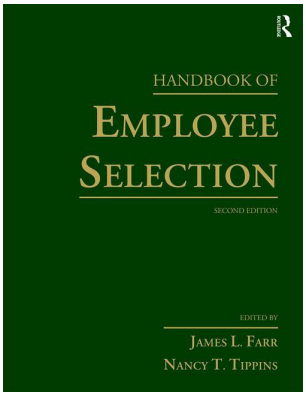

Handbook of Employee Selection

James L. Farr, Nancy T. Tippins, Walter C. Borman, David Chan, Michael D. Coovert, Rick Jacobs, P. Richard Jeanneret, Jerard F. Kehoe, Filip Lievens, S. Morton McPhail, Kevin R. Murphy, Robert E. Ployhart, Elaine D. Pulakos, Douglas H. Reynolds, Ann Marie Ryan, Neal Schmitt, Benjamin Schneider

Reliability

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-1
Dan J. Putka
Published online on: 22 Mar 2017

1

RELIABILITY

DAN J. PUTKA

Reliability and validity are concepts that provide the scientific foundation upon which we construct and evaluate predictor and criterion measures of interest in personnel selection. They offer a common technical language for discussing and evaluating (a) the generalizability of scores resulting from our measures (to a population of like measures), as well as (b) the accuracy inferences we desire to make based on those scores (e.g., high scores on our predictor measure are associated with high levels of job performance; high scores on our criterion measure are associated with high levels of job performance).[1] Furthermore, the literature surrounding these concepts provides a framework for scientifically sound measure development that, a priori, can enable us to increase the likelihood that scores resulting from our measures will be generalizable, and inferences we desire to make based upon them, supported.

Like personnel selection itself, the science and practice surrounding the concepts of reliability and validity continue to evolve. The evolution of reliability has centered on its evaluation and framing of "measurement error," as its operational definition over the past century has remained focused on notions of consistency of scores across replications of a measurement procedure (Haertel, 2006; Spearman, 1904; Thorndike, 1951). The evolution of validity has been more diverse—with changes affecting not only its evaluation but also its very definition, as evidenced by comparing editions of the *Standards for Educational and Psychological Testing* produced over the past half century by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (AERA, APA, & NCME, 2014). Relative to the evolution of reliability, the evolution of validity has been well covered in the personnel selection literature (e.g., Binning & Barrett, 1989; McPhail, 2007; Schmitt & Landy, 1993; Society for Industrial and Organizational Psychology, Inc., 2003) and will continue to be well covered in this Handbook. For this reason, this chapter will be devoted to providing an integrated, modern perspective on reliability.

In reviewing literature in preparation for this chapter, I was struck at the paucity of organizational research literature that has attempted to juxtapose and integrate perspectives on reliability of the last 50 years, with perspectives on reliability from the first half of the 20th century. Indeed, Borsboom (2006) lamented that to this day many treatments of reliability are explicitly framed or implicitly laden with assumptions based on measurement models from the early 1900s. While classical test theory (CTT) certainly has its place in treatments of reliability, framing entire treatments around it serves to "trap" us within the CTT paradigm (Kuhn, 1962). This makes it difficult for students of the field to compare and contrast—on conceptual and empirical grounds—perspectives offered by other measurement theories and approaches to reliability estimation. This state of affairs is highly unfortunate because perspectives on reliability and methods for its estimation have evolved greatly since Gulliksen's codification of CTT in 1950, yet these advances have been slow to disseminate into personnel selection research and practice.

Indeed, my review of the literature reveals what appears to be a widening gap between perspectives of reliability offered in the organizational research literature and those of the broader psychometric community (e.g., Borsboom, 2006; Raykov & Marcoulides, 2011). Couple this trend with (a) the recognized decline in the graduate instruction of statistics and measurement over the past 30 years in psychology departments (Aiken, West, & Millsap, 2008; Merenda, 2007), as well as (b) the growing availability of statistical software and estimation methods since the mid-1980s, and we have a situation where the psychometric knowledge base of new researchers and practitioners can be dated prior to exiting graduate training. Perhaps more disturbing is that the lack of dissemination of modern perspectives on reliability can easily give students of the field the impression that the area of reliability has not had many scientifically or practically useful developments since the early 1950s.

In light of the issues raised above, my aim in the first part of this chapter is to parsimoniously reframe and integrate developments in the reliability literature over the past century that reflects, to the extent of my knowledge, our modern capabilities. In laying out this discussion, I use examples from personnel selection research and practice to relate key points to situations readers may confront in their own work. Given this focus, note that several topics commonly discussed in textbook or chapter-length treatments of reliability are missing from this chapter. For example, topics such as standard errors of measurement, factors affecting the magnitude of reliability coefficients (e.g., sample heterogeneity), and applications of reliability-related data (e.g., corrections for attenuation, measure refinement) receive little or no attention here. The omission of these topics is not meant to downplay their importance to our field; rather, it just reflects the fact that fine treatments of these topics already exist in several places in the literature (e.g., Feldt & Brennan, 1989; Haertel, 2006; Nunnally, 1978). My emphasis is on complementing the existing literature, not repeating it. In place of these important topics, I focus on integrating and drawing connections among historically disparate perspectives on reliability. As noted below, such integration is essential, because the literature on reliability has become extremely fragmented.

For example, although originally introduced as a "liberalization" of CTT more than 40 years ago, generalizability theory is still not well integrated into textbook treatments of reliability in the organizational literature. It tends to be relegated to secondary sections that appear after the primary treatment of reliability (largely based on CTT) is introduced, not mentioned at all, or treated as if it had value in only a limited number of measurement situations faced in research and practice. Although such a statement may appear as a wholesale endorsement of generalizability theory and its associated methodology, it is not. As an example, the educational measurement literature has generally held up generalizability theory as a centerpiece of modern perspectives on reliability, but arguably, this has come at the expense of shortchanging confirmatory factor analytic (CFA)-based perspectives on reliability and how such perspectives relate to and can complement generalizability theory. Ironically, this lack of integration goes both ways, because CFA-based treatments of reliability rarely, if ever, acknowledge how generalizability theory can enrich the CFA perspective (e.g., DeShon, 1998), but rather link their discussions of reliability to CTT. Essentially, investigators faced with understanding modern perspectives on reliability are faced with a fragmented, complex literature.

OVERVIEW

This chapter's treatment of reliability is organized into three main sections. The first section offers a conceptual, "model-free" definition of measurement error. In essence, starting out with such a model-free definition of error is required to help clarify some confusion that tends to crop up when one begins to frame error from the perspective of a given measurement theory and the assumptions such theories make regarding the substantive nature of error. Next I overlay this conceptual treatment of error with perspectives offered by various measurement models. Measurement models are important because they offer a set of hypotheses regarding the composition of observed scores, which, if supported, can allow us to accurately estimate reliability from a sample of data and apply those estimates to various problems (e.g., corrections for

attenuation, construction of score bands). Lastly, I compare and contrast three traditions that have emerged for estimating reliability: (a) a classical tradition that arose out of work by Spearman (1904) and Brown (1910), (b) a random-effects model tradition that arose out of Fisher's work with analysis of variance (ANOVA), and (c) a CFA tradition that arose out of Joreskog's work on congeneric test models.

RELIABILITY

A specification for error is central to the concept of reliability, regardless of one's theoretical perspective, but to this day the meaning of the term "error" is a source of debate and confusion (Borsboom & Mellenbergh, 2002; Murphy & DeShon, 2000; Schmidt, Viswesvaran, & Ones, 2000). The sources of variance in scores that are designated as sources of error can differ as a function of (a) the inferences or assertions an investigator wishes to make regarding the scores, (b) how an investigator intends to use the scores (e.g., for relative comparison among applicants or absolute comparison of their scores to some set standard), (c) characteristics of the measurement procedure that produced them, and (d) the nature of the construct one is attempting to measure. Consequently, what is called error, even for scores produced by the same measurement procedure, may legitimately reflect different things under different circumstances. As such, there is no such thing as *the* reliability of scores (just as there is no such thing as *the* validity of scores), and it is possible for many reliability estimates to be calculated that depend on how *error* is being defined by an investigator. Just as if we qualify statements of validity, with statements of "validity for purpose X" or "evidence of validity for supporting inference X," so too must care be taken when discussing reliability with statements such as "scores are reliable with respect to consistency across Y," where Y might refer to items, raters, tasks, or testing occasions, or combinations of them (Putka & Hoffman, 2013, 2015). As we'll see later, different reliability estimates calculated on the same data tell us very different things about the quality of our scores and the degree to which various inferences regarding their consistency are warranted.

A convenient way to start to address these points is to examine how error has come to be operationally defined in the context of estimating reliability. All measurement theories seem to agree that reliability estimation attempts to quantify the expected degree of consistency in scores over replications of a measurement procedure (Brennan, 2001a; Haertel, 2006). Consequently, from the perspective of reliability estimation, error reflects the expected degree of inconsistency between scores produced by a measurement procedure and replications of it. Several elements of these operational definitions warrant further explanation, beginning with the notion of replication. Clarifying these elements will provide an important foundation for the remainder of this chapter.

Replication

From a measurement perspective, replication refers to the repetition or reproduction of a measurement procedure such that the scores produced by each "replicate" are believed to assess the same construct.[2] There are many ways of replicating a measurement procedure. Perhaps the most straightforward way would be to administer the same measurement procedure on more than one occasion, which would provide insight into how consistent scores are for a given person across occasions. However, we are frequently interested in more than whether our measurement procedure would produce comparable scores on different occasions. For example, would we achieve consistency over replicates if we had used an alternative, yet similar, set of items to those that comprise our measure? Answering the latter question is a bit more difficult in that we are rarely in a position to replicate an entire measurement procedure (e.g., construct two or more 20-item measures of conscientiousness and compare scores on each). Consequently, in practice, "parts" or "elements" of our measurement procedure (e.g., items) are often viewed as replicates of each other. The observed consistency of scores across these individual elements is then used

to make inferences about the level of consistency we would expect if our entire measurement procedure was replicated; that is, how consistent would we expect scores to be for a given person across alternative sets of items we might use to assess the construct of interest. The forms of replication described above dominated measurement theory for nearly the first five decades of the 20th century (Cronbach, 1947; Gulliksen, 1950).

Modern perspectives on reliability have liberalized the notion of replicates in terms of (a) the forms that they take and (b) how the measurement facets (i.e., items, raters, tasks, occasisons) that define them are manifested in a data collection design (i.e., a measurement design). For example, consider a measurement procedure that involves having two raters provide ratings for individuals with regard to their performance on three tasks designed to assess the same construct. In this case, replicates take the form of the six rater-task pairs that comprise the measurement procedure, and as such, are multifaceted (i.e., each replicate is defined in terms of specific rater and a specific task). Prior to the 1960s, measurement theory primarily focused on replicates that were defined along a single facet (e.g., replicates represented different items, different split-halves of a test, or the same test administered on different occasions).[3] Early measurement models were not concerned with replicates that were multifaceted in nature (Brown, 1910; Gulliksen, 1950; Spearman, 1910). Modern perspectives on reliability also recognize that measurement facets can manifest themselves differently in any given data collection design. For example, (a) the same raters might provide ratings for each ratee; (b) a unique, nonoverlapping set of raters might provide ratings for each ratee; or (c) sets of raters that rate each ratee may vary in their degree of overlap. As noted later, the data collection design underlying one's measurement procedure has important implications for reliability estimation, which, prior to the 1960s, was not integrated into measurement models. It was simply not the focus of early measurement theory (Cronbach & Shavelson, 2004).

Expectation

A second key element of the operational definition of reliability offered above is the notion of expectation. The purpose of estimating reliability is not to quantify the level of consistency in scores among the sample of replicates that comprise one's measurement procedure for a given study (e.g., items, raters, tasks, or combinations thereof). Rather, the purpose is to use such information to make inferences regarding (a) the consistency of scores resulting from our measurement procedure as a whole with the population from which replicates comprising our measurement procedure were drawn (e.g., the population of items, raters, tasks, or combinations thereof believed to measure the construct of interest) and (b) the consistency of the said scores for the population of individuals from which our sample of study participants was drawn. Thus, the inference space of interest in reliability estimation is inherently multidimensional. As described in subsequent sections, the utility of measurement theories is that they help us make this inferential leap from sample to population; however, the quality with which estimation approaches derived from these theories do so depend on the properties of scores arising from each replicate, characteristics of the construct one is attempting to measure, and characteristics of the sample of one's study participants.

Consistency and Inconsistency

Lastly, the third key element of the operational definition of reliability is the notion of consistency in scores arising from replicates. Defining reliability in terms of consistency of scores implies that error, from the perspective of reliability, is anything that gives rise to inconsistency in scores.[4] Conversely, anything that gives rise to consistency in a set of scores, whether it is the construct we intend to measure or some contaminate source of construct-irrelevant variation that is shared or consistent across replicates, delineates the "true" portion of an observed score from the perspective of reliability. Indeed, this is one reason why investigators are quick to note

that "true score," in the reliability sense of the word, is a bit of a misnomer for the uninitiated—it is not the same as a person's true standing on the construct of interest (Borsboom & Mellenbergh, 2002; Lord & Novick, 1968; Lumsden, 1976). Thus, what may be considered a source of error from the perspective of validity may be considered true score from the perspective of reliability.

Although an appreciation of the distinction between true score from the perspective of reliability and a person's true standing on a construct can be gleaned from the extant literature, there seems to be a bit more debate with regard to the substantive properties of error. The confusion in part stems from a disconnect between the operational definition of error outlined above (i.e., inconsistency in scores across replicates) and hypotheses that measurement theories make regarding the distributional properties of such inconsistencies, which may or may not reflect reality. For example, in the sections above I made no claims with regard to whether inconsistency in scores across replications reflected (a) unexplainable variation that would be pointless to attempt to model, (b) explainable variation that could potentially be meaningfully modeled using exogenous variables as predictors (i.e., measures other than our replicates), or (c) a combination of both of these types of variation. Historically, many treatments of reliability, whether explicitly or implicitly, have equated inconsistency in scores across replicates with "unpredictable" error (e.g., AERA, APA, & NCME, 1999, p. 27). However, nothing in the operational definition of error laid out above necessitates that inconsistencies in scores are unpredictable. Part of the confusion may lie in the fact that we often conceive of replicates as having been randomly sampled from a broader population(s) or are at least representative of some broader population(s) (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Nunnally, 1978; Tryon, 1957). From a statistical perspective, the effects associated with such replicates on scores would be considered random (Jackson & Brashers, 1994), but this does not necessitate that variation in those effects is unexplainable or beyond meaningful prediction, particularly when raters define the replicates of interest (Cronbach et al., 1972; Murphy & DeShon, 2000). Thus, one should be cautious when framing inconsistency in scores as reflecting random errors of measurement because it is often confused with the notion that such errors are beyond meaningful explanation (Ng, 1974).

Summary

This section offered a model-free perspective on error and how it has come to be operationally defined from the perspective of reliability. I adopted this strategy in part because of the confusion noted above but also to bring balance to existing treatments of reliability in the industrial-organizational (I-O) literature, which explicitly or implicitly tends to frame discussions of reliability from the CTT tradition. The language historically used in treatments of CTT makes it difficult for investigators to recognize that inconsistency in scores is not necessarily beyond meaningful explanation, although we conceive of it as random. Another reason I belabor this point is that beginning with Spearman (1904), a legacy of organizational research emerged that focuses more on making adjustment for error in our measures (e.g., corrections for attenuation), rather than developing methods for modeling and understanding error in our measures, which in part may reflect our tendency to view such error as unexplainable.

ROLE OF MEASUREMENT MODELS

The defining characteristic of a measurement model is that it specifies a statistical relationship between observed scores and unobserved components of those scores. Such unobserved components may reflect sources of consistency in scores (across replicates), whereas others may reflect sources of inconsistency. As noted earlier, the utility of measurement models is that they offer a set of hypotheses regarding the composition of observed scores, which, if supported, can allow us to accurately estimate reliability (e.g., reliability coefficients, standard errors of measurement) from a sample of data and apply those estimates to various problems (e.g.,

corrections for attenuation, construction of score bands). To the extent that such hypotheses are not supported, faulty conclusions regarding the reliability of scores may be drawn, inappropriate uses of the reliability information may occur, and knowledge regarding inconsistencies in our scores may be underutilized. In this section, I compare and contrast measurement models arising from two theories that underlie the modern literature on reliability, namely CTT and generalizability theory (G-theory).[5]

The measurement models underlying CTT and G-theory actually share some important similarities. For example, both (a) conceive of observed scores as being an additive function of true score and error components and (b) view true score and error components as uncorrelated. Nevertheless, as discussed below (cf. Generalizability Theory), certain characteristics of G-theory models enable them to be meaningfully applied to a much broader swath of measurement procedures that we encounter in personnel selection relative to the CTT models. Rather than being competing models, it is now commonly acknowledged that CTT models are simply a more restrictive, narrower version of the G-theory model, which is why G-theory is generally viewed as a "liberalization" of CTT (AERA, APA, & NCME, 1999; Brennan, 2006; Cronbach, Rajaratnam, & Gleser, 1963). Nevertheless, given its relatively narrow focus, it is convenient for pedagogical purposes to open with a brief discussion of CTT before turning to G-theory.

Classical Test Theory

Under classical test theory, the observed score (X) for a given person *p* that is produced by replicate *r* of a measurement procedure is assumed to be a simple additive function of two parts: the person's true score (T) and an error score (E).

$$X_{pr} = T_p + E_{pr} \tag{1.1}$$

Conceptually, a person's true score equals the expected value of their observed scores across an infinite set of replications of the measurement procedure. Given that such an infinite set of replications is hypothetical, a person's true score is unknowable but, as it turns out, not necessarily unestimatable (see Haertel, 2006, pp. 80–82). As noted earlier, true score represents the source(s) of consistency in scores across replicates (note there is no "*r*" subscript on the true score component in Equation 1.1)—in CTT it is assumed to be a constant for a given person across replicates. Error, on the other hand, is something that varies from replicate to replicate, and CTT hypothesizes that the mean error across the population of replicates for any given person will be zero. In addition to these characteristics, if we look across persons, CTT hypothesizes that there will be (a) no correlation between true and error score associated with a given replicate ($r_{Tp.Ep1} = 0$), (b) no correlation between error scores from different replicates ($r_{Ep1.Ep2} = 0$), and (c) no correlation between error scores from a given replicate and true scores from another replicate ($r_{Ep1.Tp2} = 0$). Although the CTT score models do not necessitate that error scores from a given replicate (or composite of replicates) be uncorrelated with scores from measures of other attributes, the latter is a key assumption underlying the use of reliability coefficients to correct observed correlations for attenuation (Schmidt & Hunter, 1996; Spearman, 1910). Essentially, this last assumption implies that inconsistency in a measurement procedure will be unrelated to any external variables (i.e., variables other than our replicates) and therefore beyond meaningful prediction. From basic statistics we know that the variance of the sum of two independent variables (such as T and E) will simply equal the sum of their variances; thus, under CTT, observed score variance across persons for a given replicate is simply the sum of true score variance and error variance.

$$\sigma^2_X = \sigma^2_T + \sigma^2_E \tag{1.2}$$

As detailed later, reliability estimation attempts to estimate the ratio of σ^2_T over $\sigma^2_T + \sigma^2_E$, not for a single replicate but rather for a measurement procedure as a whole, which as noted earlier

is often conceived as consisting of multiple replicates. Thus, reliability coefficients are often interpreted as the proportion of observed score variance attributable to true score variance, or alternatively, the expected correlation between observed scores resulting from our measurement procedure and scores that would be obtained had we based our measure on the full population of replicates of interest (i.e., hypothetical true scores).

One of the key defining characteristics of CTT is the perspective it takes on replicates. Recall that earlier I offered a very generic definition for what constitutes a replicate. I described how we often conceive of parts or elements of a measurement procedure as replicates and use them to estimate the reliability of scores produced by our procedure as a whole. As noted later, CTT-based reliability estimation procedures assume that replicates have a certain degree of "parallelism." For example, for two replicates to be considered strictly (or classically) parallel, they must (a) produce identical true scores for a given individual (i.e., T_p for Replicate A = T_p for Replicate B), (b) have identical mean observed scores, and (c) have identical error variances.[6] The commonly used Spearman-Brown prophecy formula is an example of a CTT-based estimation procedure that is based on the assumption that replicates involved in its calculation are strictly parallel (Feldt & Brennan, 1989).

It is often not realistic to expect any two replicates to be strictly parallel. For example, items on a test of cognitive ability are rarely of the same difficulty level, and raters judging incumbents' job performance often differ in their level of leniency/severity. Under such conditions, item means (or rater means) would differ, and thus, such replicates would not be considered strictly parallel. In recognition of this, CTT gradually relaxed its assumptions over the years to accommodate the degrees of parallelism that are more likely to be seen in practice. The work of Lord (1955), Lord and Novick (1968), and Joreskog (1971) lays out several degrees of parallelism, which are briefly reviewed below.

Tau-equivalent replicates produce identical true scores for a given individual but may have different error variances (across persons) and as such different observed variances. Essentially, tau-equivalent replicates relax assumptions further, in that they allow true scores produced by any given pair replicates to differ by a constant (i.e., T_p for Replicate 1 = T_p for Replicate 2 + C, where the constant may differ from pair to pair of replicates). As such, essential tau-equivalence accommodates the situation in which there are mean differences across replicates (e.g., items differ in their difficulty, and raters differ in their leniency/severity). The assumption of essential tau-equivalence underlies several types of coefficients commonly used in reliability estimation, such as coefficient alpha, intraclass correlations, and as discussed in the next section, generalizability coefficients.[7]

One thing that may not be immediately obvious from the description of essential tau-equivalence offered above is that it does not accommodate the situation in which replicates differ in true score variance (across persons). Joreskog's (1971) notion of congeneric test forms (or more generally, congeneric replicates) accommodated this possibility. Specifically, the congeneric model allows true scores produced by a given replicate to be a linear function of true scores from another replicate (i.e., T_p for Replicate 1 = $b \times T_p$ for Replicate 2 + C). As illustrated in the later section on reliability estimation, this accommodates the possibility that replicates may be differentially saturated with true score variance or be measured on a different metric.

The degrees of parallelism discussed above have implications for estimating reliability; more specifically, they have implications for the accuracy of results produced by reliability estimation methods that we apply to any given set of replicates. As discussed later, we can apply nearly any reliability estimation method derived from the classical tradition to any sample of replicates, regardless of their underlying properties; however, the estimate we get will differ in its accuracy depending on (a) the extent to which the underlying properties of those replicates conform to the assumptions above and (b) characteristics of the construct one is attempting to measure. It is beyond the scope of this chapter, and not its intent, to provide a catalog of coefficients that may be appropriate for estimating the reliability depending on the degree of parallelism among the replicates of interest, because excellent descriptions exist elsewhere in the literature (e.g., Feldt & Brennan, 1989, Table 3, p. 115; Raykov & Marcoulides, 2011). However, in reviewing treatments such as the one offered by Feldt and Brennan (1989), be cognizant that the myriad

coefficients they review (including the commonly used Spearman-Brown prophecy and coefficient alpha) were formulated to deal with scores arising from measurement procedures in which (a) replicates were defined by a single facet (e.g., replicates reflect different items or test parts) and (b) that facet was fully crossed with one's objects of measurement (e.g., all test takers are administered the same set of items, and all test takers completed the same test on two different occasions). As we will see below, application of classical reliability estimation methods in cases in which replicates are multifaceted (e.g., replicates representing task-rater pairs) or cases in which the design underlying one's measurement procedure is not fully crossed is problematic (Cronbach & Shavelson, 2004). The treatment of reliability for measurement procedures characterized by multifaceted replicates or involving noncrossed measurement designs leads naturally to the introduction of G-theory.

Generalizability Theory

G-theory liberalizes CTT in that it has mechanisms within its score models for (a) dealing with single-faceted and multifaceted replicates, (b) simultaneously differentiating and estimating multiple sources of error arising from different measurement facets (e.g., items, raters, occasions, tasks), (c) dealing with scores produced by a wide variety of data collection designs (e.g., crossed, nested, and ill-structured measurement designs), (d) adjusting the composition of true score and error depending on the generalizations one wishes to make regarding the scores, (e) adjusting the composition of true score and error depending on how one intends to use the scores (e.g., for relative comparison among applicants or absolute comparison of their scores to some set standard), and (f) relaxing some of the assumptions put on the distributional properties of true and error components proscribed under CTT. The purpose of this section will be to elaborate these features of G-theory model in a way that is relatively free from G-theory jargon, which has been cited as one reason why the unifying perspective that G-theory offers on reliability has yet to be widely adopted by organizational researchers (DeShon, 2002).

Perhaps the most visible way G-theory model liberalizes the CTT model is its ability to handle measurement procedures comprising multifaceted replicates. To illustrate this key difference between the G-theory and CTT models, let us first consider an example in which we have observed scores based on ratings of job applicants' responses to three interview questions designed to assess interpersonal skill. Say that we had the same three raters interview each applicant and that each rater asked applicants the same three questions (i.e., applicants, raters, and questions are fully crossed). Thus, we have nine scores for each applicant—one for each of our nine "replicates," which in this case are defined by unique question-rater combinations. Under CTT and G-theory, we might conceive of an applicant's true score as the expected value of his/her observed score across the population of replicates—in this case it is the population of raters and questions. However, if we were to apply the CTT score model to such replicates, it would break down because it does not account for the fact that some replicates share a rater in common and other replicates share a question in common. As such, the error associated with some replicates will be correlated across applicants, therefore violating one of the key assumptions underlying CTT measurement model (i.e., errors associated with different replicates are uncorrelated). As shown below (cf. Equation 1.4), the G-theory measurement model permits the addition of terms to the model that account for the fact that replicates are multifaceted. The insidious part of this illustration is that the situation above would not prevent us from applying estimation methods derived from CTT to these data (e.g., calculating coefficient alpha on the nine replicates). Rather, perhaps unbeknownst to the investigator, the method would allocate error covariance among replicates that share a rater or question in common to true score variance because they are a source of consistency across at least some of the replicates (Komaroff, 1997; Raykov, 2001a). That is, the CTT score model and commonly used coefficients derived from it (e.g., coefficient alpha) are blind to the possibility of multifaceted replicates, which is a direct reflection of the fact that early measurement theory primarily concerned itself with fully crossed, single-faceted measurement designs (Cronbach & Shavelson, 2004).

To account for the potential that replicates can be multifaceted, G-theory formulates its measurement model from a random-effects ANOVA perspective. Unlike CTT, which has its roots in the correlational research tradition characteristic of Spearman and Pearson, G-theory has its roots in the experimental research tradition characteristic of Fisher (1925). As such, G-theory is particularly sensitive to dealing with replicates that are multifaceted in nature and both crossed and noncrossed measurement designs. It has long been acknowledged that issues of measurement design have been downplayed and overlooked in the correlational research tradition (Cattell, 1966; Cronbach, 1957), and this is clearly evident in reliability estimation approaches born out of CTT (Cronbach & Shavelson, 2004; Feldt & Brennan, 1989). To ease into the G-theory measurement model, I start with a simple example, one in which we assume observed scores are generated by a replicate of a measurement procedure that is defined along only one facet of measurement. The observed score (X) for a given person *p* that is produced by any given replicate defined by measurement facet "A" (e.g., "A" might reflect items, occasions, raters, tasks, etc.) is assumed to be an additive function:

$$X_{pa} = b_0 + u_p + u_a + u_{pa} + e_{pa} \tag{1.3}$$

where b_0 is the grand mean score across persons and replicates of facet A; u_p is the main effect of person *p* and conceptually the expected value of *p*'s score across the population of replicates of facet A (i.e., the analogue of true score); u_a represents the main effect of replicate *a* and conceptually is the expected value of *a*'s score across the population of persons; u_{pa} represents the $p \times a$ interaction effect and conceptually reflects differences of the rank ordering of persons across the population of replicates of facet A; and lastly, e_{pa} is the residual error that conceptually is left over in X_{pa} after accounting for the other score effects.[8] As with common random-effects ANOVA assumptions, these score effects are assumed to (a) have population means of zero, (b) be uncorrelated, and (c) have variances of σ^2_P, σ^2_A, σ^2_B, σ^2_{AB}, and $\sigma^2_{Residual}$, respectively (Jackson & Brashers, 1994). The latter variance components are the focus of estimation efforts in G-theory, and they serve as building blocks of reliability estimates derived by G-theory.

Of course, the example above is introduced primarily for pedagogical purposes; the real strength of the random-effects formulation is that the model above is easily extended to measurement procedures with multifaceted replicates (e.g., replicates that reflect question-rater pairs). For example, the observed score (X) for a given person *p* that is produced by any given replicate defined by measurement facets "A" and "B" (e.g., "A" might reflect questions and "B" might reflect raters) is assumed to be an additive function.

$$X_{pab} = b_0 + u_p + u_a + u_b + u_{pa} + u_{pb} + u_{ab} + u_{pab} + e_{pab} \tag{1.4}$$

A key difference to point out between the models specified in Equations 1.3 and 1.4 is the interpretation of the main effects for individuals. Once again, u_p is the main effect of person *p*, but conceptually it is the expected value of *p*'s score across the population of replicates defined by facets A and B. Thus, although the u_p term in Equations 2.3 and 2.4 provides an analogue to the true score, the substance of true scores differs depending on the nature of the population(s) of replicates of interest. Extending this model beyond two facets (e.g., a situation in which replicates are defined as a combination of questions, raters, and occasions) is straightforward and simply involves adding main effect terms for the other facets and associated interaction terms (Brennan, 2001b).

One thing that is evident from the illustration of the G-theory model provided above is that, unlike the CTT model, it is scalable; that is, it can expand or contract depending on the degree to which replicates underlying a measurement procedure are faceted. Given its flexibility to expand beyond simply a true and error component, the G-theory model potentially affords investigators with several more components of variance to consider relative to the CTT model. For example, using the interview example presented above, we could potentially decompose variance in interview scores for applicant *p* on question *q* as rated by rater *r* into seven components.[9]

$$\sigma^2_X = \sigma^2_P + \sigma^2_Q + \sigma^2_R + \sigma^2_{PQ} + \sigma^2_{PR} + \sigma^2_{QR} + \sigma^2_{PQR,\ Residual} \tag{1.5}$$

Recall from the earlier discussion of CTT that the basic form reliability coefficients take on is $\sigma^2_T/(\sigma^2_T + \sigma^2_E)$. This fact begs the question, from the G-theory perspective, what sources of variance comprise σ^2_T and σ^2_E? As one might guess from the decomposition above, the G-theory model offers researchers a great deal of flexibility when it comes to specifying what constitutes error variance and true score variance in any given situation. As demonstrated in the following sections, having this flexibility is of great value. As alluded to in the opening paragraph of this section, the sources of variance in scores that are considered to reflect error (and true score for that matter) can differ depending on (a) the generalizations an investigator wishes to make regarding the scores, (b) how an investigator intends to use the scores (e.g., for relative comparison among applicants or absolute comparison of their scores to some set standard), and (c) characteristics of the data collection or measurement design itself, which can limit an investigator's ability to estimate various components of variance. The idea of having flexibility of specifying what components of observed variance contribute to true score and error is something that is beyond the CTT score model because it only partitions variance into two components. The following sections highlight how the G-theory model offers investigators flexibility for tailoring the composition of σ^2_T and σ^2_E to their situation.

Dependency of σ^2_T and σ^2_E on Desired Generalizations

The decision of what components of variance comprise σ^2_T and σ^2_E depends in part on the generalizations the investigator wishes to make based on the scores. To illustrate this, let us take the interview example offered above and say that the investigator was interested in (a) generalizing scores from his or her interview across the population of questions and raters and (b) using the scores to make relative comparisons among applicants who completed the interview. In such a case, variance associated with applicant main effects (σ^2_P) would comprise σ^2_T, and variance associated with interactions between applicants and each type of measurement facet (i.e., applicant-question interaction variance, σ^2_{PQ}; applicant-rater interaction variance, σ^2_{PR}; and applicant-question-rater interaction variance and residual variance, $\sigma^2_{PQR,Residual}$) would comprise σ^2_E. The relative contribution of these latter effects to error variance would be scaled according to the number of questions and raters involved in the measurement procedure. As the number of questions increases, the contribution of σ^2_{PQ} would go down (i.e., error associated with questions would be averaged away), and as the number of raters increases, the contribution of σ^2_{PR} would go down (i.e., error associated with raters would be averaged away). Specifically, the "generalizability" coefficient described above would be

$$E\rho^2 = \frac{\sigma^2_P}{\sigma^2_P + \left[\frac{\sigma^2_{PQ}}{n_Q} + \frac{\sigma^2_{PR}}{n_R} + \frac{\sigma^2_{PQR,\,Residual}}{n_Q n_R}\right]} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E} \tag{1.6}$$

where the term in brackets represents σ^2_E, n_Q is the number of interview questions, n_R is the number of raters, and $n_Q n_R$ is the product of the number of questions and raters.[10] Note that increasing the number of questions and/or raters will result in decreasing that part of error associated with questions and/or raters, respectively. The idea that G-theory allows for the scaling of these effects as a function of the number of questions and raters sampled is analogous to the role of the Spearman-Brown prophecy in CTT, in which the number of replicates that comprise a measurement procedure directly affects the estimated reliability of scores produced by that procedure (Feldt & Brennan, 1989). The key difference here is that G-theory allows one to differentiate and examine the effect that adjusting the sampling of different types of facets has for reliability (e.g., separately adjusting the number of questions and raters), whereas the Spearman-Brown prophecy does not allow such differentiation to occur. As such, applying the Spearman-Brown prophecy to estimate what the reliability of scores would be if the length of a measure is changed can greatly mislead investigators if the replicates that comprise that measure are multifaceted (Feldt & Brennan, 1989).

To illustrate, let us take the interview example offered above and say $\sigma^2_P = .50$, $\sigma^2_{PQ} = .30$, $\sigma^2_{PR} = .10$, and σ^2_{PQR},Residual = .10. Recall our interview comprises three questions and three raters (i.e., nine question-rater pairs serve as replicates). Using Equation 1.6, the estimated reliability of the average rating across questions and raters would be .78 ($\sigma^2_T = .50$, $\sigma^2_E = .14$). Now, if we were to ask what effect "doubling the length of the interview" would have on reliability, and we used the Spearman-Brown prophecy (i.e., $2E\rho^2/[1 + E\rho^2]$) to answer that question, we would achieve an estimate of .88, which is analogous to what we achieve if we replaced n_Q, n_R, and $n_Q n_R$ in Equation 1.6 with $2n_Q$, $2n_R$, and $2n_Q 2n_R$. Note that the Spearman-Brown prophecy does not provide the estimated reliability for 18 question-rater pairs (i.e., double the existing number of replicates), but rather an estimated reliability for 36 question-rater pairs (i.e., six questions × six raters). As such, in this case, the Spearman-Brown formula gives us an estimated reliability if the effective length of the interview were quadrupled rather than doubled. Another shortcoming of the Spearman-Brown formula is that it fails to account for the fact that there are multiple ways one can effectively double the length of the interview, each of which may produce a different reliability estimate. For example, we can have two questions and nine raters, which would give us 18 question-rater pairs and result in an average rating reliability of .75 on the basis of Equation 1.6. Alternatively, we can have nine questions and two raters, which would also give us 18 question-rater pairs but result in an average rating reliability of .85 on the basis of Equation 1.6. Essentially, there is no mechanism within the Spearman-Brown formula that accounts for the fact that facets may differentially contribute to error. As this example illustrates, making adjustments to the number of levels sampled for one facet (e.g., questions in this case) may have a much more profound effect on error than making adjustments to the number of levels sampled for other facets (e.g., raters) included in the design.

Returning to the discussion of the dependency of σ^2_T and σ^2_E on the generalizations one wishes to make regarding their scores, let us now say that a different investigator uses the same interview procedure described above, but instead only wished to generalize scores from the procedure across the population of raters. For example, this might be the case if the investigator feels that the questions get at different parts of the interpersonal skill construct, and as such does not wish to treat inconsistency in scores across questions (for a given applicant) as error. In such a case, variance associated with applicant main effects (σ^2_P) and a function of applicant-question interaction effects (σ^2_{PQ}) would comprise σ^2_T, and variance associated with interactions between applicants and raters (σ^2_{PR}) and the applicant-rater-questions along with residual error applicant ($\sigma^2_{PQR,Residual}$) would comprise σ^2_E (Brennan, 2001b; DeShon, 2002). In this situation, the investigator is essentially examining the consistency of scores across raters on the basis of ratings that have been averaged across the three interview questions—in G-theory this is known as fixing a facet of measurement.[11]

Dependency of σ^2_T and σ^2_E on Intended Use of Scores

Slightly varying the example above allows for illustration of the implications of how an investigator intends on using scores for the sources of variance that contribute to σ^2_T and σ^2_E. For example, let us say the interview above was conducted to determine if applicants met some minimum level of interpersonal skill. That is, rather than comparing applicants against one another, the interest is in comparing their scores to some standard of interpersonal skill. Also, let us return to the original example in which the investigator was interested in generalizing scores across the population of questions and raters. In this case, variance due to the main effects of questions and raters, as well as their interaction (i.e., σ^2_Q, σ^2_R, σ^2_{QR}), would contribute σ^2_E (in addition to sources identified earlier, σ^2_{PQ}, σ^2_{PR}, $\sigma^2_{PQR,Residual}$) because they influence the absolute magnitude of the score any given applicant receives. In the example from the previous paragraphs in which we were only interested in using scores to make relative comparisons among applicants, these effects did not contribute to error because they have no bearing on how applicants were rank ordered (i.e., question and rater main effects are constants across applicants for designs in which questions and raters are fully crossed with applicants). The potential for

such effects to contribute to σ^2_E in crossed designs (as they do in this example) is not addressed by CTT, because it is simply beyond the scope of the CTT model to handle error of that type (Cronbach & Shavelson, 2004).

Dependency of σ^2_T and σ^2_E on Characteristics of the Measurement Procedure

Critics may argue that the interview examples offered above do not reflect the reality of measurement designs faced in applied organizational research and practice. Such critics would be right. Rarely, if ever, are the measurement designs involving ratings that we confront in the applied organizational research and practice fully crossed. When we are fortunate to have two or more raters for each ratee, the orientation of raters to ratees is often what Putka, Le, McCloy, and Diaz (2008) have termed "ill-structured".[12] Specifically, the sets of raters that rate each ratee are neither identical (indicative of a fully crossed design) nor completely unique (indicative of a design in which raters are nested with ratees); rather, each ratee is rated by a set of raters that may vary in their degree of overlap. The implications of the departure of measurement designs from the fully crossed ideal is that it can limit our ability to uniquely estimate the components of variance that underlie observed scores (e.g., those illustrated in Equation 1.5), which in turn limits our flexibility for choosing which components contribute σ^2_T and σ^2_E. To illustrate this, let's consider a few variants on the interview example above.

Say that instead of having three raters rate each applicant on each interview question, a different nonoverlapping set of three raters rates each applicant (i.e., raters are nested within applicants). In this case, rater main effect variance (σ^2_R) and applicant-rater interaction effect variance (σ^2_{PR}) would be inseparable, and both will contribute to σ^2_E regardless of whether the investigator was interested in using the scores simply to rank order applicants or compare applicants' scores to some fixed standard (McGraw & Wong, 1996; Shrout & Fleiss, 1979). However, often in practice we are not dealt such nested designs—the sets of raters that may rate each ratee tend to vary in their degree of overlap. Although less "clean" than the aforementioned nested design, having some degree of overlap actually gives us an opportunity to uniquely estimate σ^2_R and σ^2_{PR} (as we are able to do in a fully crossed design) (Putka et al., 2008). Nevertheless, as was the case with the nested design, σ^2_R and σ^2_{PR} will contribute to σ^2_E, because the raters that rate each ratee are not identical, σ^2_R and σ^2_{PR} will affect the rank ordering of ratees' scores (Schmidt et al., 2000). However, unlike the nested design, the contribution of σ^2_R to σ^2_E will be dependent on the amount of overlap between the sets of raters that rate each ratee—a subtlety not widely known but pertinent to many organizational researchers who work with ratings (Putka et al., 2008).

Lastly, let's use the previous interview example one more time to provide a critical insight offered by G-theory—the notion of hidden measurement facets and their implications for interpreting the substantive nature of σ^2_T and σ^2_E. In laying out the interview example above, it was implicit that raters conducted interviews on separate occasions. However, a more common situation might be that raters sit on a panel, and as such the three questions are asked of a given applicant on the same occasion. In either case, we have measurement procedures with designs that are "notationally" identical (i.e., applicants × questions × raters); however, the variance components underlying scores produced by these interview procedures have different substantive meanings. If each rater conducted a separate interview, variance attributable to the applicant-rater interaction (σ^2_{PR}) would also reflect applicant-occasion variance (σ^2_{PO}). In other words, σ^2_{PR} would not only reflect inconsistencies in raters' rank ordering of applicants, but also inconsistency in the applicants' responses across the occasions on which the interviews were conducted. If the applicant participated in the panel interview, raters would be rating the applicants' responses on the same occasion, and as such variance attributable to the applicant-rater interaction (σ^2_{PR}) would be just that, but variance attributable to applicant main effects (σ^2_P) would also reflect applicant-occasion variance (σ^2_{PO}). This stems from the fact that raters are observing a given applicant on the same occasion, and as such occasion of measurement serves as a source of consistency in raters' ratings that would not be present if raters conducted separate interviews. In both of the examples above, σ^2_{PO} is not separable from the other source of variance with which it is confounded. In the case of separate interviews, raters covary with occasions; in the case of panel

interviews, occasions are not replicated for a given applicant. Thus, these examples illustrate how a measurement facet can hide in different ways to influence the substantive meaning of σ^2_E (in the case of the separate interview) and σ^2_T (in the case of the panel interviews).

The examples above also illustrate an important point—just because we cannot isolate or estimate a source of variance underlying observed scores does not mean those sources of variance are not present and influencing our scores (Brennan, 2001b; DeShon, 1998; Feldt & Brennan, 1989; Schmidt & Hunter, 1996). Indeed, it is interesting to take the concept of hidden facets and use them to frame some common measurement issues in personnel selection. For example, the magnitude of person-rater interaction variance (σ^2_{PR}) in job performance ratings has been found to be quite large (e.g., Schmidt et al., 2000; Scullen et al., 2000). However, if raters are viewing the performance of individuals on (a) different occasions, (b) different tasks, and/or (c) different tasks on different occasions, then part of what we typically label person-rater interaction variance may actually also reflect several other sources of variance (e.g., person-occasion interaction variance, person-task interaction variance, and person-task-occasion interaction variance). In other words, the hidden facets of occasion and task might help explain the sizable person-rater interaction effects often found in job performance ratings. In the context of assessment centers, hidden facets might partially explain the common finding of the dominance of exercise effects over dimension effects (Lance, 2008). For example, dimensions within exercises share an occasion of measurement in common (and sometimes share raters as well), whereas dimensions in different exercises do not. As such, all else being equal we would expect scores for dimensions within exercises to be more consistent with each other than with scores for dimensions in different exercises. Thus, what is interpreted as an exercise effect in the context of assessment center ratings may partially be explained by hidden occasion and rater facets of measurement that increase consistency among dimension scores within exercises relative to dimension scores across exercises (e.g., Cronbach, Linn, Brennan, & Haertel, 1997). These examples illustrate the potential utility of framing common measurement issues through the lens of hidden facets illuminated by G-theory.

Summary

This section described perspectives on observed scores adopted by two measurement theories that dominate current discussions of reliability. Through a single example, I illustrated the many ways in which G-theory liberalizes not only the score model offered by CTT but also the perspective it offers on reliability. By no means did the discussion fully illustrate how G-theory is applied or how reliability coefficients based on G-theory are calculated. For such details, the reader is referred to other treatments (Brennan, 2001b; DeShon, 2002; Haertel, 2006; Shavelson & Webb, 1991). Nor did this section illustrate how *very* different conclusions regarding the reliability of scores can be depending on (a) the generalizations an investigator wishes to make regarding those scores, (b) how an investigator intends to use those scores (e.g., for relative comparison among applicants or absolute comparison of their scores to some set standard), or (c) the measurement design the investigator uses to gather data that gives rise to the scores (see Putka & Hoffman [2013, 2015] for concrete illustrations of the consequences of these decisions for the magnitude of reliability estimates for assessment center and job performance ratings, respectively). Nevertheless, given space constraints, this was not my intent. Rather, I tried, in a way that was relatively free of G-theory jargon, to show how G-theory offers a way for framing and dealing with measurement situations that CTT was designed to handle, as well as those that CTT was never really designed to handle—a key reason why G-theory currently underlies modern perspectives on reliability (Cronbach & Shavelson, 2004).

ESTIMATION OF RELIABILITY

The previous sections outlined conceptual and model-based perspectives on reliability and measurement error. This section addresses how these concepts and models translate into methods for estimating reliability. Reliability is often summarized in terms of (a) coefficients ranging

from 0 to 1 or (b) standard errors of measurement (SEMs) expressed in a raw score metric. My focus is on the former, partly out of page limits and partly because the latter can typically be calculated from components of the former.[13] As noted earlier, under CTT and G-theory the goal of reliability estimation is to estimate the ratio σ^2_T / (σ^2_T + σ^2_E). The following sections discuss methods for estimating this ratio and components of it. My intent here is not to provide a catalog of different types of reliability coefficients, nor is my intent to provide a cookbook on how to estimate reliability in any given situation. Indeed, as should be clear from the previous section, doing so would not be fruitful given that the composition of σ^2_T and σ^2_E in any given situation partly reflects the aims of the individual investigator. Rather, I focus on comparing and contrasting different historical traditions on estimating reliability, examining the pros and cons of each, and speaking to their equifinality under certain conditions.

The extant literature on reliability estimation is characterized by a multitude of loosely organized coefficients and estimation methods. Historically, the psychometric literature tended to organize discussions of reliability estimation in terms of categories or types of reliability (e.g., test-retest reliability, split-half, parallel-forms, coefficients of equivalence, stability, precision; Cronbach, 1947; Gulliksen, 1950). With the advent of G-theory, psychometricians have slowly gravitated away from categories or types of coefficients that characterized early test theory because "the categories may now be seen as special cases of a more general classification, generalizability coefficients" (AERA et al., 1999, p. 27). As Campbell (1976) noted, the G-theory model "removes the somewhat arbitrary distinctions among coefficients of stability, equivalence, and internal consistency and replaces them with a general continuum of representativeness" (p. 202). Interestingly, this movement toward a unitarian perspective on reliability has temporally coincided with the movement from trinitarian to unitarian perspectives on validity (Brennan, 2006). Ironically, unlike our views on validity, our formal treatments of reliability estimation in organizational research have remained focused on categories or types of reliability coefficients (e.g., Aguinis, Henle, & Ostroff, 2001; Guion, 1998; Le & Putka, 2007; Ployhart, Schnider, & Schmitt, 2006; Schmidt & Hunter, 1996).[14] Rather than continuing to bemoan the current state of affairs, I offer an alternative way of framing discussions of estimating reliability that may help bring organizational research, practice, and pedagogy more in line with modern psychometric thought. Before doing so, I offer a quick example to help illustrate the rationale behind the structure offered below.

When calculating existing types of reliability coefficients, such as a simple Pearson correlation calculated between two replicates (Brown, 1910; Spearman, 1910), coefficient alpha (Cronbach, 1951), or intraclass correlation (ICC; Shrout & Fleiss, 1979)—with which most investigators are familiar—it is important to remember that these are just sets of mathematical operations that can be applied to any set of replicates of our choosing (e.g., raters, items, tasks, occasions). They will all produce, to varying degrees of quality (depending on the properties of the underlying data and construct being measured), estimates of the ratio $\sigma^2_T/(\sigma^2_T + \sigma^2_E)$. As noted above, the substantive meaning of σ^2_T and σ^2_E will depend in large part on the types of replicates to which the mathematical operations are applied. For example, if we apply them to replicates defined as items, σ^2_T will reflect consistency across items; if we apply them to replicates defined as occasions, σ^2_T will reflect consistency across occasions; if we apply them to replicates defined as raters, σ^2_T will reflect consistency across raters; and so on and so forth.[15] Unfortunately, our literature has a tendency to associate certain types of coefficients with certain types of replicates (e.g., coefficient alpha with items, ICCs with raters). This is unfortunate and misleading, because this simply reflects the type of replicate with which these procedures happened to be introduced by earlier authors. Computationally, the procedures are blind to the types of replicates to which they are applied, and many are algebraically identical (Cronbach & Shavelson, 2004; Feldt & Brennan, 1989). For example, alpha is a specific type of ICC, and all ICCs can be framed as generalizability coefficients (Brennan, 2001b; McGraw & Wong, 1996). The following discussion is organized around three traditions for estimating reliability. The classical tradition largely attempts to estimate reliability directly, with little attention toward estimating components of it. More modern traditions (e.g., those based on random-effects models and CFA models) attempt to generate estimates of σ^2_T and σ^2_E, or components of them, which gives investigators flexibility to combine components in different ways to calculate reliability estimates appropriate for

their situation and achieve a better understanding of the sources of error (and true score) in their measures.

Classical Tradition

This classical tradition has its roots in using Pearson correlation between two replicates (e.g., split-halves of a single test, tests administered on two different occasions) to estimate reliability (Brown, 1910; Spearman, 1910). It is based on the premise that the correlation between two strictly parallel replicates (e.g., split-halves of a test, the same test administered on two occasions) equals the proportion of observed score variance attributable to true scores from a single replicate. If applied to split-halves of a test, the Spearman-Brown prophecy formula would then be used to "step-up" the said correlation to arrive at an estimate of reliability for scores produced by the full test. The primary strength of estimation methods based on this tradition is their simplicity and widespread familiarity. Pearson correlations are easy to calculate and widely used in selection research and practice (Schmidt & Hunter, 1996).

Early psychometricians realized that the Spearman-Brown approach described above becomes unwieldy in situations dealing with more than two replicates (e.g., a 10-item conscientiousness scale). Specifically, they realized that depending on which split-halves of their test they calculated their correlation on, they would get a different estimate of reliability (Kuder & Richardson, 1937). In light of this difficulty, researchers developed alternative approaches to estimating reliability that were a function of replicate variances (e.g., item variances) and observed score variances (e.g., variance of the full test). These approaches provided a computationally simple solution that could easily accommodate measures involving two or more single-faceted replicates and are reflected in Kuder and Richardson's (1937) KR-20, Guttman's (1945) set of lambda coefficients, and Cronbach's coefficient alpha (Cronbach, 1951).[16, 17] Another positive characteristic of these latter approaches relative to the Spearman-Brown prophecy is that they only necessitate replicates be essentially tau-equivalent, as opposed to strictly parallel (Novick & Lewis, 1967), although subsequent research has found that alpha is robust to violations of essential tau-equivalence (Haertel, 2006).

Unfortunately, all of the classical estimation approaches described above, from Spearman-Brown through coefficient alpha, are limited in some important ways. As noted earlier, the CTT model on which these coefficients are based was developed for use with measurement procedures involving single-faceted replicates that were fully crossed with one's objects of measurement (Cronbach & Shavelson, 2004). The simplicity of calculating a Pearson *r*, the Spearman-Brown prophecy, and alpha belies interpretational and statistical problems that arise if one attempts to apply them to replicates that are (a) not fully crossed with one's objects of measurement or (b) multifaceted in nature. As Cronbach and Shavelson (2004) noted in their discussion of the possibility of applying alpha to replicates that are not fully crossed, "Mathematically, it is easy enough to substitute scores from a nested sample matrix by simply taking the score listed first for each (person) as belonging in Column 1, but this is not the appropriate analysis" (p. 400). Nevertheless, application of such classical estimation methods, regardless of a procedure's underlying design, has been common practice in organizational research (Viswesvaran, Schmidt, & Ones, 2005).

To illustrate the problems that arise when classical estimation methods are applied to measurement designs that are not fully crossed, consider an example in which job incumbents are each rated by two raters on their job performance. Some incumbents may share one or more raters in common, whereas others may share no raters in common. In this case, standard practice is to (a) randomly treat one rater for each ratee as "rater 1" and the other as "rater 2," (b) assign the ratings of "rater 1" to column 1 and the ratings of "rater 2" to column 2 in a data set, (c) calculate the Pearson correlation between columns to estimate the reliability of a single-rater's ratings, and then (d) use the Spearman-Brown prophecy on the said correlation to estimate the reliability for the average rating (Viswesvaran et al., 2005). Putka et al. (2008) elaborated on several problems with this common practice, namely (a) the estimates derived from this process can differ depending on the assignment of raters to columns 1 and 2 for each ratee; (b) Pearson

r fails to account for the fact that residual errors are nonindependent for ratees who share one or more raters in common, which leads to a downward bias in estimated true score variance (σ^2_T) (Kenny & Judd, 1986); and (c) the Spearman-Brown prophecy inappropriately scales the contribution of rater main effect variance to error variance (σ^2_E) as a function of the number of raters per ratee, rather than the amount of overlap between sets of raters that rate each ratee, leading to an overestimate of σ^2_E (see also Brennan, 2001b, p. 236). In addition to these points, Putka and his colleagues offer a solution for dealing with this type of design that is based on the random-effects model tradition of estimating reliability (discussed later).

Second, with regard to the problem of applying classical methods to multifaceted replicates, the task-rater example presented earlier clearly showed the hazards of blindly applying alpha to replicates of such nature. However, it would be a fallacy to suggest that investigators who adopt classical methods would actually apply alpha or other classical methods in such a manner. Indeed, early psychometricians seemed acutely aware of the limitations of the CTT model, and they attempted to deal with the inability of the CTT model to account for multifaceted replicates by calculating different types of coefficients. For example, Cronbach (1947) discussed the coefficient of equivalence and stability (CES), which was calculated by correlating two different forms of a measure completed by the same respondents on two different occasions (i.e., replicates defined by form-occasion combinations). Cronbach later realized that emergence of the G-theory score model in the 1960s eliminated the need to "mix and match" pairs of replicates like this and provided a generalized solution that applied regardless of whether one was dealing with single-faceted or multifaceted replicates and regardless of whether one was dealing with crossed or noncrossed designs (Cronbach & Shavelson, 2004).

Although the tradition of using coefficients such as CES to deal with multifaceted replicates has faded in psychometrics, it has continued to characterize organizational research and practice, because we have continued to frame problems of reliability in a way that, for better or worse, resembles the psychometric literature of the 1940s. For example, Schmidt and his colleagues have demonstrated how, in the context of fully crossed designs, one can calibrate different sources of error in scores (e.g., error arising from inconsistencies across items, occasions, raters, etc.) through the addition and subtraction of Pearson correlations applied to different types of replicates (Schmidt et al., 2000). Indeed, for fully crossed designs, Schmidt and others illustrated how one can arrive at estimates for at least some of the variance components estimable based on the random-effects model underlying G-theory (Le, Schmidt, & Putka, 2009; Schmidt, Le, & Ilies, 2003). However, it is important to note that the calibration methods based on the classical coefficients alluded to above will not be able to estimate all components of variance that a given measurement design may support estimating, even if the design is fully crossed. For example, such methods cannot be used to estimate the unique contribution of facet main effects (e.g., rater main effects, question main effects) or interactions among facets (e.g., question-rater effects). Lacking this flexibility is unfortunate, particularly if one is interested in (a) comparing scores to standards (e.g., cutoff score) rather than simply making relative comparisons among individuals or (b) simply gaining a more comprehensive understanding of the sources of variance underlying scores. Remember that the CTT score model that gave rise to the classical coefficients discussed above was never designed to account for main effects of measurement facets, largely because they were assumed not to exist (e.g., recall parallel measures have equal means) and because they were not of interest in the problems that Spearman and other early psychometric researchers concerned themselves with (Cronbach & Shavelson, 2004).

Random-Effects Model Tradition

If one has a measurement procedure involving multifaceted replicates, or the design that underlies the procedure is something other than fully crossed, a natural choice for estimating reliability is based on variance components generated by fitting a random-effects model to one's data (Jackson & Brashers, 1994; Searle et al., 1992). The modern random-effects model has its root in the work of Fisher's early work on the ANOVA model and ICCs (Fisher, 1925). Work by Hoyt (1941) and Ebel (1951) provided early examples of using the ANOVA framework for estimating

reliability for single-faceted replicates. Of particular note was Ebel's (1951) work on ratings in which he dealt with crossed and nested measurement designs. This early work branched in two directions, one that manifested itself in today's literature on ICCs (e.g., McGraw & Wong, 1996; Shrout & Fleiss, 1979) and the other that developed into G-theory (Cronbach et al., 1972). Although rarely acknowledged in the ICC literature on reliability estimation, G-theory encompasses that literature. ICCs and reliability coefficients produced under G-theory (i.e., G-coefficients) are nothing more than ratios of variance components; for example, $\sigma^2_T/(\sigma^2_T + \sigma^2_E)$. G-theory simply acknowledges that these ICCs can take on many more forms than those discussed by McGraw and Wong (1996) and Shrout and Fleiss (1979), and, per the earlier discussion on G-theory, offers a comprehensive framework for constructing a reliability estimate that is appropriate given one's situation.

As alluded to in the earlier treatment of G-theory, when it was originally developed, the primary approach of estimating variance components that contributed to σ^2_T and σ^2_E was the random-effects ANOVA model. This same approach to estimating variance components underlies the modern literature on ICCs (e.g., McGraw & Wong, 1996). Unfortunately, estimating variance components using ANOVA-based procedures can be an arduous process that until recently and without highly specialized software involved numerous manipulations of the sums of squares resulting from ANOVA tables (e.g., Cronbach et al., 1972; Shavelson & Webb, 1991). Relative to calculating coefficients arising from the classical tradition, the difference in simplicity of estimating reliability could be substantial. Indeed, this may be a large reason why G-theory never gained traction among organizational researchers. However, since the 1960s several advances in random-effects models have made estimation of variance components much simpler and resolved many problems associated with ANOVA-based estimators of variance components (DeShon, 1995; Marcoulides, 1990; Searle et al., 1992). Unfortunately, this knowledge has been slow to disseminate into the psychometric and I-O literature, because many still seem to equate G-theory with ANOVA-based variance component estimation procedures that characterized G-theory upon its introduction to the literature.

Procedures for the direct estimation of variance components that underlie all reliability coefficients are now widely available in common statistical packages (e.g., SAS, SPSS, R) and allow investigators to estimate variance components with a few clicks of a button. DeShon (2002) and Putka and McCloy (2008) provided clear examples of the ease with which variance components can be estimated within SAS and SPSS. As such, modern methods of variance component estimation are far easier to implement than (a) procedures characteristic of the early G-theory literature and (b) the calibration techniques discussed by Schmidt et al. (2000), which would require an investigator to engage in a series of manipulations with various types of coefficients arising out of the classical tradition. In addition to offering parsimony, modern methods of variance component estimation have another key advantage: they can readily deal with missing data and unbalanced designs characteristic of organizational research (DeShon, 1995; Greguras & Robie, 1998; Marcoulides, 1990; Putka et al., 2008). In contrast, ANOVA-based variance component estimators characteristic of the early G-theory literature are not well equipped to handle such messy designs. Indeed, when confronted with such designs, advocates of G-theory have often suggested discarding data to achieve a balanced design for purposes of estimating variance components (e.g., Shavelson & Webb, 1991)—with modern methods of variance component estimation, the need for such drastic steps has subsided. The most notable drawback of modern methods of variance component estimation—largely based on full or restricted maximum likelihood—is that they can involve rather substantial memory requirements for large measurement designs (Bell, 1985; Littell, Milliken, Stroup, & Wolfinger, 1996). In some cases, such requirements may outstrip the memory that Windows-based desktop computers can currently allocate to programs for estimating variance components (e.g., SAS and SPSS).

The strengths of reliability estimation methods based on the random-effects model tradition relative to the classical tradition are substantial. First, they fully encompass classical methods in that they can be used to estimate reliability for measurement procedures involving single-faceted replicates that are fully crossed with one's object of measurement. Second, unlike classical methods, they can easily be used to formulate reliability estimates for measurement procedures involving multifaceted replicates in which the facets are crossed, nested, or any

combination thereof. Third, the random-effects tradition provides investigators with not only coefficients but also the variance components that underlie them. As Cronbach and Shavelson (2004) stated: "Coefficients (reliability) are a crude device that do not bring to the surface many subtleties implied by variance components" (p. 394). Variance components allow researchers to get a much finer appreciation of what comprises error than simply having one omnibus estimate of error. Readers interested in learning more about formulation of reliability estimates via variance components estimated by random-effects models—or more generally, G-theory—are referred to DeShon (2002), Haertel (2006), Putka et al. (2008), and Shavelson and Webb (1991). For a concrete illustration of a wide variety of reliability estimates that can be formulated for assessment center and job performance ratings, see Putka and Hoffman (2013) and (2015), respectively. For a more thorough technical presentation, one should consult Brennan (2001b).

Confirmatory Factor Analytic Tradition

Although G-theory is often espoused as a conceptual centerpiece of modern psychometrics (along with item response theory, or IRT), it is important to separate the conceptual perspective G-theory offers on reliability from the estimation methods (random-effects models) it proscribes. Such a distinction is important because although the conceptual perspective offered by G-theory can serve as a parsimonious way to frame the problem of building a reliability coefficient appropriate for one's situation (regardless of whether one uses classical methods, random-effects methods, or CFA methods to derive estimates of such coefficients), the random-effects model that undergirds G-theory and classical methods of estimating reliability share a key drawback. Specifically, they offer no clear mechanism for (a) testing or dealing with violations of CTT and G-theory measurement model assumptions and (b) specifying or testing alternative factorial compositions of true score—both of which have fundamental implications for the interpretation of reliability estimates.[18] It is in this regard that CFA approaches to estimating reliability are strong (McDonald, 1999).

Unlike reliability estimation approaches born out of the classical and random-effects traditions, CFA-based approaches force investigators to be specific about the substantive nature of the latent structure underlying their replicates (indicators, in CFA terms). For example, CFA forces them to face questions such as:

- Is the covariance shared among replicates (i.e., true score variance from the perspective of classical and random-effects approaches) accounted for by a single latent true score factor or multiple latent factors?
- Do indicators of the latent true score factor(s) load equally on that/those factor(s) (i.e., are they at least essentially tau-equivalent) or is their heterogeneity in factor loadings (i.e., suggesting they are not at least essentially tau-equivalent)?
- What proportion of true score variance (as defined in CTT and G-theory) reflects the effects of a single latent factor, as opposed to residual covariances?

Although such questions have implications for reliability estimates arising from the classical and random-effect traditions, neither of these traditions has a built-in mechanism for addressing them. In essence, they ascribe all shared covariance among replicates to a latent entity (e.g., true score), regardless of whether it stems from a single factor or multiple factors. Thus, in some ways CFA can be seen as a way of clarifying the factorial composition of true score variance as conceived by CTT and G-theory measurement models. One may argue that such clarification is more an issue of validity rather than reliability (e.g., Schmidt et al., 2000); however, as discussed in the following paragraphs, the dimensionality of the focal construct of interest has implications for the accuracy of reliability estimates based on the classical and random-effects traditions (Lee & Frisbie, 1999; Rae, 2007; Rogers, Schmitt, & Mullins, 2002).

The CFA tradition of reliability estimation arose out of Joreskog's (1971) work on the notion of congeneric tests discussed earlier. To illustrate, consider a situation in which we administer a 10-item measure of agreeableness to a sample of job applicants. In this case, our replicates are single-faceted and defined in terms of items, and those items are fully crossed with our objects

of measurement—applicants. From the CFA perspective, we might view the replicates as indicators of a latent factor representing true score, and then fit a model to the data such that the variance of the latent factor is set to one, and the factor loadings and unique variances are freely estimated. On the basis of such a model, the estimated reliability of the sum of the *k* replicates can be obtained via

$$\omega = \frac{\left(\sum_{i=1}^{k}\lambda_i\right)^2}{\left(\sum_{i=1}^{k}\lambda_i\right)^2 + \sum_{i=1}^{k}\theta_{ii}} = \frac{\hat{\sigma}^2_T}{\hat{\sigma}^2_T + \hat{\sigma}^2_E}, \quad (1.7)$$

where λ_i represents the estimated factor loading for the *i*th of *k* replicates, and θ_{ii} represents the estimated unique variance for the *i*th replicate (McDonald, 1999; Reuterberg & Gustafsson, 1992).[19] As with the application of classical and random-effects approaches to reliability, the substantive meaning of σ^2_T and σ^2_E based on this formulation will differ depending on the type of replicates to which they are applied (e.g., items, raters, occasions, tasks, etc.). Thus, if applied to replicates defined by items, the estimate of σ^2_T provided by the squared sum of loadings will reflect consistency among items (attributable to the latent true score factor). If applied to replicates defined by raters, the estimate of σ^2_T provided by the squared sum of loadings will reflect consistency among raters (again, attributable to the latent true score factor).

A key benefit of the CFA approach described above is that it will allow one to impose constraints on parameter estimates (e.g., the λ_s and θ_s) that allow one to test various assumptions underlying the CTT and G-theory score models (see Joreskog & Sorbom, 2001, pp. 124–128; Reuterberg & Gustafsson, 1992; Werts, Linn, & Joreskog, 1974). If replicates are strictly parallel (an assumption underlying the Spearman-Brown prophecy; Feldt & Brennan, 1989), then they should have equal factor loadings (i.e., $\lambda_1 = \lambda_2 = \lambda_k$) and equal unique variances (i.e., $\theta_{11} = \theta_{22} = \theta_{kk}$). If replicates are tau-equivalent or essentially tau-equivalent (an assumption underlying alpha and coefficients based on the random-effects tradition), then they should have equal factor loadings but their unique variances can differ. To the extent that factor loadings vary across replicates (i.e., the replicates are not at least essentially tau-equivalent), most reliability estimates based out of the classical and random-effects tradition (e.g., alpha) will tend to be slightly downward biased (Novick & Lewis, 1967).[20] Nevertheless, this common claim is based on the premise that the replicates on which alpha is estimated are experimentally independent—from a CFA perspective this would imply there are no unmodeled sources of covariance among replicates after accounting for the latent true score factor (Komaroff, 1997; Raykov, 2001a; Zimmerman, Zumbo, & LaLonde, 1993). In light of the fact that many constructs of interest to organizational researchers are heterogeneous (e.g., situational judgment) or clearly multidimensional (e.g., job performance), application of the formula shown in Equation 1.7 would be questionable because it implies that a single common factor accounts for the covariance among replicates, which in practice may rarely be true.

The observation above brings us to a critical difference between the CFA-based formulation of reliability noted in Equation 1.7 and those based on the classical and random-effects traditions—the former often specifies a single latent factor as the sole source of covariance among replicates, and as such only variance in replicates attributable to that factor is treated as true score variance (for a more general, CFA-based alternative, see Raykov & Shrout, 2002). Recall from the operational definition of true score offered earlier and the perspective on true score offered by the CTT and G-theory score models that true score reflects all sources of consistency across replicates. As Ghiselli (1964) noted,

> The fact that a single term . . . has been used to describe the amount of the trait an individual possesses should not be taken to imply that individual differences in scores on a given test are determined by a single factor. (p. 220)

The implications of this are that whereas the CFA formulation above ignores any covariance among replicates that is left over after extracting a first latent factor, classical coefficients such

as alpha and coefficients derived from the random-effects tradition lump such covariance into the estimate of true score variance (Bost, 1995; Komaroff, 1997; Maxwell, 1968; Raykov, 2001a; Smith & Luecht, 1992; Zimmerman et al., 1993).

This characteristic of the CFA approach offered above presents investigators with a dilemma: Should residual covariance observed when adopting such an approach be treated as (a) error variance (σ^2_E) or (b) a source of true score variance (σ^2_T)? In estimates of reliability based on the classical tradition, one does not have much of an option. True score variance as estimated under the classical tradition reflects any source of consistency in scores, regardless of whether it stems from a first common factor, or what, in CFA terms, would be viewed as residual covariance or correlated uniquenesses (Komaroff, 1997; Scullen, 1999). Similarly, under the random-effects tradition, true score variance reflects any source of consistency in score, not accounted for by a variance component reflecting one of the facets of measurement (e.g., items, raters, occasions). However, with CFA, researchers have the flexibility to distinguish between true score variance that (a) arises from a common factor hypothesized to reflect a construct of interest and (b) reflects residual covariance among replicates after extracting the first factor (Raykov, 1998, 2001b). Although in theory having this flexibility is valuable because it allows one insight into the substance of true score variance, it also has practical benefits in that it can allow investigators to tailor a reliability coefficient to their situation depending on the nature of the construct they are assessing. To illustrate this flexibility, I offer three examples as follows that selection researchers and practitioners may encounter.

First, let us say one (a) designs a measurement procedure to assess a unidimensional construct, (b) uses a fully crossed measurement design comprising replicates defined by a single facet of measurement (e.g., items) to assess it, (c) fits the single-factor CFA model described above to the resulting data, and (d) finds evidence of residual covariance. Assuming there is no pattern to the residual covariance that would suggest the presence of additional substantively meaningful factors, the investigator would likely desire to treat the residual covariance as a source error variance (σ^2_E) rather than a source of true score variance (σ^2_T).[21] Fortunately, such residual covariance can be easily incorporated into Equation 1.7 by replacing the term corresponding to the sum of unique variances with a term that reflects the sum of unique variances and residual covariances or by simply replacing the denominator with observed score variance (Komaroff, 1997; Raykov, 2001a). If one were to calculate alpha on these same data, or fit a simple random-effects model to estimate σ^2_T, such residual covariance would be reflected in σ^2_T as opposed to σ^2_E and thus would produce a reliability estimate that is higher than the modified omega-coefficient described here when the sum of the residual covariances are positive (lower when the sum is negative) (Komaroff, 1997). It is important to note that the comparison made here between alpha and modified omega is based on the assumption that the replicates in the analysis are at least essentially tau-equivalent. If the replicates are not at least essentially tau-equivalent, then this would lower the estimate of alpha, thus either partially or completely offsetting any positive bias created by the presence of residual covariance (Raykov, 2001a).

As another example, let us say one (a) designs a measurement procedure to assess a relatively heterogeneous, but ill-specified construct (e.g., situational judgment), and again (b) uses a fully crossed measurement design comprising replicates defined by a single facet of measurement (e.g., scenarios) to assess it, (c) fits the single-factor CFA model described above to the resulting data, and (d) finds evidence of residual covariance. In this case, the investigator may choose to treat the residual covariance as a source of true score variance (σ^2_T) rather than error variance (σ^2_E). Unlike the first example, given the heterogeneous nature of the situational judgment construct, the investigator would not likely expect the covariance among scenarios to be accounted for by a single factor. For example, the investigator may hypothesize that the scenarios comprising the assessment vary in the degree to which various combinations of individual differences (e.g., interpersonal skill, conscientiousness, and general mental ability) are required to successfully resolve them. As such, scores on scenarios may differentially covary depending on the similarity of the individual difference profile required to resolve them. Under such conditions, one would need more than a single factor to account for covariation among replicates, but given the ill-structured nature of situational judgment construct, the investigator may not find strong evidence for a simple factor structure. As was the case with treating residual covariances as σ^2_E in the previous

example, Equation 1.7 can easily be modified to treat residual covariances as σ^2_T by adding a term to the squared sum of loadings that reflects the sum of all residual covariances, specifically

$$\varpi' = \frac{\left(\sum_{i=1}^{k}\lambda_i\right)^2 + \sum_{i}^{k}\sum_{\substack{j \\ i\neq j}}^{k}\mathrm{Cov}(e_i, e_j)}{\left(\sum_{i=1}^{k}\lambda_i\right)^2 + \sum_{i}^{k}\sum_{\substack{j \\ i\neq j}}^{k}\mathrm{Cov}(e_i, e_j) + \sum_{i=1}^{k}\theta_{ii}} = \frac{\hat{\sigma}^2_T}{\hat{\sigma}^2_T + \hat{\sigma}^2_E}. \quad (1.8)$$

Note that treating residual covariance as σ^2_T represents a departure from how the CFA literature on reliability estimation has generally espoused treating such covariance when estimating reliability (e.g., Komaroff, 1997; Raykov, 2001b). Nevertheless, the perspective offered by these authors is largely based on the assumption that the investigator is assessing a unidimensional construct. If one were to calculate alpha on such replicates, or fit a random-effects model to estimate σ^2_T, the covariance among residuals noted above would contribute to σ^2_T as opposed to σ^2_E and as such would produce a coefficient similar in magnitude to what is provided by Equation 1.8 (Komaroff, 1997).

Lastly, and as a third example, let us say one (a) designs a measurement procedure to assess a multidimensional construct (e.g., job performance), (b) uses a fully crossed measurement design comprising replicates defined by a single facet of measurement (e.g., items) to assess it, and (c) samples content for the measure in a way that allows one to distinguish between different dimensions of the construct (e.g., samples items corresponding to multiple job performance dimensions). In this situation, one might be interested in estimating the reliability of scores on each dimension of the construct separately, as well as estimating the reliability of a composite score based on the sum of dimension-level scores (e.g., an overall performance score). To achieve a reliability estimate for scores on the overall composite, the single-factor CFA model described would clearly not be appropriate. Rather, a multifactor model may be fitted in which each factor reflects dimensions of the construct being targeted by the measure. Indicators would be allowed to load only on those factors they are designed to reflect, and the reliability and true score variance of the overall composite score would be a function of factor loadings and factor covariances (Kamata, Turhan, & Darandari, 2003; Raykov, 1998; Raykov & Shrout, 2002). Any residual covariance among indicators associated with a given factor could be treated as noted in the earlier examples (i.e., treated as σ^2_T or σ^2_E) depending on how the investigator views such covariance in light of the substantive nature of the target construct and particular measurement situation. Such multifactor models could also be used to simultaneously generate separate estimates of reliability of scores for each dimension of the construct (Raykov & Shrout, 2002).

Although classical and random-effects traditions do not concern themselves with the factorial composition of true score covariance as the CFA tradition does, estimation methods arising out of the former traditions have developed to deal with reliability estimation for scores produced by measures that clearly reflect the composite of multiple dimensions. Such methods have typically been discussed under the guise of (a) reliability estimation for measures stratified on content (e.g., items comprising the measure were sampled to assess relatively distinct domains such as deductive and inductive reasoning) or, more generally, (b) reliability estimation for composite scores (Cronbach, Schoneman, & McKie, 1965; Feldt & Brennan, 1989). Here "composites" do not necessarily refer to a compilation of items thought to reflect the same construct (i.e., replicates), but rather compilations of measures designed to reflect distinct, yet related constructs (e.g., proficiency with regard to different requirements for a trade or profession) or different components of a multidimensional construct (e.g., task and contextual performance). Scores produced by such component measures may differ in their reliability and their observed relation with one another. Sensitivity to such issues is clearly seen in classical formulas for the reliability of composites such as stratified coefficient alpha (Cronbach et al., 1965) and Mosier's (1943) formula for the reliability of a weighted composite.[22] In the case of stratified alpha and Mosier's coefficient, σ^2_T for the overall composite score reflects the sum of σ^2_T for each component of the composite and the sum of covariances between replicates (e.g., items, raters) comprising different components.[23] The fact that covariances between replicates from different components of the

composite contribute to true score variance has a very important implication: these estimates will likely produce inflated estimates of reliability in cases in which measures of each component share one or more elements of a facet of measurement (e.g., raters, occasions) in common.

For example, consider a situation in which one gathers job performance ratings on two dimensions of performance for each ratee—task performance and contextual performance. Assume that, for any given ratee, the same two raters provided ratings of task performance and contextual performance. In this case, the measures of task and contextual performance share raters in common and as such are "linked" (Brennan, 2001b). Thus, the covariation between task and contextual performance in this example reflects not only covariance between their true scores but also covariance arising from the fact that they share a common set of raters. Were we to apply stratified alpha or Mosier's formula to estimate the reliability of the composite score produced by summing across the two dimensions (using inter-rater reliability estimates for each component in the aforementioned formulas), covariance attributable to having a common set of raters would contribute to true score variance, thus artificially inflating the reliability estimate (assuming we wish to generalize the measures across raters). Stratified alpha and Mosier's formula are based on the assumption that errors of measurement associated with components that comprise the composite are uncorrelated; to the extent they are positively correlated—a likely case when components share one or more elements of a facet of measurement in common—the estimates they provide can be substantially inflated (Rae, 2007). Outside of multivariate G-theory, which is not widely used or discussed in the organizational research literature (Brennan, 2001b; Webb & Shavelson, 1981), there appear to be no practical, straightforward analytic solutions to this situation on the basis of classical and random-effects estimation traditions.

Multifaceted Replicates and Noncrossed Measurement Designs in CFA

In all of the CFA examples offered above, the discussion assumed that the source(s) of extra covariation among replicates beyond the first factor was due to multidimensionality, or more generally heterogeneity in the construct being measured. However, as the example from the previous paragraph illustrated, such covariation can also arise from the characteristics of one's measurement design. For example, such extra covariation can also arise if the replicates that serve as indicators in a CFA are multifaceted and share a measurement design element (e.g., a rater, an occasion) in common. This brings us to another critical point regarding the CFA-based approach to reliability estimation discussed above. When Joreskog (1971) originally formulated the congeneric test model upon which many CFA-based estimates of reliability are grounded, it was based on a set of replicates defined along a single facet of measurement (e.g., items), and that facet was assumed to be fully crossed with the objects of measurement (e.g., persons). However, as noted above, when replicates are multifaceted, those replicates that share a level of a given facet in common (e.g., replicates that share a common rater or occasion of measurement) will covary above and beyond any substantive factors (e.g., interpersonal skill, job performance) that underlie the replicates (DeShon, 1998).

There are numerous ways to account for multifaceted replicates within the CFA framework; however, only recently have they begun to find their way into the literature (e.g., DeShon, 1998; Green, 2003; Le et al., 2009; Marcoulides, 1996; Marsh & Grayson, 1994). Many of the methods being espoused for handling multifaceted replicates in the context of CFA have their roots in the literature on modeling of multitrait-multimethod data (e.g., Kenny & Kashy, 1992; Widaman, 1985). For example, in the context of the interview example offered earlier, we might fit a model that not only includes a latent factor corresponding to the construct of interest (e.g., interpersonal skill) but also specifies latent factors that correspond to different raters or interview questions (e.g., all indicators associated with rater 1 would load on a "Rater 1" factor, all indicators associated with rater 2 would load on a "Rater 2" factor). Alternatively, one might allow uniqueness for those indicators that share a rater or question in common to covary and constrain those that do not go to zero (e.g., Lee, Dunbar, & Frisbie, 2001; Scullen, 1999). By fitting such models, one can derive estimates of variance components associated with various elements of one's measurement design (e.g., person-rater effects, person-question effects) that

resemble what is achieved by fitting a random-effects model to the data described earlier (e.g., DeShon, 2002; Le et al., 2009; Marcoulides, 1996; Scullen et al., 2000). As illustrated earlier in the discussion of G-theory, these variance components can then be used to construct reliability coefficients appropriate for one's situation.

Unfortunately, as was the case with using classical reliability coefficients to calibrate various sources of error in scores (e.g., Schmidt et al., 2000), CFA-based approaches to variance component estimation have a few drawbacks. First, they do not lend themselves easily to estimating variance attributable to (a) facet main effects (e.g., rater main effects, question main effects) or (b) interactions among measurement facets (e.g., rater-question interaction effects). Although it is possible to estimate the effects above, this would require calculating covariances among persons (i.e., persons as columns/variables) across facets of measurement of interest (e.g., raters, question) as opposed to the typical calculation of covariances among question-rater pairs (i.e., question-rater pairs are treated as columns/variables) across objects of measurement (e.g., persons).[24] Furthermore, it is not clear how CFA could be leveraged to deal with designs that are more ill-structured in nature (e.g., Putka et al., 2008). For example, recall the example earlier where we had performance ratings for a sample of incumbents that were rated by multiple raters, and the raters that rated each incumbent varied in their degree of overlap. When confronted with such designs in the past applications of CFA, organizational researchers have generally resorted to random assignment of raters to columns for each ratee (e.g., Mount, Judge, Scullen, Sytsma, & Hezlett, 1998; Scullen et al., 2000; Van Iddekinge, Raymark, Eidson, & Attenweiler, 2004). As noted earlier, the drawback of doing this is that it can produce results that (a) vary simply depending on how raters are assigned for each ratee and (b) fail to account for the nonindependence of residuals for incumbents that share a rater in common, which downwardly biases estimates of true score variance (Kenny & Judd, 1986; Putka et al., 2008).

Lastly, for better or worse, the literature on CFA offers myriad ways to parameterize a model to arrive at variance component estimates, each of which has various strengths and weaknesses that are still in the process of being ironed out (e.g., Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Marsh & Grayson, 1994). With the random-effects model discussed above, the number of alternative parameterizations (at least as currently implemented in common statistical software such as SAS and SPSS) is quite limited. The difficulty this creates when using CFA to estimate variance components is determining which parameterization is most appropriate in a given situation, because a clear answer has not emerged and will likely ultimately depend on characteristics of the construct being measured and characteristics of one's measurement situation (Marsh & Grayson, 1994). This is complicated by the fact that the choice of which parameterization is adopted often may be less a matter of substantive considerations and more a reflection of the parameterization that allowed the software fitting the model to converge to an admissible solution (Lance et al., 2004). Ultimately, such nuances, coupled with the complexity of CFA-based approaches to variance component estimation, may limit the utility of such approaches for reliability estimation in general selection research and practice.

Summary: Comparison and Equifinality of Estimation Traditions

On the basis of the examples, one might ask which tradition best serves the needs of personnel selection research and practice? My answer would be no single tradition currently satisfies all needs. Table 1.1 summarizes characteristics of the reliability estimation traditions discussed above.

Beyond their simplicity and familiarity, classical approaches do not appear to have much to offer. Modern random-effects approaches address not only measurement situations that classical approaches were initially designed to handle (e.g., those involving single-faceted replicates and fully crossed designs) but also those situations that classical approaches were not designed to handle (i.e., procedures involving multifaceted replicates and/or noncrossed designs). Couple this with the ease with which variance components can now be estimated using widely available software (e.g., SPSS, SAS), as well as the consistency of the random-effects model with modern psychometric

TABLE 1.1

Relative Advantages and Disadvantages of Reliability Estimation Traditions

		Random	
Characteristics of Estimation Tradition	*Classical*	*Effects*	*CFA*
Perceived simplicity	Yes	No	No
Widely discussed in organizational literature on reliability estimation	Yes	No	No
Easily implemented with standard statistical software (e.g., SPSS, SAS)	Yes	Yes	No[a]
Direct and simultaneous estimation of variance components underlying α^2_T and α^2_E	No	Yes	Yes
Straightforward to apply to nested and ill-structured measurement designs confronted in applied organizational research and practice	No	Yes	No
Capacity to isolate and estimate variance attributable to facet main effects and interactions among facets	No	Yes	No[b]
Offers mechanism for testing and dealing with violations of CTT and G-theory measurement model assumptions	No	No[c]	Yes
Offers mechanism for specifying and testing alternative factorial compositions of true score	No	No	Yes

[a] Potential exception is PROC CALIS within SAS.
[b] As noted in text, such effects could be estimated by fitting CFA models to covariances calculated across measurement facets (e.g., question, raters, question-rater pairs) as opposed to objects of measurement (e.g., persons).
[c] SAS and SPSS now offer users the ability to fit "heterogeneous variance" random-effect models, which for some designs can be used to assess various equivalence assumptions underlying the CTT and G-theory measurement models (e.g., Is α^2_T for Rater 1 = α^2_T for Rater 2?).

perspectives on reliability (i.e., G-theory; AERA, APA, & NCME, 2014; Brennan, 2006), and it appears the random-effects tradition has much to offer. Nevertheless, the classical and random-effects traditions suffer from two similar drawbacks in that their estimation procedures offer no clear mechanism for (a) testing or dealing with violations of CTT and G-theory measurement model assumptions on which their formulations of reliability are based and (b) specifying or testing alternative factorial compositions of true score. The latter drawback can make the interpretation of reliability estimates difficult because of ambiguity of what constitutes true score, particularly for measures of heterogeneous constructs. This is where the CFA tradition can offer an advantage; however, this advantage does not come freely—its price is added complexity.

For single-faceted replicates that are fully crossed with one's objects of measurement, CFA methods are straightforward to apply and clear examples exist (e.g., Brown, 2006; McDonald, 1999). For multifaceted replicates, a systematic set of examples has yet to be provided for investigators to capitalize on, which is complicated by the fact that the CFA models can be parameterized in numerous different ways to arrive at a solution (Marsh & Grayson, 1994). This has a tendency to restrict such solutions to psychometrically savvy researchers and practitioners. Moreover, for the ill-structured measurement designs discussed by Putka et al. (2008), which are all too common in selection research involving ratings (e.g., assessment centers, interviews, job performance), it is not clear how the CFA models would overcome the issues raised. Thus, we have a tradeoff between the ease with which modern random-effects models and software can deal with multifaceted measurement designs of any sort and the model fitting and testing capabilities associated with CFA, which can not only check on measurement model assumptions but also refine our specification (and understanding) of true score for measures of heterogeneous constructs.

Although I have treated reliability estimation approaches arising out of classical, random effects, and CFA traditions separately, it is important to recall how we began this section: all of these traditions can be used to arrive at the same ratio—$\sigma^2_T/(\sigma^2_T + \sigma^2_E)$. The substantive meaning of σ^2_T and σ^2_E will depend on the type of replicates examined, the nature of the measurement procedure, and the construct that one is assessing. Nevertheless, all of these traditions can potentially be leveraged to arrive at an estimate of this ratio and/or components of it. How

they arrive at those estimates, the assumptions they make in doing so, and the strengths and weaknesses of the methodologies they use is what differentiates them. In cases in which one has a measurement procedure comprising single-faceted replicates or multifaceted replicates in which facets are fully crossed with one's objects of measurement and one is interested solely in using scores to make relative comparisons among objects of measurement (e.g., persons), much literature has accumulated indicating that these traditions can produce very similar results, even in the face of moderate violation of common tau-equivalence assumptions (e.g., Le et al., 2009; Reuterberg & Gustafsson, 1992).

For example, Brennan (2001b) and Haertel (2006) show how random-effects ANOVA models may be used to estimate variance components and form reliability coefficients that are identical to the types of reliability coefficients from the classical tradition (e.g., alpha, coefficients of equivalence and stability). Marcoulides (1996) demonstrated the equivalence of variance components estimated based on CFA and random-effects ANOVA models fitted to a multifaceted set of job analysis data. Le et al. (2009) illustrated how one can arrive at similar variance component estimates using functions of Pearson correlations, random-effects models, and CFA models. Lastly, Brennan (2001b) and Hocking (1995) demonstrated how it is possible to generate variance component estimates without even invoking the random-effects or CFA models but simply calculating them as functions of observed variance and covariances (in some ways akin to Schmidt et al., 2000). Each of these works illustrate that, under certain conditions, the three traditions discussed can bring investigators to similar conclusions. However, as illustrated, nuances regarding the (a) generalizations one wishes to make regarding their scores, (b) the intended use of those scores (e.g., relative comparisons among applicants vs. comparisons of their scores to a fixed cutoff), (c) characteristics of one's measurement procedure itself (e.g., nature of its underlying design), and (d) characteristics of the construct one is attempting to measure (e.g., unidimensional vs. multidimensional, homogeneous vs. heterogeneous) make some of these approaches more attractive than others under different circumstances. Ideally, the well-informed investigator would be in a position to capitalize on the relative strengths of these traditions when formulating reliability and variance component estimates of interest given his/her situation.

Lastly, regardless of what theoretical perspective one adopts on reliability estimation (i.e., whether it is more grounded in CTT or G-theory), I encourage future researchers to think critically and attempt to evaluate whether the data they are attempting to apply those theories to conform to the score models underlying their estimates of reliability. My sense is that within I-O psychology and the organization sciences, we often take the notion that our data conform to the assumptions implied by score models underlying reliability estimates as a given, but rarely do we take the time to seriously evaluate such claims.

CLOSING THOUGHTS ON RELIABILITY

Perspectives on reliability and methods for its estimation have evolved greatly over the last 50 years, but these perspectives and methods have yet to be well integrated (Brennan, 2006). One potential reason for this lack of integration may stem from the historical disconnect between experimental and correlation research traditions (Cronbach, 1957), which continues to manifest itself today, particularly in our approaches to reliability estimation (Cronbach & Shavelson, 2004). Another potential reason for this lack of integration may stem from the recognized decline in the graduate instruction of statistics and measurement over the past 30 years in psychology departments (Aiken et al., 2008; Merenda, 2007). For example, in reviewing results of their study of doctoral training in statistics, measurement, and methodology in PhD psychology programs across North America, Aiken et al. (2008) lament:

> We find it deplorable . . . the measurement requirement occupies a median of only 4.5 weeks in the PhD curriculum in psychology. A substantial fraction of programs offered no training in test theory or test construction; only 46% of programs judge that the bulk of their graduates could assess the reliability of their own measures.
>
> (p. 43)

Under such conditions, it makes it nearly impossible for faculty to comprehensibly integrate and discuss implications of developments in the areas above into classroom discussions of psychometrics; almost out of necessity, we limit ourselves to basic treatment of age-old perspectives on measurement. Couple this observation with the explosion of new statistical software and availability of new estimation methods since the mid-1980s, and it creates a situation where staying psychometrically current can be a challenge for those in academe, as well as those in practice. Of course, also complicating the trends is the course of normal science, which leads us to pursue incremental research that refines measurement models and the perspectives on reliability they offer but does not emphasize integration of models and perspectives (Kuhn, 1962). Such a lack of integration among psychometric models and perspectives is unfortunate because it can serve as a source of parsimony, which is critical when one has limited time to devote to such topics in the course of graduate instruction and in the course of applied research and practice. I hope this treatment has brought some degree of parsimony to what have often been treated as disparate, loosely related topics. Furthermore, I hope it casts developments in the area of reliability in a novel light for selection researchers and practitioners and encourages us to explore and capitalize on modern methods for framing reliability, error, and their underlying components.

NOTES

1. Throughout this chapter I use the term "scores" to generically refer to observed manifestations of a measurement procedure—thus, scores might be ratings, behavioral observations, test scores, etc.
2. As we discuss later, the degree to which replicates are assumed to "assess the same construct" differs across measurement theories. The degree of similarity among replicates has been discussed under the rubric of degrees of part-test similarity (Feldt & Brennan, 1989) and degrees of parallelism (Lord, 1955). At this point, further discussion of this issue is unnecessary, but we will revisit this issue when discussing the role of measurement models in reliability estimation.
3. A key exception here is Cronbach's (1947) treatment of a coefficient of equivalence and stability.
4. Actually, this is a bit of an overstatement. As alluded to in the opening paragraph, error, in any given situation, will be partly dependent on the generalization(s) the investigator wishes to make regarding scores. In some cases, investigators may choose not to treat a given source of inconsistency in scores as error. For example, this might occur in the context of performance ratings where inconsistencies in job incumbents' scores across different dimensions of performance may be viewed as acceptable by the investigator (Scullen, Mount, & Goff, 2000). This example illustrates why the investigator is a critical part of defining error in any given situation. We will touch upon this topic again later when we discuss generalizability theory.
5. Readers may question the omission of item response theory (IRT) from the subsequent discussion. Like Brennan (2006), we tend to view IRT models as "scaling" models rather than "measurement" models because they do not have a built-in explicit consideration of measurement error. Furthermore, the focus of applications of IRT is often on estimation/scaling of a latent trait, ability of interest, or calibration of item parameters rather than the isolation and quantification of measurement error (see Brennan, 2006, pp. 6–7). Although I am not downplaying the obvious importance of IRT for psychometrics and personnel selection, I felt it was beyond the scope of this chapter to address IRT while still addressing reliability as I have done herein. For a recent, parsimonious treatment of IRT, see Yen and Fitzpatrick (2006).
6. Given condition (a) and (c) such replicates will also have identical observed score variances.
7. Actually, this statement is a bit of a misnomer, because coefficient alpha and intraclass correlations simply represent specific computational forms of a broader class of coefficients known as generalizability coefficients (Cronbach & Shavelson, 2004).
8. The highest order interaction term and the residual term in G-theory models are confounded because such designs essentially amount to having one observation per cell. Thus, in practice, it is not possible to generate separate estimates of variance in X attributable to these two effects.
9. Two notes here: First, as we will discuss in the following sections, one's ability to estimate each of these components will be limited by the measurement design underlying one's measurement procedure. The example here assumes a fully crossed design, which will often not be the case in practice. Second, note that in Equation 1.5 we combine variance components for the applicant-question-rater interaction and residual terms; this reflects the fact that these sources of variance will not be uniquely estimable.

10. Although labeled as a "generalizability" coefficient, note that this formula provides an estimate of σ^2_T over $\sigma^2_T + \sigma^2_E$, and as such may be considered an estimate of reliability.
11. Note the idea of fixing a facet of measurement for purposes of estimating σ^2_T and σ^2_E in the context of G-theory is different from modeling a factor or covariate as fixed in the context of mixed-effects models (DeShon, 2002; Searle, Casella, & McCulloch, 1992).
12. Another common ratings design faced in practice (particularly with job performance ratings) is one in which ratees are nested with raters (e.g., each group of incumbents is rated by their respective group supervisor). In this case, each ratee has only one rater, and as such there is no way to distinguish between the σ^2_{PR} (typically considered a source of error) and σ^2_P (typically considered true score variance). Thus, estimating inter-rater reliability on the basis of data structured in this manner is not possible.
13. We refer the interested readers to Brennan (1998), Haertel (2006), and Qualls-Payne (1992) for modern treatments of SEMs in the context of CTT and G-theory. One advantage of SEMs over reliability coefficients is that they can be tailored to individuals being measured (e.g., differential amounts of error depending on individuals' level of true score), whereas reliability coefficients are typically associated with groups of individuals. The latter is often cited as one benefit of IRT-based perspectives on measurement over CTT- and G-theory-based perspectives; however, CTT and G-theory also offer methods for generating individual-level SEMs (Haertel, 2006).
14. My speculation on why this occurred is (a) the perceived complexity and jargon-loaded nature of G-theory (DeShon, 2002), (b) the overarching dominance of the correlational research tradition underlying selection research and practice (Cronbach, 1957; Dunnette, 1966; Guion, 1998), and (c) the steady decline of teaching psychometrics and statistics in graduate programs since the 1970s (Aiken et al., 2008; Merenda, 2007).
15. Given the discussion raised earlier, σ^2_T in any of these examples may also reflect variance attributable to one or more hidden facets of measurement.
16. On a historical note, Cronbach did not *invent* coefficient alpha per se—Guttman's (1945) L_3 coefficient and Hoyt's (1941) coefficient are algebraically identical to alpha and were introduced long before Cronbach's (1951) landmark article.
17. We should note that a subtle difference between Pearson *r*-based indices of reliability and those noted here (i.e., KR-20, Gutman's lambdas, alpha) is that the latter assess the additive relationship between replicates, whereas Pearson *r* assesses the linear relationship between replicates. Differences in the variances of replicates will reduce alpha and other additive reliability indices, but they will have no effect on Pearson *r*-based indices because the latter standardizes any variance differences between replicates away (McGraw & Wong, 1996).
18. One potential caveat to this regards the fairly recent ability of SAS and SPSS to fit random-effects models that allow for heterogeneity in variance component estimates (e.g., Littell et al., 1996; SPSS Inc., 2005). Such capabilities might be leveraged to test parallelism assumptions underlying the CTT and G-theory score models.
19. McDonald (1999) refers to this coefficient as "omega" (p. 89).
20. This downward bias arises from the fact that most reliability estimates based on these traditions rely on the average covariance among replicates to make inferences regarding the magnitude of true score variance for each replicate. To the extent that replicates are not essentially tau-equivalent, this average covariance will tend to underestimate true score variance for each replicate (a component of true score variance of the composite of replicates), thus leading to a slight underestimation of reliability when all other assumptions are met (e.g., uncorrelated errors among replicates) (Feldt & Brennan, 1989; Raykov, 2001a).
21. Even if there is a pattern to the residual covariances, the investigator might still wish to treat them as contributing to σ^2_E if they reflect an artifact of the particular measurement situation (e.g., Green & Hershberger, 2000). This raises an important point: The examples offered here are for illustration; they are not prescriptions for future research and practice. Ultimately, the individual investigator decides how to treat residual covariance given the characteristics of the measurement situation he or she faces.
22. Note Mosier's (1943) formula is equivalent to the formula for stratified coefficient alpha if elements comprising a composite are equally weighted.
23. Note that all else being equal, stratified alpha will tend to be higher (appropriately so) than coefficient alpha applied to the same data if between-component item covariances are lower than within-component item covariances—likely a common occurrence in practice for measures of multidimensional constructs (Haertel, 2006; Schmitt, 1996). In both cases the denominator of these coefficients is the same (observed variance); what changes is how true score variance for each *component of the composite* is estimated (these in turn are part of what contribute to σ^2_T *for the overall composite*). For

stratified alpha, $\sigma^2_{T_{}}$ *for any given component* is a function of the average covariance among items within that component, for alpha, $\sigma^2_{T_{}}$ *for any given component* is a function of the average covariance among *all* items, regardless of component. As such, if between-component item covariances are lower than within-component item covariances, $\sigma^2_{T_{}}$ *for any given component* will be lower if alpha is applied to the data rather than stratified alpha; in turn, the estimate $\sigma^2_{T_{}}$ *for the overall composite* produced by alpha will also be lower.

24. Interested readers are referred to Hocking (1995) and Brennan (2001b, pp. 166–168).

REFERENCES

Aguinis, H., Henle, C. A., & Ostroff, C. (2001). Measurement in work and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology: Vol. 1: Personnel psychology* (pp. 27–50). London, England: Sage.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Graduate training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*, 32–50.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bell, J. F. (1985). Generalizability theory: The software problem. *Journal of Educational and Behavioral Statistics, 10*, 19–29.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478–494.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425–440.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30*, 505–514.

Bost, J. E. (1995). The effects of correlated errors on generalizability and dependability coefficients. *Applied Psychological Measurement, 19*, 191–203.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22*, 307–331.

Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295–317.

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

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: American Council on Education and Praeger.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Campbell, J. P. (1976). Psychometric theory. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 185–222). New York, NY: John Wiley & Sons.

Cattell, R. B. (1966). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 67–128). Chicago, IL: Rand McNally.

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika, 12*, 1–16.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 292–334.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671–684.

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, NY: John Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373–399.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

Cronbach, L. J., Schoneman, P., & McKie, D. (1965). Alpha coefficient for stratified-parallel tests. *Educational & Psychological Measurement, 25*, 291–312.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and its successor procedures. *Educational and Psychological Measurement, 64*, 391–418.

DeShon, R. P. (1995, May). *Restricted maximum likelihood estimation of variance components in generalizability theory: Overcoming balanced design requirements.* Paper presented at the 10th annual conference of the Society of Industrial and Organizational Psychology, Orlando, FL.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*, 412–423.

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.

Dunnette, M. D. (1966). *Personnel selection and placement*. Oxford, England: Wadsworth.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407–424.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8*, 38–60.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York, NY: American Council on Education and Macmillan.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd.

Ghiselli, E. E. (1964). *Theory of psychological measurement*. New York, NY: McGraw-Hill.

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8*, 88–101.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7*, 251–270.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360 degree feedback ratings. *Journal of Applied Psychology, 83*, 960–968.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Guttman, L. A. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255–282.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Hocking, R. R. (1995). Variance component estimation in mixed linear models. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 541–571). New York, NY: Marcel Dekker.

Hoyt, C. (1941). Test reliability obtained by analysis of variance. *Psychometrika, 6*, 153–160.

Jackson, S., & Brashers, D. E. (1994). *Random factors in ANOVA*. Thousand Oaks, CA: Sage.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133.

Joreskog, K. G., & Sorbom, D. (2001). *LISREL 8: User's reference guide*. Lincolnwood, IL: Scientific Software International.

Kamata, A., Turhan, A., & Darandari, E. (April 2003). *Estimating reliability for multidimensional composite scale scores*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99*, 422–431.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165–172.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated errors on coefficient alpha. *Applied Psychological Measurement, 21*, 337–348.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: The University of Chicago Press.

Lance, C. L. (2008). Why assessment centers don't work the way they're supposed to. *Industrial and Organizational Psychology, 1*, 84–97.

Lance, C. L., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89*, 377–385.

Le, H., & Putka, D. J. (2007). Reliability. In S. G. Rogelberg (Ed.), *Encyclopedia of industrial and organizational psychology* (Vol. 2, pp. 675–678). Thousand Oaks, CA: Sage.

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.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests comprised of testlets. *Educational and Psychological Measurement, 61*, 958–975.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*, 237–255.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15*, 325–336.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27*, 251–280.

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports, 66*, 102–109.

Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling, 3*, 290–299.

Marsh, H. W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling, 1*, 116–146.

Maxwell, A. E. (1968). The effect of correlated error on reliability coefficients. *Educational and Psychological Measurement, 28*, 803–811.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

McPhail, S. M. (Ed.) (2007). *Alternative validation strategies: Developing new and leveraging existing validity evidence*. San Francisco, CA: John Wiley and Sons.

Merenda, P. F. (2007). Psychometrics and psychometricians in the 20th and 21st centuries: How it was in the 20th century and how it is now. *Perceptual and Motor Skills, 104*, 3–20.

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika, 8*, 161–168.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait rater and level effects in 360-degree performance ratings. *Personnel Psychology, 51*, 557–576.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873–900.

Ng, K. T. (1974). Spearman's test score model: A restatement. *Educational and Psychological Measurement, 34*, 487–498.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1–13.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and theory* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

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.

Putka, D. J., & Hoffman, B. J. (2015). 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, NY: Taylor & Francis.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*, 959–981.

Putka, D. J., & McCloy, R. A. (February 2008). *Estimating variance components in SPSS and SAS: An annotated reference guide*. Retrieved March 23, 2009, from http://www.humrro.org/djp_archive/Estimating Variance Components in SPSS and SAS.pdf

Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement, 29*, 213–225.

Rae, G. (2007). A note on using stratified alpha to estimate the composite reliability of a test composed of interrelated nonhomogeneous items. *Psychological Methods*, *12*, 177–184.

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22*, 375–385.

Raykov, T. (2001a). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*, 69–76.

Raykov, T. (2001b). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54*, 315–323.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9*, 195–212.

Reuterberg, S. E., & Gustafsson, J. E. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement, 52*, 795–811.

Rogers, W. M., Schmitt, N., & Mullins, M. E. (2002). Correction for unreliability of multifactor measures: Comparison of alpha and parallel forms of approaches. *Organizational Research Methods, 5*, 184–199.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods, 8*, 206–224.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*, 901–912.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.

Schmitt, N., & Landy, F. J. (1993). The concept of validity. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations* (pp. 275–309). San Francisco, CA: Jossey-Bass.

Scullen, S. E. (1999). Using confirmatory factor analyses of correlated uniquenesses to estimate method variance in multitrait-multimethod matrices. *Organizational Research Methods, 2*, 275–292.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956–970.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York, NY: Wiley.

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Smith, P. L., & Luecht, R. M. (1992). Correlated effects in generalizability studies. *Applied Psychological Measurement, 36*, 229–235.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

SPSS. (2005). *Linear mixed-effect modeling in SPSS: An introduction to the mixed procedure (Technical Report LMEMWP-0305)*. Chicago, IL: Author.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54*, 229–249.

Van Iddekinge, C. H., Raymark, P. H., Eidson, C. E., Jr., & Attenweiler, W. J. (2004). What do structured selection interviews really measure? The construct validity of behavior description interviews. *Human Performance*, *17*, 71–93.

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

Webb, N. M., & Shavelson, R. J. (1981). Multivariate generalizability of general educational development ratings. *Journal of Educational Measurement, 18*, 13–22.

Werts, C. E., Linn, R. L., & Joreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement, 34*, 25–32.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1–26.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger.

Zimmerman, D. W., Zumbo, B. D., & LaLonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53*, 33–49.