Clarifying the Scope of Generalizability Theory for Multifaceted Assessment

Duncan J. R. Jackson[1], George Michaelides[2], Chris Dewberry[3], and Paul Englert[4]

[1]King's Business School, King's College London
[2]Norwich Business School, University of East Anglia
[3]Independent Scholar
[4]Department of Psychology, Nanyang Technical University, Singapore

Generalizability theory (G theory) continues to be underutilized in applied psychological research, both in New Zealand and internationally, possibly due to uncertainties about the types of questions that it can be used to address. G theory and its associated random effects model basis is often positioned as an approach limited to the study of reliability. In contrast, latent variable theory, and its confirmatory factor analytic (CFA) basis, is used more widely to address issues of validity whilst controlling for reliability. This study clarifies the types of questions to which G theory can be applied by testing whether there is any justification for differences in interpretation between results based on G theory and latent variable theory. We reanalyzed data from an operational assessment center (N = 214 managerial assessees) and found comparable aggregated effects, generalizability coefficients, and latent scores across the G theory and latent variable theory approaches, suggesting that both can be applied to problems related to reliability and structural validity.

Keywords: *Psychological assessment, multifaceted assessment, generalizability theory*

INTRODUCTION

In applied psychology research and practice, the measurement of job-relevant characteristics is often complex and multifaceted. For example, assessment centers (ACs) utilize a complex design in which the ratings assigned to participants are a function of multiple interacting influences, such as raters, rating items, performance dimensions, and management simulation exercises (Lance, Foster, et al., 2007). Other relevant examples include personality assessments, job performance measures, situational judgment tests, and gamified assessments (Christian et al., 2010; Gnambs, 2015; Jackson, Kim, et al., 2016). Multifaceted assessments are widely applied in New Zealand as well as internationally (Krause & Thornton, 2009; Taylor et al., 2002). More general concerns about exercising statistical control over AC scores are relevant to the indigenous Māori population of New Zealand. Investigations into subgroup differences in this context have been explored in previous work on ACs (Jackson & Englert, 2011) and in other measures used in employee selection (Guenole et al., 2003).

The complex, multifaceted design of many organizational measurement systems presents a considerable challenge to those seeking to establish the extent to which they are valid and reliable. Generalizability theory (G theory) was originally developed specifically to address multifaceted measurement designs (Cronbach et al., 1972; Cronbach et al., 1963) and is therefore well-suited to such procedures commonly observed in organizations. Fairly recent developments around the application of G theory to ill-structured measurement designs broaden its applicability, given how common these designs are in organizations (Putka et al., 2008). Yet, compared to applications of the more widely applied latent variable theory, G theory retains the status of the "underdog" with fewer research studies employing its use. As rough indication, a recent no-limits search of Business Source Complete with the keywords "generalizability theory" and "organization" only returned 92 hits. Replacing the former search term with "confirmatory factor analysis" increased the hit rate to 1,028.

In latent variable theory (e.g., Borsboom, 2008), CFA is routinely used to examine both reliability and validity. On the other hand, in G theory (e.g., Brennan, 2001), random effects models (REMs) are often used but, conceptually, their application is routinely restricted to an examination of reliability. In this article, we explain the relative advantages of G theory and REMs over latent variable theory and CFA in the psychometric evaluation of multifaceted measurement systems. We discuss a possible reason why G theory has been underutilized, particularly in examining issues relating to validity. One noteworthy explanation in this respect is a concern that the REMs utilized in analyzing G theory models may not produce results which are comparable to those generated with CFA. Directly addressing this issue, we examine the extent to which REMs and CFAs produce equivalent outcomes by reanalyzing a real-world data set using both approaches.

Conceptions of Validity

The precise meaning of validity is complex and the focus of ongoing debate (Borsboom et al., 2004). Putka

and Sackett (2010, p. 39) define validity as "the degree to which evidence supports inferences one proposes to draw about the target of assessment". Central to this definition is that the researcher is compelled to provide sufficient evidence to support the validity-related claims they make about their measurement procedure (Eignor, 2013).

Multiple forms of evidence might be used to support the case for the existence of hypothetical constructs (Campbell & Fiske, 1959; Cronbach & Meehl, 1955), including face, content, predictive, discriminant, and convergent validity elements. Another common form of validity evidence concerns the structure of ratings or responses within a given assessment procedure, akin to the concept of "substantive coherence" internal to the measure itself (e.g., Finch & French, 2015, p. 152). For example, the researcher's focus might be on the extent to which ratings in an AC support the assumption that the raters are evaluating candidates on performance dimensions rather than on exercise performance (e.g., Lance et al., 2004). Such evidence facilitates an understanding about how measures function internally. An understanding about the internal structure of measures offers insights into why criterion-related relationships with external measures might be evident and so can be used effectively in conjunction with other forms of validity evidence (Putka & Sackett, 2010). Thus, while structure represents a single form of evidence, it might nevertheless be critical, particularly if the researcher investigates how the measurement structure interacts with other forms of validity evidence.

Conceptions of Reliability

Reliability is traditionally defined as being concerned with measurement error, or variance that interferes with the assessment of constructs focal to the researcher's aims (Borsboom & Mellenbergh, 2002; Schmidt et al., 2000). Putka and Sackett (2010) summarize contemporary, operational perspectives on reliability as relating to replication, expectation, and consistency. *Replication* refers to the reproducibility of an observation relating to a given construct. *Expectation* refers to the ability to infer from (a) observations (e.g., items, raters) used in a procedure to a hypothetical population of observations deemed as admissible for measuring a construct of interest, and (b) observations in a sample to those in a population of participants. *Consistency* refers to those elements of the measurement procedure that replicate and thus either contribute to construct measurement (i.e., an estimate of true score variance, see Spearman, 1907) or, less desirably, to some consistent but construct-irrelevant source of variation. Conversely, elements of the measurement procedure that fail to replicate contribute to undesirable error variance in observations.

In classical test theory and in G theory, reliability is represented by the ratio of true score (referred to as universe score in G theory) score to total variance (i.e., sources of universe score / sources of universe score + sources of error, see Crocker & Algina, 1986). This ratio is often referred to as a generalizability coefficient or G coefficient (Shavelson & Webb, 1991). In classical test theory, reliability is typically estimated for different purposes or perspectives on reliability using separate reliability coefficients. For example, coefficient alpha is applied to questions about internal consistency, whereas test-retest reliability coefficients are applied to questions about temporal stability (Nunnally & Bernstein, 1994). In contrast, G theory allows the researcher to estimate and thus control for multiple perspectives on reliability simultaneously (e.g., in G theory it is possible to estimate effects relevant to internal consistency and temporal stability within the same analysis, Cronbach et al., 1972). This can present a more controlled perspective on reliability, particularly in complex, multifaceted measurement designs.

Applications of Latent Variable Theory and CFA

Many of the measurement procedures used in applied psychology reflect a simple measurement design often involving items, constructs, and respondents. During the early and middle parts of the 20th century, classical test theory was applied to this type of design. The central assumption of classical test theory is that a person's score on a test is a function of their true score on a latent construct or trait (e.g., conscientiousness) plus error. Here, error is viewed as being a consequence of multiple unmeasured variables associated with test administration, the candidate, and the test itself.

More recently, the development of CFA has made it possible to separate a general estimate of error into separate components, allowing a more detailed test of latent variable theory (Brown, 2006; Lance et al., 2002). By combining the error of measurement associated with each of the items involved in assessing individuals on a particular latent trait with residual error, an overall index of the reliability of the measure in evaluating that trait is obtained via CFA (Brown, 2006). Further, by examining how well relevant data sets fit the model proposed to measure the latent trait, CFA can be used to assess one form of validity evidence relating to the structure of the instrument (Borsboom & Mellenbergh, 2002).

CFA is of considerable utility in examining the reliability and validity of relatively simple measurement designs. However, its application can be limited in more complex measurement systems of the type often used in organizations. For example, in structured interviews (e.g., Saunders & Townsend, 2016), two or more raters may evaluate candidates against groups of items nested within several dimensions (e.g. communication skills, teamwork etc.). As the number of variables involved in a measurement design grows, so do the complexities involved in establishing the validity and reliability of that design. Here, the reliability of interviews depends on multiple, systematic measurement components, including raters, items, dimensions, all possible interactions between these elements, and residual error due to other unknown influences.

G Theory as an Approach to Reliability

In the organizational literature, as well as in others, CFA is widely applied as an indication of construct-related evidence (e.g., Borsboom, 2008; Brown, 2006; Eid et al., 2008; Lance, Foster, et al., 2007; Lance, Woehr, et al., 2007). However, less clarity surrounds the purpose of G theory in addressing issues concerning reliability or validity. At its inception, G theory was primarily presented as a framework for understanding *reliability* in multifaceted measurement. Cronbach et al. (1963)

described G theory as a "liberalization of reliability theory" (p. 137), and primarily framed their arguments for the development of the theory in terms of reliability. As they developed G theory, Cronbach et al. (1972, p. 15) further positioned it as being concerned primarily with reliability, drawing attention to the flexible approach it provides, in that, based on judicious reasoning, theory, or research evidence, researchers can specify multiple sources of universe score (the G theory analogue of true score) and error. Classical test theory, on the other hand, usually offers no such flexibility (Brennan, 2000).

Other researchers and methodologists followed Cronbach et al. (1972) in presenting G theory as being principally concerned with the study of reliability. Brennan (2001) discusses the idea that conditions of measurement influence error or variability in scores, and that it is possible for researchers using G theory to quantify such influences. On summarizing the aims of G theory, Brennan states that "historically these types of issues have been subsumed under the heading "reliability". Generalizability theory liberalizes and extends traditional notions of reliability" (p. 2). Similarly, and consistent with the Cronbach et al. description, Shavelson and Webb (1991) make reference to the focus in G theory on the dependability of scores. They state that the G coefficient often reported in G theory analyses is "analogous to classical test theory's reliability coefficient" (p. 2).

In some of the most recent treatments of G theory in organizational contexts, researchers continue to frame the approach as a perspective on score reliability. Putka and Hoffman (2013, p. 115) separated measurement error in a G theory model into components classified as "reliable and unreliable". Similarly, Putka and Hoffman (2014) framed their chapter on the application of G theory to job performance measures as a perspective on reliability. Akin to the perspective presented by Putka and colleagues, Jackson, Michaelides, et al. (2016) and Jackson et al. (2020) presented their G theory models as perspectives on reliable and unreliable sources of variance related to ACs and multisource performance ratings respectively, implying that G theory primarily concerns reliability. LoPilato et al. (2015, p. 693) defined G theory as a "statistical framework for identifying factors that affect the reliability of measurements". Woehr et al. (2012) stated that "Typically, G-theory is introduced and discussed in the context of reliability estimation" (p. 15).

G theory as an Approach to Validity

Although, the descriptions offered above suggest that G theory is primarily concerned with reliability, not all researchers describe the approach as being restricted to the reliability domain, and indeed several scholars position it primarily as an approach towards summarizing validity evidence. Arthur et al. (2000, p. 819) had as one of their research objectives "to recommend and demonstrate the use of generalizability theory analysis to assess convergent/discriminant validity" in the context of AC ratings. They expanded on this description, noting that "Evidence of construct-related validity is derived from the extent to which variance associated with the constructs of interest (measurement focus) is large relative to the variance associated with conditions of measurement[1]". Lievens (2001b, p. 203) aimed to "shed light on the issue of assessment center construct validity" using G theory as a basis. Similarly, Lievens (2001a) applied G theory, in part, to examine evidence of "discriminant validity" in ratings from assessor training (p. 259). In the context of multitrait-multimethod matrices (MTMMs), Woehr et al. (2012) investigated the question: "How do the variance components stemming from G-theory relate to the traditional notions of construct-related validity?" (p. 141), and demonstrated how effects estimated via G theory have analogs in classic work on MTMMs (e.g., Campbell & Fiske, 1959). Highhouse et al. (2009) described G theory as "an especially powerful method for gathering construct validity evidence" (p. 784).

The Applicability of G Theory

While we discuss reliability and validity separately above, this distinction is not altogether clear in the psychometric literature, with Campbell and Fiske (1959) describing it in terms of "regions on a continuum" (p, 83). This idea is reflected in the developmental stages of G theory, where Cronbach et al. (1963) noted that "the theory of 'reliability' and the theory of 'validity' coalesce" in the context of G theory[2] (p. 157), and Cronbach et al. (1972) stated that "generalizability theory blurs the distinction between reliability and validity" (p. 380). An elaboration of this latter statement was offered by Brennan (2000). In a typical G theory-based analysis, multiple, systematic facets[3] are isolated in a data set. Brennan suggests that some of these facets might be associated with validity (e.g., Participant × Trait interactions) and others with reliability (e.g., Participant × Item interactions).

Notwithstanding these observations, recent and historical perspectives on G theory suggest that the approach is primarily concerned with reliability (e.g., Cronbach et al., 1972; Cronbach et al., 1963; Jackson et al., 2020; Jackson, Michaelides, et al., 2016; LoPilato et al., 2015; Putka & Hoffman, 2013, 2014; Putka & Sackett, 2010; Thompson, 2003). This perhaps limits its perceived usefulness. Therefore, and only for the purposes of comparison in this paper, we begin by assuming the popular perspective that the purpose of G theory is to summarize reliability evidence. In Table 1, we present effects relevant to an example task-based AC model (Jackson et al., 2010; Thoresen & Thoresen, 2012) and

[1] We infer here that "conditions of measurement" refer to those measurement conditions not specified as relating to constructs of interest (e.g., variance related to items, raters, etc.).

[2] Here, Cronbach et al. (1963, p. 157) specifically refer to the idea that the universe of admissible observations is a construct domain introduced by the researcher that has potential "explanatory or predictive power". A G theory analysis therefore offers suggestions about "how validily one can interpret a measure as representative of a certain set of possible measures" (see p. 157).

[3] A *facet* is any systematic source other than participants that contributes to variance in scores (e.g., items, raters, etc).

Table 1. Generalizability Theory and Confirmatory Factor Analytic Perspectives on Task-Based Assessment Center Effects

Effect	Common interpretation	G theory perspective, often associated with[a]:	CFA perspective, often associated with:
p	General factor	Reliability (relative[b])	Validity, structural
pc	Role-exercise-dependent interaction	Reliability (relative)	Validity, structural
pi:c,e	General, item, and role-exercise interaction + residual variance	Reliability (relative)	Reliability
c	Role-exercise main effect	Reliability (absolute[c])	NA
i:c	Item-in-role-exercise main effect	Reliability (absolute)	NA

Note. In task-based assessment centers, exercise factors represent role constructs of interest and are not considered to be method or mode effects. p = participant; c = role-exercise construct; i = rating item; e = residual error variance; G theory = generalizability theory; CFA = confirmatory factor analysis; NA = non-applicable. Exercise and item main effect estimates are unavailable in typical CFA output. [a]We acknowledge that there is considerable variability in the literature here, with several authors positioning G theory as capable of summarizing validity evidence. We refer here to some of the original (e.g., Cronbach et al., 1963) and the most recent (e.g., Putka & Hoffman, 2013) perspectives on this issue. [b]Variance components associated with relative decisions apply where the aim is to evaluate the score of an individual in terms of how it relates to scores from a larger group (e.g., norm-referenced scores). [c]Absolute decisions, on the other hand, are concerned with cut-off scores (e.g., pass/fail criteria), which are arguably less common in studies of organizations (see Shavelson & Webb, 1991 for further discussion on relative versus absolute decisions).

compare the hypothetical interpretation of these effects from a reliability-oriented G theory perspective against a more widely applied latent variable theory perspective analog. Of the three effects that are available for comparison across the two perspectives, only one, that for residual error, shares the same interpretation across the G theory and latent variable methodological frameworks.

Assuming that the reliability and validity concepts are meaningfully distinguished from one another, cross-theory differences in the interpretation of effects raises a conundrum. We suggest that evidence for reliability should be interpreted according to an accepted definition of reliability, regardless of the approach used to garner that evidence. Likewise, evidence for validity should be interpreted as it relates to an accepted definition of validity, and the status of such evidence should not depend on the approach used in its collection. Variability in the interpretation of effects in this respect could impede progress in understanding organizational phenomena.

Comparing G theory- and Latent Variable Theory-Related Methods

Why is it that output from methods associated with traditional and recent perspectives on G theory is framed as an examination of reliability (e.g., LoPilato et al., 2015), whereas output from methods associated with latent variable theory is often interpreted as it relates to an examination of validity (e.g., Borsboom, 2008)? It is possible that that the REMs popularly applied in G theory versus the CFAs in latent variable theory simply produce fundamentally different results. Output from these methods could lend itself more towards an interpretation based in reliability in G theory, and validity in latent variable theory.

Several researchers have replicated results from REMs using constrained CFA models (Marcoulides, 1996; Raykov & Marcoulides, 2006; Woehr et al., 2012). Notably, in the context of MTMMs used in organizations, Woehr et al. replicated the variance estimates in a univariate[4] REM model with a constrained CFA model. Thus, the capacity for CFA to reproduce REM results is known. However, the ability to reproduce the same variance estimates across REM and CFA addresses only a component of the problem discussed here. Two key issues here are how those results are interpreted (i.e., as reliability and/or validity evidence); and whether there is any justification for interpreting results differently based on the method from which they have been derived, and the specific theoretical framework on which a given method is based.

A relevant consideration is that in G theory, aggregation formulae are often applied to REM variance estimates in a manner that is not typical or even clearly possible in a traditional latent variable theory framework via CFA. Aggregation can greatly influence relative effect size in a measurement model (Kuncel & Sackett, 2014; Putka & Hoffman, 2013). The effect estimates in a REM are orthogonal and this statistical property enables aggregation formulae to be selectively applied to relevant effects (Brennan, 1992, 2001; Searle et al., 2006). In principle, it is possible to apply G theory-based aggregation formulae to estimators generated via a CFA constrained in a manner analogous to a corresponding REM, and then to compare outcomes from both types of analysis. It would be possible here to establish whether there is any justification for interpreting effects differently across methods, given the application of formulae usually applied in G theory.

An issue related to aggregation formulae in G theory centers on G coefficients. The G coefficient is widely applied to analyses invoking the G theory framework (Brennan, 2001). Whether applying G coefficients based on REM versus CFA estimators makes a difference to statistical outcomes is currently unclear. If, overall, REM and constrained CFA results are similar, the justification for interpreting one type of analysis differently from another, depending on whether the researcher takes a G theory or latent variable theory perspective, is weakened.

Both REMs and CFAs are used to indicate variance associated with constructs in G theory and latent variable theory respectively (Borsboom, 2008; Cronbach et al., 1972). It is possible to generate latent scores (sometimes referred to as factor scores) for these construct effects both in REMs and in CFAs. Latent scores are defined as an estimate of a participant's relative standing on a construct

[4] We focus on univariate REMs , given the similarities between multivariate REMs and their widely-criticized correlated uniqueness CFA analog (Lance et al., 2002; Woehr et al., 2012).

of interest. In conceptual terms, latent scores provide an indication of what a participant's score would have been on the construct of interest, had it been possible to measure it directly (Brown, 2006) and are relevant to constructs evaluated via multifaceted assessment. A consideration of latent scores in REMs and G theory is rare (however, see Ward, 1986) and we were unable to find any sources where REM- and CFA-derived latent scores had been compared. Such a comparison could shed light on what is perhaps the core purpose of many multifaceted measures: their capacity to produce intended construct scores. Differences in effect size and patterns of intercorrelation between latent scores generated through REM versus CFA might offer suggestions about the basis for differences in the interpretation of their respective outputs. This could, in turn, highlight whether there are fundamental differences between REMs and CFAs that justify restrictions in the scope of application associated with G theory.

Summary

Under a latent variable theory perspective, CFA is regularly considered to be concerned with structural validity as well as reliability (Borsboom, 2008; Borsboom et al., 2004; Eid et al., 2008; Kleinmann & Köller, 1997; Lance, Woehr, et al., 2007). Historical and recent perspectives on G theory, via the interpretation of REMs, position it primarily as a perspective on reliability (Thompson, 2003). The latter perspective restricts the scope of G theory relative to latent trait theory, in terms of the types of research questions that it can address. This might have limited the popularity of G theory, despite the fact that the REMs commonly used by G theorists are, in many circumstances, more accommodating of the complex research designs often encountered in organizational research (Michalak et al., 2019; Soltani et al., 2005). If a comparison between estimators generated using a REM and CFA reveals little difference in outcomes, even when considering aggregation, G coefficients, and latent scores, then this would call into question differences in interpretation from G theory versus latent variable theory standpoints. In keeping with these arguments, we propose the following, three Research Questions (RQs):

> *RQ1:* When comparing aggregated results across REM, constrained CFA, as well as traditional CFA output, is there any justification for interpreting effects differently across methods as they relate to reliability or validity?

In RQ1, as we expand on below, our intention is to create two analyses: one based on a REM and the other based on a CFA, that constrain their estimates in a similar manner. The intention is to create variance component estimates that are directly comparable, but that have been generated using different estimation processes.

> *RQ2*: Do G coefficients based on REM and CFA return similar outcomes?

In RQ2, our aim is to use the variance components mentioned for RQ1 to generate G coefficients that are directly comparable across estimates based on REM and CFA.

> *RQ3*: Do latent scores based on REM and CFA generate similar effects and patterns of intercorrelation?

In RQ3, we aim to produce latent scores that are directly comparable for analyses based on REM and CFA so that they can be contrasted against one another and correlated.

METHOD

Our data-analytic aims in this study center on providing a comparison between effects generated using REMs and effects generated using CFA. For this purpose, we reanalyzed a subset of data from Jackson et al. (2010). Our interest here was in testing a model with a small number of effects so that it could be easily reproduced in both REMs and CFAs and to maintain simplicity and brevity. In the original study, the authors analyzed data from a task-based AC, which is a simplified version of a traditional AC, where role constructs that are assessed within each exercise. Thus, scores for each exercise in a task-based AC represent role-exercise constructs (Jackson, 2012). We provide a brief description of participants and materials below. A full description of the AC under scrutiny is available in the Jackson et al. (2010) article. We note here that our aims are not oriented towards contributing to the literature on the structural characteristics of ACs and our inclusion of data related to a task-based AC is incidental.

Participants

A total of 214 managerial assessees from New Zealand participated in the study (we removed ratings from 1 participant due to incomplete data, bringing our analysis *N* to 213). The mean age of participants was 45.53 (*SD* = 10.33) and 54% of the sample were men, 46% were women. The organization under scrutiny specialized in postal, insurance, credit, banking, and administrative services. Assessees were evaluated by 19 assessors ranked one level above assessees and 4 additional assessors who were employed as consultant psychologists. We could not estimate assessor-related effects because the ratio of assessors to assessees was set at 1:2 to reduce costs for the participant organization. However, recent research across multiple samples suggests that assessor-related effects tend to be small (see Jackson, Michaelides, et al., 2016; Putka & Hoffman, 2013), assuming that assessors are adequately and appropriately trained. In the present case, assessors were trained using a frame-of-reference training (FORT) procedure, as recommended in the assessor training literature (Gorman & Rentsch, 2009; Pulakos, 1986). Training lasted for a 2-day period and covered familiarization with assessment materials, common rater errors, and mock assessments with related FORT discussions.

AC Characteristics

AC ratings related to (a) a group discussion and oral presentation based on managing new staff (i.e., the management role), (b) a group discussion and oral presentation on selecting new staff (i.e., the human resource selection role), and (c) a group-based problem-solving exercise (i.e., the contextualized problem-solver role). Thus, the role-exercises included 3 levels

represented for each exercise. For each role-exercise construct, 7 behavioral descriptor items (21 items in total, e.g., *uses objective and non-emotive language when delivering feedback to others*) were retained for analysis. Behavioral descriptors were rated on a scale ranging from 1 (certainly below standard) to 10 (certainly above standard). All exercises were developed based on competency and inductive job analyses (Tett et al., 2000; Williams & Crafts, 1997).

Analyses

Our primary interest was in comparing two analogous models: one based on REMs, the other on CFA. The first model comprised a REM (see Searle et al., 2006) with restricted maximum likelihood (REML) estimation as a representation of the models typically used in contemporary studies using G theory (e.g., Putka & Hoffman, 2013, 2014). A total of 3 main effects were estimated in this model, relating to participant assessees (*p*), role-exercise constructs (*c*), and rating items nested in role-exercise constructs (*i*:*c*)[5]. Taking interactions between effects into account, this resulted in a total of 5 effects that could be estimated within the REM model, each of which is listed and described in Table 1.

The second model that we tested was based on a CFA constrained to enable estimation in a manner analogous to that relevant to the REM (Marcoulides, 1996; Raykov & Marcoulides, 2006). This involved constraining the CFA model to have equal latent factor variances and unique variances. All factor covariances and error covariances were constrained to zero and all factor loadings were constrained to 1 (see Woehr et al., 2012, p. 144, Figure 2 caption). It was possible to estimate 3 effects with this approach, including the analogs of the main effect for *p*, the *pc* interaction, and an estimate for residual variance (see Table 1 for a description of these effects). To add a supplementary perspective, we tested a regular CFA model with correlated latent factors (as depicted in Figure 1).

To the REM and CFA variance estimates, we applied aggregation and G coefficient formulae based on those from the extant G theory literature (Brennan, 2001; Jackson et al., 2020; Putka & Hoffman, 2013; Shavelson & Webb, 1991). We extracted latent scores relating to role-exercise constructs from both the REM and the constrained CFA and correlated the two sets of latent scores. For the REM analysis, latent scores were derived from random intercepts relating to Participant × Exercise interactions (e.g., Liu et al., 2008). For CFA, latent scores were represented for each construct by the average of the product of each item response and its associated factor loading (e.g., Brown, 2006). The REM was conducted using the lmer function in lme4 for R (Bates et al., 2015). The CFA was conducted using lavaan for R (Rosseel, 2012). G coefficients were specified such that the effects for p and pc defined universe score. This is because p represents general individual differences, which is routinely of focal interest in an evaluation approach (Shavelson & Webb, 1991). The pc interaction represents individual differences on the focal constructs of interest, and thus represents a source of value to the evaluation instrument (Putka & Hoffman, 2013). The residual effect was specified as contributing to error.

RESULTS

To provide a perspective on goodness-of-fit, we tested the model shown in Figure 1, which represents the standard CFA model implied in the task-based AC literature with correlated, latent role-exercise constructs (e.g., Jackson et al., 2010; Thoresen & Thoresen, 2012). The model converged within expected parameters and model fit was acceptable according to criteria specified in Brown (2006)[6], $\chi^2 = 245.74(165)$, $p < .001$; comparative fit index (CFI) = .970; Tucker-Lewis index (TLI) = .962; root-mean-square error of approximation (RMSEA) = .045; standardized root-mean-square residual (SRMR) = .043. Averaged, squared standardized loadings suggested effect sizes for the general factor = .19, role-exercise constructs = .40, and unique variance = .41. Averaged,

Table 2. Comparison of Generalizability Theory and Confirmatory Factor Analytic Effects

Effect	Generalizability theory estimates					Constrained CFA estimates			
	VC	Total %	BP %	Aggregation formula	BP % Aggregated	VC	BP %	Aggregation formula	BP % Aggregated
σ_p^2	.4224	26.08	27.71	σ_p^2	45.11	.4200	27.69	σ_p^2	45.09
σ_{pc}^2	.4159	25.68	27.28	σ_{pc}^2	44.42	.4140	27.29	σ_{pc}^2	44.44
$\sigma_{pi:c,e}^2$	.6862	42.37	45.01	$\sigma_{pi:c,e}^2/n_{i:c}$	10.47	.6830	45.02	$\sigma_{pi:c,e}^2/n_{i:c}$	10.47
σ_c^2	.0216	1.33	–	–	–	–	–	–	–
$\sigma_{i:c}^2$	.0736	4.55	–	–	–	–	–	–	–
G					.90				.90

Note. p = participant main effect (or general performance effect); c = exercise-role construct; i = item; e = residual error; G = generalizability coefficient; VC = variance component; BP = between-participant variance. Dashes indicate non-applicability. Generalizability theory estimates are derived from the variances in a random effects model. Confirmatory factor analysis (CFA) estimates are derived from variances in a CFA model constrained as described below. The residual in the CFA analysis is estimated using the formula $1 - \sigma_p^2 + \sigma_{pc}^2$. Constraints imposed on the CFA model included equal c factor variances, equal unique variances, c factor covariances constrained to zero, c factor and p factor covariances constrained to zero, all error covariances constrained to zero, and all factor loadings constrained to one.

[5] In G theory notation, the presence of a colon (:) indicates a level of nesting. For example *i*:*c* implies that items are nested in constructs.

[6] CFA-related goodness-of-fit was not estimated for the constrained models that follow because associated fit indices can "reflect types of misfit that have little or no bearing on the accuracy of G-theory model parameter estimates" (Woehr et al., 2012, p. 158).

Table 3. Factor Score Correlations

REM latent scores			
	c1	c2	c3
c1			
c2	.61		
c3	.57	.42	
CFA latent scores			
c1			
c2	.61		
c3	.58	.48	
REM with CFA latent scores			
c1	.99	.59	.57
c2	.60	.99	.45
c3	.57	.45	.99

Note. c = role-exercise construct, G theory = generalizability theory, REM = variance components analysis, typically used in G theory; CFA = confirmatory factor analysis. In the bottom matrix, REM latent scores appear on the vertical axis. REM estimates based on latent scores for $\sigma_p^2 + \sigma_{pc}^2$, where p = participant. CFA latent scores were estimated from a correlated 3c model.

squared standard covariances among role-exercise constructs = .08.

The models used in REMs, and often as a basis for G theory, offer a somewhat different perspective on observed data than that associated with CFA. To allow for comparison between the CFA and REM analyses, we constrained the CFA model in Figure 1 as described in the note in Table 2, in keeping with guidance provided in the methodological literature (Marcoulides, 1996; Raykov & Marcoulides, 2006; Woehr et al., 2012). Table 2 shows a comparison between variance components from on a REML-based REM and variance components from an analogous, restricted CFA model. Both models converged acceptably.

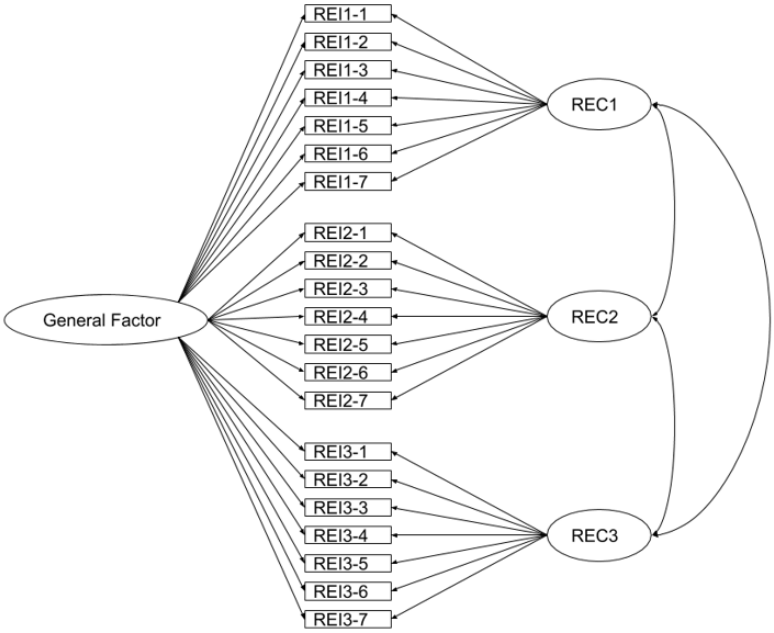
We applied formulae to REM and analogous CFA estimates in Table 2 based on those commonly applied in the G theory literature (see Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991), so as to approximate the effects of aggregation on variance estimates. In the present case, only aggregation to role-exercise scores was considered, because this is of focal interest in task-based ACs (Jackson et al., 2005; Lance, 2012). A total of 5 effects were available for the REM, which included 2 main effects that were not relevant to between-participant comparisons. The remaining 3 effects were relevant to between-participant comparisons and were available in both the REM and CFA analyses.

The results presented in Table 2 show outcomes that are almost identical when comparing across the G theory and CFA analyses. With respect to effect size, both pre- and post-aggregated results only differed by a maximum of .02 of a percentage point, thus indicating near zero differences between outcomes generated by the two analytic approaches (in response to RQ 1). As expected for the task-based approach to ACs, the majority of variance on aggregation was associated with general performance (around 45%, regardless of estimation approach) and Participant × Role-Exercise Construct interactions (around 44%, again regardless of estimation approach). Table 2 shows that formulae for aggregation commonly applied in the G theory literature can be applied in the same way to constrained CFA variance components with practically the same outcomes. It follows that G coefficients, estimated on both REM and CFA variance components (see RQ 2), in both cases, analogs of $(\sigma_p^2 + \sigma_{pc}^2)/[\sigma_p^2 + \sigma_{pc}^2 + (\sigma_{pi:c,e}^2/n_{i:c})]$, resulted in identical outcomes to 2dp at .90 (where p = participant, c = exercise-role construct, i = item, and e = residual error).

The results shown in Table 2 and the results of the traditional CFA model (shown in Figure 1) also reflect similar outcomes. However, the different methods need to be considered with respect to their treatment of data. In the REMs that act as the basis for G theory (shown in Table 2), it is assumed that any latent constructs under consideration do not share any common variance beyond that which is already accounted for by the general effect, σ_p^2 (Marcoulides, 1990; Raykov & Marcoulides, 2006). In the case of the traditional CFA depicted in Figure 1, the effect size for role-exercise construct loadings = 39.63% and for the general factor = 19.38% (based on average, squared standardized loadings). The summary role-exercise effect here was of a greater magnitude than that presented in Table 2 for the G theory model (see results prior to aggregation: role-exercise effect = 27.71%, general factor = 27.28%). This is because the traditional CFA estimate for the role-exercise effect includes variance shared between role-exercise constructs. Once these method-specific idiosyncrasies are acknowledged, even the results of the traditional CFA are similar to those presented in Table 2 for the G theory analyses given that larger role-exercise effects are expected from a traditional CFA.

To provide an additional perspective on the outcomes above, we extracted latent scores for role-exercise constructs based on both REM and CFA estimates (see RQ 3). Table 3 shows three matrices, which display correlations between (a) REM latent scores, (b) CFA latent scores, and (c) REM latent scores and CFA latent scores. When comparing the separate REM and CFA outcomes (i.e., a and b above), it is clear in Table 3 that the two modes of estimation make very little difference to how the latent scores intercorrelate. The largest of these differences was between role-exercise constructs 2 and 3 (r = .42 versus r = .48). When expressed in terms of a percentage of variance explained, this is a near-zero difference (i.e., < .004%). REM and CFA latent scores (i.e., c above), shown in the diagonal of the bottom matrix in Table 3, correlated at a uniform .99 for all 3 role-exercise constructs. This provides further evidence that the results across G theory and analogous CFA methods are, for practical purposes, almost identical.

Figure 1. Task-based assessment center confirmatory factor analysis model, showing role-exercise indicators (REI1-1 through REI3-7), role-exercise latent constructs (REC1 through REC3), and a general factor.



DISCUSSION

G theory has never reached the status of a mainstream methodological approach in applied psychology, despite a lengthy history and wide applicability to the complex measurement designs routinely found in organizations (Cronbach et al., 1972; DeShon, 2002; Putka & Hoffman, 2014). We posit that a key reason for this lack of uptake is because of uncertainties about what types of research questions G theory can be used to address. Both historically (Cronbach et al., 1963), and in recent organizational research (Jackson, Michaelides, et al., 2016; Putka & Hoffman, 2013, 2014), G theory has been characterized as an approach towards summarizing reliability evidence. However, some researchers position the approach as being relevant to summarizing validity evidence (Arthur et al., 2000; Highhouse et al., 2009; Lievens, 2001a, 2001b; Woehr et al., 2012). In contrast to the differing perspectives on the purpose of G theory, much more agreement is apparent about the role of CFA and its capacity to summarize structural validity-related evidence whilst also acknowledging reliability (e.g., Brown, 2006). It might therefore be no coincidence that CFA is more widely applied in the discipline (e.g., Lance et al., 2004; Lance et al., 2002) than is G theory (e.g., Murphy & DeShon, 2000).

We compared results from a G theory model based on a REM of a task-based AC (Jackson et al., 2010) with analogous results generated through a CFA model constrained to match the outcomes generated through the REM. Comparison of the REM and CFA outcomes, including those relating to aggregation formulae often applied in G theory (RQ 1), G coefficients (RQ 2), and latent scores (RQ 3), revealed that the two methods provided practically identical results (see Tables 2 and 3). We found that a regular CFA model with correlated latent factors suggested conclusions similar to those based on the REM.

Our results suggest that REM, the technique normally adopted when G theory is applied, provides a perspective that is analogous to that provided by CFA, and that there is, therefore, no cogent justification for cross-method differences in the interpretation of specific effects. Cronbach et al. (1972) stated that G theory blurs the reliability-validity distinction. Brennan (2000) suggested that Cronbach et al. referred here to the idea that G theory can address (a) sources of variance often considered to be about validity and (b) sources of variance often considered to be about reliability. Our results are consistent with Brennan's interpretation, and we offer the extension that irrespective of whether a G theory or CFA approach is used, any sources of variance related to observations (e.g., items, assessors) are likely to concern reliability, whereas any sources of variance related to the equivalent of latent constructs (e.g., dimensions, personality constructs, role-exercise constructs) are likely to concern structural validity.

In a G theory model, distinctions between sources of variance as they relate to validity or reliability might be straightforward in many cases because each effect is presented separately and can, potentially, be meaningfully categorized. For example, with reference to the between-participant effects listed in Table 2, the effects σ_p^2 and σ_{pc}^2 are concerned with the equivalent of CFA latent constructs and thus could be categorized as relating to validity evidence. The former of these effects represents the CFA analog of a general performance effect or *positive manifold* (e.g., Ree et al., 2015). The latter interaction represents the CFA equivalent of role-exercise latent constructs (Jackson, 2012). In contrast, the $\sigma_{pi:c,e}^2$ effect includes the influence of indicator items, and it could therefore be argued that this effect relates to reliability evidence.

What is less clear, perhaps, is how G coefficients should be conceptualized. If we accept the classification of effects as sources of either reliability or validity evidence as described above, then G coefficients combine aspects of both reliability and validity. That said, there is often a predictable pattern to how G coefficients are constructed in that validity-related effects commonly define the numerator and reliability-related effects commonly define the denominator in G coefficient equations. This is certainly the case in the present example where the G coefficient $(\sigma_p^2 + \sigma_{pc}^2)/[\sigma_p^2 + \sigma_{pc}^2 + (\sigma_{pi:c,e}^2/n_{i:c})]$ contains validity-related effects in the numerator and the reliability-related effect in the

denominator[7]. Thus, one interpretation of the G coefficient could be the ratio of structural validity to reliability evidence.

The finding of a relatively large proportion of variance associated with what is presumed to be a latent construct does not guarantee, in any way, the validity of the measure being applied (Putka & Sackett, 2010). It does suggest a systematic source of variance that is potentially relevant to the internal structure of the assessment procedure, which, we suggest, could count as one, limited, source of validity evidence. The possibility still exists, however, that this systematic source of variance might, in fact, be irrelevant to the construct(s) of interest. Other sources of evidence will be necessary to determine the nature of such effects, whether they relate to what was intended for measurement, and whether they relate meaningfully and as expected to externally measured constructs (see Strauss & Smith, 2009).

Implications

Our results suggest that G theory and CFA deal with sources of evidence for both reliability *and* structural validity. In future research involving G theory, researchers using either methodological approach could classify effects as they pertain to reliability or validity evidence, to assist in developing a clear and consistent understanding of the structure of multifaceted measures that does not depend on methodological context.

Our findings highlight the idea that the theoretical principles of G theory apply with the use of methods such as CFA, just as much as they apply when using REM. REM appears to have become synonymous with G theory, but, in fact, G theory is not REM. The "statistical machinery" (Brennan, 1997, p. 15) used to generate effects in G theory is secondary to the theory itself. As suggested in this paper, at least some G theory models can be adequately estimated using CFA. There are likely other statistical methods that could be used as a basis for G theory. Even within REM, there are different options that researchers can choose from to estimate effects, including those based on REML, ANOVA-analogous, or Bayesian estimators (Brennan, 2001). The main issue here, though, is that G theory should be thought of as a theoretical framework that is not anchored to a specific statistical method. While REM represents the most common basis for G theory, its aggregation formulae, G coefficients, and latent scores can be used with other statistical foundations, as we demonstrate with CFA.

Our results suggest that consideration should be given to the advantages and disadvantages of using one statistical basis over another for G theory. The benefits of employing CFA include that it can provide multiple perspectives on a data set, including a model constrained such that it is similar to a REM as well as a regular CFA model with correlated latent constructs. The latter model can provide more detail than REMs about each specific construct under scrutiny, as well as GFIs for the model as a whole (Le et al., 2009; Woehr et al., 2012). However, particularly with studies involving large numbers of effects, REMs might present a more practical approach than CFA because fewer parameters require estimation in REMs. Moreover, organizational measurement often requires the use of raters (e.g., in job performance evaluation or ACs). The presence of multiple raters might present a measurement design that is ill-structured (i.e., neither perfectly crossed nor nested, see Putka et al., 2011; Putka et al., 2008). While REML or Bayesian estimators in REM can handle ill-structured designs, there is often no practical way to address such designs in CFA (Putka et al., 2011; Putka et al., 2008).

Limitations

A limitation of our study is the simplicity of the model used to demonstrate comparisons between REM and CFA. However, we purposely chose a simple model (i.e., a model with a small number of effects) to facilitate an explanation of G theory, which is often described as conceptually complex (DeShon, 2002). Moreover, a small number of effects allows for direct comparisons between REM and CFA models, where such comparisons might not be practical with models that contain many effects. For example, it can be impractical to estimate effects related to raters with CFA because doing so could require a latent variable for each of potentially large numbers of raters (Jackson et al., 2020). The downside to the application of a simple model, however, is that we are unable to show from this study how different combinations of effects might contribute to universe score and error variance. Nonetheless, we are confident that the reader will be able to extrapolate in principle from the basic design presented here to more complex designs used in other operational assessment procedures.

For our G theory model, we could have explored alternatives to the REML estimators that we used. For example, Bayesian estimators have been recommended for more complex designs in the AC literature (Jackson, Michaelides, et al., 2016) and in the literature on multisource performance ratings (Jackson et al., 2020). Bayesian approaches provide an effective approach towards defining variability around effect estimates in the form of credible intervals that relate to a full posterior distribution (Gelman, 2006). However, empirical evidence suggests that G theory analyses based on Bayesian or REML estimators provide results that are similar or identical, assuming that none of the effects are fenced at zero (Jackson, Michaelides, et al., 2016; LoPilato et al., 2015). No fenced estimates were present in our analyses.

Conclusion

G theory is underutilized in applied psychology. We see this as an oversight because it is well suited to many of the measurement designs encountered in organizations, both in New Zealand and internationally. G theory could therefore help inform on theory and practice in organizational measurement. More clarity is needed on the types of research questions that G theory can be used to address, albeit those concerning reliability and/or validity evidence. Our results suggest that G theory can

[7] Note that the object of measurement here is participants (p) and at least some effects relating to p almost always define universe score. In G theory, it is possible to combine different elements of universe score and error, but the onus is on the researcher to justify this classification.

be used to evaluate both reliability and structural validity evidence in a similar manner to how CFA is routinely applied. Effects representing observations can be categorized as relating to reliability and effects representing analogs of latent constructs can be categorized as relating to one type of structural validity evidence, similar to the latent constructs addressed by CFA (see Strauss & Smith, 2009).

References

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*(4), 813-835. https://doi.org/10.1016/S0149-2063(00)00057-X

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

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives, 6*, 25-53. https://doi.org/10.1080/15366360802035497

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables and constructs: A comment on Schmidt and Hunter. *Intelligence, 30*(6), 505-514. https://doi.org/10.1016/S0160-2896(02)00082-X

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity [Psychometrics & Statistics & Methodology 2200]. *Psychological Review, 111*(4), 1061-1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brennan, R. L. (1992). *Elements of generalizability theory*. American College Testing (ACT) Publications.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice, 16*(4), 14-20. https://doi.org/10.1111/j.1745-3992.1997.tb00604.x

Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*(1), 5-10. https://doi.org/10.1111/j.1745-3992.2000.tb00017.x

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.

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

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*(1), 83-117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth.

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. https://doi.org/10.1037/h0040957

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137-163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

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

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods, 13*(3), 230-253. https://doi.org/10.1037/a0013219

Eignor, D. R. (2013). The standards for educational and psychological testing [Professional Ethics & Standards & Liability 3450]. *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.*, 245-250. https://doi.org/10.1037/14047-013

Finch, W. H., Jr., & French, B. F. (2015). *Latent variable modeling with R*. Routledge/Taylor & Francis Group.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515-533. https://doi.org/10.1214/06-BA117A

Gnambs, T. (2015). Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences, 84*, 84-89. https://doi.org/10.1016/j.paid.2014.08.019

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*(5), 1336-1344. https://doi.org/10.1037/a0016476

Guenole, N., Englert, P., & Taylor, P. J. (2003, Jun). Ethnic group differences in cognitive ability test scores within a New Zealand applicant sample. *New Zealand Journal of Psychology, 32*(1), 49-54. <Go to ISI>://000184192100007

Highhouse, S., Broadfoot, A., Yugo, J. E., & Devendorf, S. A. (2009). Examining corporate reputation judgments with generalizability theory. *Journal of Applied Psychology, 94*(3), 782-789. https://doi.org/10.1037/a0013934

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). 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., Kim, S., Lee, C., Choi, Y., & Song, J. (2016). Simulating déjà vu: What happens to game performance when controlling for situational features? *Computers in Human Behavior, 55*, 796-803. https://doi.org/10.1016/j.chb.2015.10.031

Jackson, D. J. R., Michaelides, G., Dewberry, C., Schwencke, B., & Toms, S. (2020). The implications of unconfounding multisource performance ratings. *Journal of Applied Psychology, 105*(3), 312-329. https://doi.org/10.1037/apl0000434

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, 101(7), 976-994. https://doi.org/10.1037/apl0000102
Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance, 18*(3), 213-241. https://doi.org/10.1207/s15327043hup1803_2
Jackson, D. J. R., Stillman, J. A., & Englert, P. (2010). Task-based assessment centers: Empirical support for a systems model. *International Journal of Selection and Assessment, 18*(2), 141-154. https://doi.org/10.1111/j.1468-2389.2010.00496.x
Kleinmann, M., & Köller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. *Journal of Social Behavior and Personality, 12*(5), 65-84.
Krause, D. E., & Thornton, G. C., III. (2009, Oct). A cross-cultural look at assessment center practices: Survey results from Western Europe and North America. *Applied Psychology: An International Review, 58*, 557-585. https://doi.org/DOI 10.1111/j.1464-0597.2008.00371.x
Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology, 99*(1), 38-47. https://doi.org/10.1037/a0034147
Lance, C. E. (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). Routledge/Taylor & Francis Group.
Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance, 20*(4), 345-362. https://doi.org/10.1080/08959280701522031
Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89*(2), 377-385. https://doi.org/10.1037/0021.9010.89.2.377
Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods, 7*(2), 228-244. https://doi.org/10.1037%2F1082-989X.7.2.228
Lance, C. E., Woehr, D. J., & Meade, A. W. (2007). A Monte Carlo investigation of assessment center construct validity models. *Organizational Research Methods, 10*(3), 430-448. https://doi.org/10.1177/1094428106289395
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*(1), 165-200. https://doi.org/10.1177/1094428107302900
Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*(2), 255-264. https://doi.org/10.1037/0021-9010.86.2.255
Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 22*(3), 203-221. https://doi.org/10.1002/job.65
Liu, X., Rong, J., & Liu, X. (2008). Best linear unbiased prediction for linear combinations in general mixed linear models. *Journal of Multivariate Analysis, 99*(8), 1503-1517. https://doi.org/https://doi.org/10.1016/j.jmva.2008.01.004
LoPilato, A. C., Carter, N. T., & Wang, M. (2015). Updating generalizability theory in management research: Bayesian estimation of variance components. *Journal of Management, 41*(2), 692-717. https://doi.org/10.1177/0149206314554215
Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports, 66*(2), 379-386. https://doi.org/10.2466/PR0.66.2.379-386
Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling, 3*(3), 102-109. https://doi.org/10.1080/10705519609540045
Michalak, R. T., Kiffin-Petersen, S. A., & Ashkanasy, N. M. (2019). I feel mad so I be bad': The role of affect, dissatisfaction and stress in determining responses to interpersonal deviance. *British Journal of Management, 30*(3), 645-667. https://doi.org/10.1111/1467-8551.12286
Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*(4), 873-900. https://doi.org/10.1111/j.1744-6570.2000.tb02421.x
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior & Human Decision Processes, 38*(1), 76-91. https://doi.org/10.1016/0749-5978(86)90027-0
Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology, 98*(1), 114-133. https://doi.org/10.1037/a0030887
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). Taylor & Francis.
Putka, D. J., Lance, C. E., Le, H., & McCloy, R. A. (2011). A cautionary note on modeling multitrait-multirater data arising from ill-structured measurement designs. *Organizational Research Methods, 14*(3), 503-529. https://doi.org/10.1177/1094428110362107
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*(5), 959-981. https://doi.org/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). Routledge.
Raykov, T., & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing, 6*(1), 81-95. https://doi.org/10.1207/s15327574ijt0601_5
Ree, M. J., Carretta, T. R., & Teachout, M. S. (2015). Pervasiveness of dominant general factors in organizational measurement. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*(3), 409-427. https://doi.org/10.1017/iop.2015.16

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Saunders, M. N. K., & Townsend, K. (2016). Reporting and justifying the number of interview participants in organization and workplace research [Research Methods & Experimental Design 2260]. *British Journal of Management, 27*(4), 836-852. https://doi.org/http://dx.doi.org/10.1111/1467-8551.12182

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*(4), 901-912. https://doi.org/10.1111/j.1744-6570.2000.tb02422.x

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

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

Soltani, E., Van Der Meer, R., & Williams, T. M. (2005). A contrast of HRM and TQM approaches to performance management: Some evidence. *British Journal of Management, 16*(3), 211-230. https://doi.org/10.1111/j.1467-8551.2005.00452.x

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology, 18*(2), 161-169. https://doi.org/10.2307/1412408

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1-25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Taylor, P. J., Keelty, Y., & McDonnell, B. (2002). Evolving personnel selection practices in New Zealand organizations and recruitment firms. *New Zealand Journal of Psychology, 32*, 49-54.

Corresponding Author
Duncan Jackson
Email: duncan.jackson@kcl.ac.uk

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

Thompson, B. (2003). A brief introduction to generalizability theory. In B. Thompson (Ed.), *Score reliability* (pp. 43-58). Sage.

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). Routledge.

Ward, D. G. (1986). Factor indeterminacy in generalizability theory. *Applied Psychological Measurement, 10*(2), 159-165. https://doi.org/10.1177/014662168601000206

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). Davies-Black Publishing.

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*(1), 134-161. https://doi.org/10.1177/1094428111408616