d) = 3; number of situations (n_s) = 8; number of dimension-situation units ($n_{d\times s}$) = 22; total number of items (n_i) = 26.