The SAGE Handbook of Quantitative Methodology for the Social Sciences

Responsible Modeling of Measurement Data for Appropriate Inferences: Important Advances in Reliability and Validity Theory

Contributors: David Kaplan
Print Pub. Date: 2004
Print ISBN: 9780761923596
Online ISBN: 9781412973502
DOI: 10.4135/9781412973502
Print pages: 74-93

Responsible Modeling of Measurement Data for Appropriate Inferences: *Important Advances in Reliability and Validity Theory*

Bruno D. Zumbo, André A. Rupp

4.1. Introduction

If the statistical, conceptual, and practical activities of measurement were a crop seeded by Spearman, Yule, Pearson, and others working the early fields of social and behavioral research, we could proudly say that those seedlings have resulted in a bountiful harvest. The annual yield of measurement research continues to grow, and the number of new journals and books devoted to and surveying the field and reporting advances has increased over the past decade. The goal of indexing data quality has a longstanding tradition in statistical modeling, and its ubiquity in psychometric modeling thus comes as no surprise, which is why research in reliability and validity theory continues to be of relevance today, as a quick glance at the reference list of this chapter reveals. Before we begin to describe the process of harvesting the statistical crops that have been sown, however, let us first take a look at the analyst's task in measurement itself.

Analysts of test data are typically faced solely with an array of numbers, which often consists of 0s and 1s when all items on a test are scored dichotomously. It is the objective of the analyst to use this array for a variety of meaningful inferences about the examinees and the measurement instrument itself, which should be appreciated as a daunting task. Statistical modeling has always been concerned with decomposing observational values into a component that is *deterministic* and a component that is *stochastic* so that relationships between manifest and unobserved variables can be explicitly stated and uncertainty about model parameters can be estimated and used to qualify the inferences that are possible under a given model. Psychometric models are, of course, descendants of this tradition (see Goldstein & Wood, 1989; McDonald, 1982; Mellenbergh, 1994; Rupp, 2002) but are unique because they are located at

the *intersection* of examinee and item spaces, *both* of which are typically of interest to measurement specialists. For example, classical test theory (CTT) (e.g., Lord & Novick, 1968) generically decomposes the observed score into a deterministic part (i.e., true score) and a stochastic part (i.e., error), generalizability theory (*g*-theory) (e.g., Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) further unpacks the stochastic part and redefines part of error as systematic components, and item response theory (IRT) (e.g., Lord, 1980; van der Linden & Hambleton, 1997) reformulates the two model components by inducing latent variables into the data structure. Structural equation models (SEM) (e.g., Muthén, 2002) and exploratory as well as confirmatory factor analysis models (EFA and CFA, respectively) (e.g., McDonald, 1999) decompose the covariance matrix of multivariate data into deterministic (i.e., reproduced covariance matrix) and stochastic (i.e., residual matrix) components, which is a model that can be equivalently written as a formulization involving latent variables.

Even though latent trait indicator values and observed composite scores are typically highly correlated, the injection of a latent continuum into the data matrix has given us the property of item and examinee parameter invariance for perfect model fit across populations and conditions, has allowed us to define conditional standard errors of measurement similar to *g*-theory (Brennan, 1998b), and has opened up the road for adaptive testing through the use of item and test information functions (e.g., Segall, 1996; van der Linden & Hambleton, 1997). Still, these advances have not come without a price. Improvements in the level of modeling and in quantifying measurement error have come at the expense of large sample sizes that are typically required for parameter estimation in both frequentist and Bayesian frameworks (see Rupp, Dey, & Zumbo, in press). For example, categorical data, particularly dichotomous data, require the use of estimation methods such as weighted least squares, which make, for example, reliability estimates based on small sample sizes suspect (Raykov, 1997a).

The focus in this chapter is on reliability and validity, two topics that have generated many papers and books, even if one were to focus on the past 25 years only. As it is nearly impossible to review all of the developments in a single book chapter, we aim to provide a broad overview of recent developments in reliability and validity theory and periodically provide more detail to demonstrate the vast array of measurement methodologies and approaches currently available to aid us in illuminating our

understanding of social and behavioral phenomena. We will view these developments through a statistical modeling lens to highlight the consequences of choosing—perhaps even abusing—a particular modeling framework for inferential decisions.

We assume a basic exposure to measurement and test theory, but we will define basic key terms. For an accessible overview and advances in the statistical basis of reliability theory, the interested reader can consult Feldt and Brennan (1989), Knapp (2001), and Traub (1994), and for validity theory and practice, the reader can consult Messick (1995) and the papers in Zumbo (1998). Because our chapter presumes a working knowledge of modeling frameworks used in practical measurement problems, the reader might refer to Hambleton, Swaminathan, and Rogers (1991) or Lord (1980) as useful references for IRT, Kaplan (2000) or Byrne (1998) as useful references for structural equation modeling, and Comrey (1973), Everitt (1984), or McDonald (1999) as useful references for factor analysis (FA) methods.

Our discussion begins with an overview of frequently used key terms in the measurement literature to aid the understanding of our subsequent discussions, clarify some common misconceptions, and allow for more precise statements. We then present some important and practically relevant findings from the literature on reliability theory in roughly the past decade, with a strong focus on developments for reliability coefficients, standard errors of measurement, and other local quantifiers of measurement error. Finally, a section on validity theory illustrates how models for cognitively diagnostic assessment have forced measurement specialists to rethink their approaches to defining and measuring what constitutes valid inferences from test scores. But first, let us lay some groundwork with a brief discussion of terminology relevant for modeling data from measures.

4.2. Commonly Used and Misunderstood Terms in Measurement

Although the definitions presented in this section are fundamental, it is remarkable how often they are used inconsistently in the measurement literature. This is probably partly an artifact of inconsistent historical usage but can also be traced back to a discrepancy

that typically exists between the everyday usage of these terms and their precise meaning in a mathematical modeling context.

First, there is the word *reliability* itself. In nonacademic contexts, *reliable* is commonly understood to mean “a consistent dependability of judgment, character, performance, or result” (see Braham, 1996, p. 1628). For applied measurement specialists, reliability is a desired property of tests, which should be dependable measurement instruments of the constructs that they are supposed to measure or dependable measurement instruments for performance evaluation (Klieme & Baumert, 2001). Even though these notions are intuitively appealing, they are relatively imprecise and need to be translated into properties that can be mathematically tested and estimated through sample quantities. Consequently, reliability in a non-mathematical sense is often understood to be so much more than reliability in a strictly mathematical sense because, under the latter lens, reliability is basically translated into the estimation of a coefficient based on variance components in a statistical model. Such a *reliability coefficient* assesses consistent *scores* but, per se, says little about the assessment instrument itself, related inferences, and social consequences because those aspects are embedded in the larger value-laden ethical and social context of test use (Messick, 1995).

As Zimmerman and Zumbo (2001) note, formally, test data are the realization of a stochastic event defined on a product space # = #*I* x #*J*, where the orthogonal components, #*I* and #*J*, are the probability spaces for items and examinees, respectively. The joint product space can be expanded to include other spaces as well, such as spaces induced by raters or occasions, a concept that was formalized in *g*-theory from an observed-score perspective and the facets approach to measurement from an IRT perspective. Hence, modeling of test data minimally requires sampling assumptions about items and examinees, as well as the specification of a stochastic process that is supposed to have generated the data (for readers interested in a measure-theoretic Hilbert-space approach to the analysis of test data, see Zimmerman & Zumbo, 2001). Therefore, two distinct directions of generalizability are typically of interest, which require an understanding of the reliability and validity properties of scores and inferences. First, it is of interest to make statements about the functioning of a particular assessment instrument for groups of examinees who share characteristics with those examinees who have already been scored with it. Second, it is of interest to make statements about the functioning of item sets that share characteristics with

those items that are already included on a particular test form. For example, it is often of interest to show that the scores and resulting inferences for different examinee groups are comparably reliable and valid if the same instrument is administered to the different groups, a parallel version of the instrument is administered to the different groups, or selected subsets of items are administered to the different groups. This also specifically implies that researchers should report estimates of reliability coefficients and other parameters for their own data, rather than relying on published reports from other data, and that comparable validity needs to be continually assessed rather than being taken for granted based on a single assessment calibration. Let us take a look at some commonly used terms to describe the process of modeling assessment data.

It is useful to first distinguish between *test-level models* (e.g., CTT, *g*-theory models), in which modeling takes place at the observed total-score level, and *item-level models* (e.g., IRT for binary or rating scale item data and factor analysis models for continuous item data), in which modeling takes place at the item-score level along with the total-score level. For the latter models, the *primary* modeling unit is the *item*, which can be a written, aural, or graphical stimulus that entices examinees to produce behavioral responses. Yet the seemingly unambiguous notion of an item is rather fluid and context dependent. For example, items can be collected, either naturally through their placement alongside reference information on an assessment or statistically through definition, into item bundles or *testlets*, which can then be treated as a single item in subsequent mathematical analyses (note that potential response dependencies can be modeled explicitly as well; see Bradlow, Wainer, & Wang, 1999; Wang, Bradlow, & Wainer, 2002). Moreover, in other testing contexts with complex work products the definition of a single item can become extremely challenging if not impossible, and it might be preferable and necessary in the future to think of *measurement opportunities* more generally instead. For a recent description of the variety of items currently being used in measurement practice, see Zenisky and Sireci (2002).

Items can be assembled for different purposes such as personality trait assessment or knowledge assessment, and it is the latter scenario that typically leads to instruments that are commonly called *tests*. In addition, the term *scale* is also often used in the social science literature on personality assessment interchangeably with the term *questionnaire*. The terms *test, scale*, and *measure* are used interchangeably in this

chapter, but it is acknowledged that *tests* are, in common language, used to imply some educational achievement or knowledge test with correct or incorrect responses.

A subject's response to an item then becomes a *behavioral observation* in an abstract conceptual sense that needs to be quantified with a *score*, which, in turn, becomes a *statistical observation*. Some typical forms of scores are the (weighted) linear composite or *total score* that arises from individual items being scored *dichotomously* or *polytomously*. Measurement specialists then resort to specific *modeling frameworks* to account for the fact that behavioral observations are imperfect representations of the *latent variable* whose relative absence or presence the assessment instrument is supposed to quantify and, as such, contain *measurement error*. Indeed, the choice of measurement model has fundamental implications for how measurement error is viewed, and these differences lead modelers to choose particular model-specific statistics to quantify this error. Error, then, albeit a universally present phenomenon of observed behavioral responses, is conceived and quantified differently in alternate micro-universes created by different modeling frameworks. Interestingly, the well-known psychometric statement $X = T + E$ is axiomatic for all models in such frameworks.

In any modeling framework, the observable or *manifest* scores created by the interaction of examinees with items on a measure are considered to be *indicators* or markers of unobservable or *latent* variables. In this chapter, we will use the term *latent variable* to refer to a random variable that is deliberately constructed or derived from the responses to a set of items and that constitutes the building block of a statistical model (e.g., # scores in IRT or factor scores in FA). In other words, the scores are indicators of the latent variable, which is itself supposed to be an indicator of an underlying *latent trait* that is inherent in the examinees and supposedly tapped into by the items. However, these quantities and objects are not identical: The latent variable is a *psychometric* construction, whereas the latent trait is a *psychological* phenomenon. Put in a nutshell, a *construct* is defined with reference to a *nomological network* of other phenomena, empirical findings, and theories linking latent variables to abstract constructs (Embretson, 1983; Messick, 1995), whereas a latent variable is a mathematical construction. This often leads to confusion for applied specialists when psychometric dimensionalities of tests do not coincide with believed psychological dimensionalities, although this apparent discrepancy is perfectly expected if the precise distinction above is made.

Let us illustrate this distinction with an example. The Center for Epidemiological Studies-Depression (CES-D) is a 20-item scale introduced originally by Lenore S. Radloff to measure depressive symptoms in the general population. If we were studying the measurement properties of the CES-D via CFA models or IRT, the items would be considered indicators of a latent variable (which most researchers would call “depression”), but the latent variable is *not* depression itself as it is merely a mathematical construction. The latent variable is *related* to the construct of depression, however, which is defined as per the complex interrelations of ideas, definitions, and empirical findings in the clinical literature. Likewise, if one were empirically scoring the CES-D by summing the item responses, the resultant composite scale score is *not* depression itself either but again only *related* to that construct as an observable indicator of it. Even more precisely, the score is an indicator of the *severity* of depressive symptoms.

It should be noted, however, that the measurement literature is generally somewhat vague and inconsistent in its use of the term *latent variable*. The term has a number of different meanings in the measurement and statistics literature, each of which can lead to quite different variables. There are at least three common uses of relevance to this chapter. The first definition, which is the closest description of a (unobserved) latent variable in classical test theory, is that latent variables are real variables that could, in principle, be measured (e.g., proficiency or knowledge in a domain, such as mathematics, or level of depressive symptomatology). A second form of a latent variable is when observed scores arise by recording whether an underlying variable had values above or below fixed thresholds (e.g., a response to a Likert-type question). The former definition can be conceptualized within a framework of the latter definition, although it does not necessarily have to be so. The third definition, which is the most commonly used meaning in the social sciences, describes a latent variable as a constructed variable that comes prior to the items (or indicators) of which we measure. With item responses at hand and the use of a statistical model, one can predict a score on this latent variable for each person in the sample. This third meaning is most commonly used in factor analysis, latent variable modeling, and covariance structure models and is therefore the one used in this chapter. In terms of the psychometric approach to factor analysis, a latent variable is a reason for or summary of behavioral or cognitive manifestations. In the statistical framework, a latent variable is defined by

local or conditional independence (statistical entity with no real theoretical purpose). Statistically, it is assumed that if two variables are correlated, they have something unobserved in common (i.e., the latent variable). Therefore, uncorrelated errors (i.e., the residual correlation among the items over and above the factors) are a key defining feature of latent variable models.

Finally, it is useful to differentiate between observed-score and latent variable models. When an observed composite score is decomposed into two independent additive components, true score and error, without any further assumptions about the structure of the true score, researchers have termed this *CTT.* At the same time, different sets of assumptions about the error structure and true scores for repeated assessments and different sampling schemes for items and examinees have led to the definition of *parallel, essentially parallel, #-equivalent, essentially #-equivalent*, and *congeneric* test scores. Moreover, if no particular statistical model is assumed for the responses, models in CTT are typically referred to as *weak true-score models*, and if a statistical model is assumed (e.g., binomial, compound binomial), they are referred to as *strong true-score models*. If the relationship of the observed score to the true-score and error components is of a specific functional form that depends on at least one latent variable and can be formulated in a *generalized linear (latent variable) model framework*, we typically speak of latent variable models. Latent variables belong to the class of unobservable random variables, but they are a specific subset because their existence is *postulated*, and their *metric* is established through the specification of the model and the parameter estimation strategy. If response data are modeled at the item level, measurement specialists refer to these models as IRT models, which have become increasingly popular in the past two decades due to increasing computer power and their flexible mathematical formulation. It is interesting to note that there is no substantive theory in IRT but that, generally, the model *is* the theory, which, some argue, makes the rational link between the latent variable and the underlying construct it potentially indexes harder to establish as one can alternatively conceive of a latent variable as a mere data-processing filter that allows for ordered inferences about examinees and items (see Junker, 1999). In general, observed and latent variable frameworks benefit from one another and are compatible as, for example, methods of covariance structure analysis that are well suited to test assumptions about error structures associated with CTT.

At this point, it is important to take a small sidebar to highlight an essential difference between factor analysis (as it is commonly used) and IRT in item calibration. Although FA and IRT can be written as generalized linear latent variable models, the statistical estimation problem is compounded in IRT because the item responses are binary or ordered polytomous random variables, and the estimation strategy necessitates the estimation of the latent variable score for each individual in order to estimate the parameters of the item response function (i.e., calibrate the items). This is in stark contrast to most factor analysis models, wherein the latent variable is integrated out of the estimation equation by, in essence, marginalizing over the latent variable (i.e., focusing on reproducing the observed covariance matrix).

Once items have been calibrated, examinees have been scored, and quantifiers of measurement error have been computed, inferences are being made grounded in the mathematical model that was used. Ideally, those inferences ought to be accurate and result in *fair conferences* for the examinees and the assessment discipline. Investigations of the *degree* to which scores are consistent across administration conditions fall under the umbrella term of *reliability theory*, whereas investigations of the *degree* to which inferences made from test scores and the consequences of decisions based thereon are appropriate fall under the umbrella term of *validity theory*. Specifically, reliability is a question of *data quality*, whereas validity is a question of *inferential quality*. Of course, reliability and validity theory are interconnected research arenas, and quantities derived in the former bound or limit the inferences in the latter. This is seen explicitly in CTT statistics, for example, where it can be easily shown that a validity correlation coefficient is never greater than the square root of the test reliability coefficient. Moreover, to increase both reliability of scores and validity of inferences, a surge in models for *cognitively diagnostic assessment* has forced measurement specialists to refocus their attention on the *cognitive processes* that examinees are engaged in when responding to items. This has led to a renewed dissection of what forms of *evidence* support valid inferences and has brought the focus of investigations back to the examinees.

The title of this chapter was chosen to highlight that, when dealing with matters of reliability and validity, we are, in essence, dealing with matters of making inferences from test or scale scores. In other words, data on reliability and validity gathered in the process of measurement aid social and behavioral researchers in judging the

appropriateness and *limitations* of their inferences from the test or scale scores. In the next section, we provide an overview of reliability theory and the statistical properties of test and scale scores. In the section that follows, we provide an overview of validity theory and then end the chapter with some pointers to future developments.

4.3. A Unified Look at Reliability and Error of Measurement as a Basis for Valid Inferences

Quantifying measurement error can take different forms, depending on the scoring framework that is used for modeling the data. Traditionally, CTT has been used predominantly by test developers as well as applied specialists. In CTT, reliability is quantified using *reliability coefficients*, and uncertainty in scores is quantified using unconditional and conditional *standard error of measurement*. In recent years, the ever-growing literature on latent variable models, particularly IRT models, might seem to suggest to some that CTT models are passe. This would be an inappropriate perception of testing reality, however, fueled more by academic research practice than by testing practice across a wide range of situations, and we will thus briefly address this controversy. For example, Brennan (1998a) writes, “Classical test theory is alive and well, and it will continue to survive, I think, for both conceptual and practical reasons” (p. 6).

Nevertheless, the growing interest in IRT by theoreticians and practitioners alike over the past 30 years has been nothing short of spectacular. This is evidenced in the number of sessions at measurement and testing conferences and the large proportion of publications in measurement and testing journals devoted to theoretical developments or applications of IRT. Although it is true that IRT is frequently being used in moderate- to large-scale testing programs and projects, CTT statistics continue to be widely used in the development and evaluation of tests and measures in many areas of the educational, social, and behavioral sciences that are concerned with tests and measures of limited volume of production and distribution. For example, an overwhelming majority of tests and measures reviewed in source books such as the

Mental Measurements Yearbook series, produced by the Buros Institute of Mental Measurements, or the *Measures of Personality and Social Psychological Attitudes* book by Robinson, Shaver, and Wrightsman (1991) predominantly report CTT statistics. The primary reason for using CTT in small-volume testing programs and in research environments is the large sample sizes that are needed when one seeks to apply latent variable modeling approaches such as IRT and SEM (e.g., Bedeian, Day, & Kelloway, 1997; Bentler & Dudgeon, 1996; Junker, 1999). With observed-score measures being alive and well, it is thus worthwhile to investigate the recent developments that have taken place on these measures in the past decade. We will start appropriately with one of the oldest and most versatile indicators of score consistency, the reliability coefficient.

4.3.1. Recent Developments in the Theory of Reliability Coefficients

In the past 10 years, particularly due to the impact of increasing computer power, psychometric modeling has seen an explosion of sophisticated models that require the computer-intensive simultaneous estimation of numerous model parameters that has fueled a rethinking of the role of reliability coefficients. It is worth stating, though, that the dominating role of entities such as the information function in IRT has not changed modelers' desire for *conceptual reliability*. It has, however, changed the ways in which we look at the *mathematical formalization of reliability*.

As stated before, reliability is typically measured by a reliability coefficient, often denoted #*XX*', which in CTT or observed-score models is defined as the ratio of true-score variance to observed-score variance or the proportion of variation in the data that can be explained by differences among individuals or objects of measurement. Because the observed score is decomposed into two additive unobserved components, leading to ambiguities about the relative contribution of each unobserved component to total observed variance, the reliability coefficient cannot be computed directly. Instead, estimators have to be defined that provide reliability coefficient estimates based on test data from one or multiple measurement occasions. However, it is noteworthy that the definition of a reliability coefficient itself, in the context of multiple measurement occasions, poses subtle challenges to measurement specialists, who

have been haunted for more than 40 years by complications that arise from *difference scores*. Some have called for a ban in difference scores because of their supposed low reliability, but today this ban has been lifted. It is recognized that although the frequently cited limitations of difference scores are real, these limitations mostly hold for restrictive situations and that there are many scenarios for which difference scores are most appropriate (Zumbo, 1999b).

A reliability coefficient is a particularly natural index in observed-score models, and the definition of a reliability coefficient in latent variable models such as IRT or SEM is much more artificial. For both latent variable models and observed-score models, the formulization of conditional measurement error and information is a natural pathway that connects different models. Yet the reliability coefficient is intricately related to the error of measurement. For example, variance ratios in random-effects models prevalent in *g*-theory or the asymptotic variance of the ability trait distribution in IRT models depend directly on quantities that measure the error in the associated models. Nevertheless, the reliability coefficient itself is sometimes preferred as an index of the amount of measurement uncertainty inherent in test scores because it is unitless and is a single informative number that is practically easy to compute and included in most standard software packages (see Feldt & Brennan, 1989). Moreover, it is easily interpreted. Let us now turn to a few commonly encountered estimators of the population reliability coefficient.

4.3.2. Estimators of the Reliability Coefficient and Their Properties

> A fundamental fact concerning unreliability is that, in general, it cannot be estimated from only a single trial. Two or more trials are needed to prove the existence of variation in the score of a person on an item, and to estimate the extent of such variation if there is any. The experimental difficulties in obtaining independent trials have led to many attempts to estimate the reliability of a test from only a single trial by bringing in various hypotheses. Such hypotheses usually do not afford a real solution, since ordinarily they cannot be verified without the aid of at

> least two independent trials, which is precisely what they are intended to avoid. (Guttman, 1945, p. 256)

It is typically argued that reliability estimators fall into three distinct classes: (a) internal consistency coefficients, (b) alternative-forms reliability coefficients, and (c) test-retest coefficients. However, because reliability coefficients that involve multiple occasions for testing or rating can be estimated using intra-class coefficients, it seems more appropriate to distinguish only internal consistency coefficients and intra-class coefficients. Moreover, the intra-class coefficient in CTT is essentially a Spearman-Brown extrapolation of Cronbach's # (Feldt, 1990), which is itself the average of all split-half internal consistency correlation coefficients under appropriate model assumptions (Cronbach, 1951) and is, as such, preferred over a split-half coefficient computed for some arbitrary random split. Cronbach's # can be computed from data on a single administration of a test and does not require parallel forms, a test-retest scenario, or multiple judges for which an intra-class correlation coefficient can be used. For tests or items that are at least essentially # equivalent with uncorrelated errors, # equals the correlation coefficient, and for congeneric tests, it is a lower bound (Lord & Novick, 1968; see Komaroff, 1997).

Coefficient # is among the most commonly reported statistics in all of social and behavioral sciences. What makes it so useful to researchers and test developers? First, it provides a conservative lower bound estimate of the theoretical reliability in the worst of situations (i.e., when essential # equivalence does not hold). That is, the proportion of observed-score variance that is due to true individuals' differences is in truth at least the magnitude of coefficient #. Second, it provides this estimate without having to resort to repeated testing occasions and without necessitating parallel forms of a test. Third, it is easily computed and available on most statistical computer programs. The biggest limitation of coefficient # is that it results in an undifferentiated error of measurement. Generalizability theory, on the other hand, acknowledges that there are several sources for measurement error, which depend on the various factors modeled in the measurement experiment, and that one may want to model these various sources. Of course, it should be noted that in differentiating the error of measurement, one is actually also redefining the consistent or true-score part of the data.

It seems that Guttman's fears were not warranted and that we have overcome the problem of estimating reliability, a property of scores from repeated administrations, from scores from a single administration. Unfortunately, the situation may not be that simple if assumptions underlying the scoring model used are violated. In considering the assumptions of measurement models (and particularly uncorrelated errors), Rozeboom (1966) reminds us in his classic text on test theory that statistical assumptions are *empirical commitments:*

> However pleasant it may be to shuffle through the internal statistics of a compound test in search of a formula which gives the closest estimate of a test's reliability under conditions of uncorrelated errors, this is for practical applications like putting on a clean shirt to rassle a hog. (p. 415)

More than 35 years ago, Maxwell (1968) showed analytically that correlated errors lead to biased estimates of the correlation coefficient if an intra-class correlation coefficient is used as an estimator and argued that this bias is most likely to be an overestimate. It has been confirmed via simulation studies that Cronbach's # underestimates #*XX′* under violation of essential # equivalence and that it overestimates #*XX′* if errors are correlated (Zimmerman, Zumbo, & LaLonde, 1993; see Raykov, 1998b, for composite tests and Zumbo, 1999a, for a simulation framework), but these effects can be partly attenuated if both assumptions are violated simultaneously (Komaroff, 1997). Nevertheless, it appears that # is relatively robust against moderate violations of these assumptions (see Bacon, Sauer, & Young, 1995; Feldt, 2002). Similar results have been found for *g*-theory designs with multiple time points. In such designs, underestimation was present for uncorrelated errors with increasing variances over time, overestimation was present for correlated errors with equal variances over time, and both directions of estimation bias were present for correlated errors with unequal variances over time (Bost, 1995). It is important to note that correlated errors may arise for a variety of reasons. Given the advent of new item formats, one of the most common reasons for correlated errors is linked items. That is, historically, measurement specialists have advocated that items be disjoint statements that would not result in extra covariation in latent variable modeling due to item format. Items that are linked, however, may induce extra covariation among the items that appear as correlated errors (for an example, see Higgins, Zumbo, & Hay, 1999). We recommend that

researchers faced with correlated errors arising from item format see Gessaroli and Folske (2002) for a useful, yet general, approach for estimating reliability.

In latent variable modeling, correlated errors are equivalent to introducing an additional latent variable (i.e., factor) that loads on the manifest variables (e.g., MacCallum, Wegener, Uchino, & Fabrigar, 1993; Raykov, 1998a). Today, FA methods, particularly CFA, continue to be useful tools to assess the degree of correlated errors (e.g., Reuterberg & Gustafsson, 1992) and have recently been used to construct adjusted #s that reduce and sometimes eliminate the inflation effect (Komaroff, 1997). Moreover, SEM allows for the estimation of a reliability coefficient for congeneric tests that is not a lower bound for the true reliability coefficient (unlike Cronbach's #) (Raykov, 1997a), along with a bootstrap estimation of its standard error that does not depend on normality assumptions (Raykov, 1998b). Unfortunately, large sample sizes are required for the stable estimation of model parameters, and not all estimation methods are recommendable (see Coenders, Saris, Batista-Foguet, & Andreenkova, 1999). Researchers need to be aware of the additional assumptions that are required for proper estimation in a covariance structure analysis (Bentler & Dudgeon, 1996). Among these are multivariate normality of the response data required for some estimation approaches, which is unlikely to hold for categorical data, and large sample sizes required for asymptotic theory, which are unlikely to exist for small-scale assessments.

Estimating reliability coefficients and assessing model assumptions has also been done for more than three decades using FA methods (e.g., Feldt, 2002; Fleishman & Benson, 1987; Jöreskog, 1970, 1971; Kaiser & Caffrey, 1965). It has been shown repeatedly that the assumption of uncorrelated errors, coupled with unidimensionality and the use of the simple total score in observed-score modeling, corresponds to an orthogonal factor model with a single dominant factor that has loadings for each item in the test. Under this model, the reliability coefficient is estimated as the sum of squared loadings (i.e., the communalities) divided by the sum of squared loadings plus error loadings (i.e., communalities plus unique variances).

Along with FA models, SEMs allow for flexible testing of multiple assumptions such as type of model (i.e., parallel, # equivalent, congeneric), correlation of errors, invariance across time, and invariance across subgroups (e.g., Feldt, 2002; Fleishman & Benson, 1987; Raykov, 1997a, 1997b, 1998a, 1998b, 2000, 2001). In an SEM framework,

the reliability coefficient can be estimated as an internal parameter or an external parameter of the model, and test or item weights can either be preset by the investigator or estimated as factor loadings simultaneously with all other model parameters. The general approach for testing assumptions about error structures using SEM requires at least four items or tests due to the identification requirements of the model so that all hypothesis tests, including the one about congenerity, can be performed (e.g., Raykov, 1997a). In addition to coefficient #, the omega coefficient with equal and unequal weights has been proposed; unequal weights are preferred by some authors because the coefficient never increases when items are dropped. Note, however, that reliability estimates are not necessarily recommended as sole yardsticks for test construction (Bacon et al., 1995). More recently, SEM has been advocated by some to model the type of correlation structure via integrated time-series models, but the practical utility of that approach remains limited at this point (Green & Hershberger, 2000). Finally, note that, just as attenuated correlation coefficients have been shown to be sensitive to the true-score distributions for examinees (Zimmerman & Williams, 1997), coefficient # is sensitive to the score distribution of examinees, which has led to the proposal of a robust generalization of # that is insensitive to tail fluctuations in this distribution (Wilcox, 1992).

So what is a practitioner to do when coefficient # needs to be estimated? It appears that for small sample sizes, sophisticated latent trait models would provide unreliable results, and the effort of estimating these is probably not worth it. If the sample size is large (e.g., at least 200 examinees for moderate tests as a guiding principle) and one has complex item formats, then latent trait models such as SEM may be useful to estimate reliability and related quantities. It is important to always be aware of the model assumptions that are lurking in the background when choosing a particular scoring model (Zumbo, 1994), however, and for larger sample sizes and high-stakes assessment scenarios, these should be investigated to obtain the most accurate estimate of reliability and measurement error. We recommend Gessaroli and Folske's (2002) approach.

4.3.3. Hypothesis Tests for Reliability Coefficients

The intra-class correlation coefficient, which can be used for test-retest, parallel forms, subtest, and inter-rater reliability, has found wide applications in social and behavioral research (Alsawalmeh & Feldt, 1992). Its distribution theory and the distribution theory for Cronbach's # have recently been developed in more detail (Feldt, 1990; van Zyl, Neudecker, & Nel, 2000). Hence, approximate tests have been developed for two independent intra-class reliability coefficients (Alsawalmeh & Feldt, 1992), two independent coefficient #s (Alsawalmeh & Feldt, 1999; Charter & Feldt, 1996), and two dependent coefficient #s (Alsawalmeh & Feldt, 2000). Similarly, tests for disattenuated correlation coefficients can be easily formulated in an SEM framework (Hancock, 1997).

Note, however, that not all distributional results are easily applicable across a wide range of situations. For example, the asymptotic distribution of the maximum likelihood (ML) estimator of # derived by van Zyl et al. (2000) requires no assumptions about the covariance structures of the items; yet, as an asymptotic result, it requires large sample sizes. Furthermore, the multivariate normal distribution of the item response data is unlikely to hold for dichotomously scored items.

Because the meaningful interpretation of hypothesis test results depends on the power of the test, it is essential to understand that the power of a test is not a function of the reliability coefficient but a relation of it (Williams, Zimmerman, & Zumbo, 1995; Zimmerman, Williams, & Zumbo, 1993a, 1993b). As these authors remind us, power is a function of the absolute value of observed variance, and its relative decomposition is irrelevant, even though it influences the magnitude of the reliability coefficient. However, formulas for computing the power and required sample size of a test for comparing coefficient #s for two populations can indeed depend on the direct magnitude of the respective sample values for the coefficient #s due to the sampling theory involved (Feldt & Anken-mann, 1998). In summary, the class of statisticaltests for population reliability coefficients has been broadened, and even though the individual papers need to be referred to for the exact ways of conducting the tests, these tests are often not difficult.

4.3.4. Maximizing Reliability Coefficients and Composite Scores

It has long been acknowledged that Cronbach's # is not an indicator of test homogeneity or unidi-mensionality (e.g., Green, Lissitz, & Mulaik, 1977; Miller, 1995), and violations of the assumption of test homogeneity have been researched (e.g., Feldt & Qualls, 1996). If tests are measuring several related constructs, modelers in CTT deal with this by constructing composite test scores that receive appropriate weights using a table of specifications. Using a composite-score analysis instead of a total-score analysis may have a strong effect on the reliability estimate for the data, though. Formulas exist, most commonly for congeneric tests, which maximize reliability measures under different conditions (e.g., Armstrong, Jones, & Wang, 1998, for coefficient #; Goldstein & Marcoulides, 1991, and Sanders, Theunissen, & Baas, 1989, for generalizability coefficients; Knott & Bartholomew, 1993, for a normal factor model; Li, 1997, for a composite score; Li, Rosenthal, & Rubin, 1996, for cost considerations; Rozeboom, 1989, for using regression weights on a criterion variable; Segall, 1996, for linearly equated tests; Wang, 1998, for congeneric models).

Maximizing reliability is akin to determining the ideal sample size for a designed experiment under power considerations, and so, just as in traditional statistical design, practical consideration will eventually be the ultimate determining factor for test construction or the analysis method as some tests proposed to maximize reliability seem to have unrealistic characteristics (e.g., 700 multiple-choice items; see Li et al., 1996). In addition, most formulas for composite reliability coefficients require knowledge of the componential reliability coefficients. If reliability information is not available on the subcomponents that are supposed to be weighted, a multivariate covariance structure analysis approach may be called for, and formulas for weights that maximize reliability have been derived for some cases (Wang, 1998).

Coefficient # and intra-class correlation coefficients are not the only means of indexing measurement precision. In fact, they are only single numbers that capture the quality of the scores in a rather superficial sense. To obtain more precise information about how

measurement error actually affects the scores and hence decisions about examinees, we need to turn to score-level measures of precision.

4.3.5. Local Estimates of Precision in Scores

Scoring test data eventually brings about consequences for examinees. These consequences are mathematically dependent on accurately estimating the error associated with examinees' scores, which is most crucial for examinees with an observed score somewhere around the cut-score in criterion-referenced assessment or along the entire conntinuum for norm-referenced assessment. It has long been recognized that the score error is not constant along the continuum, even though in early work in CTT, unconditional raw-score SEM was reported and used. However, responsible data analysts and decision makers are aware that score error varies along the ability continuum, and more evidence from different estimation methods has been accumulated in the past decade to support this. Generally speaking, for observed-score models, curves depicting the conditional SEM will be somewhat inverse U-shaped, with smaller standard errors near the upper and lower tails of the true-score continuum and larger standard errors in the center of the true-score continuum. In contrast, the local precision curve for a test analyzed via IRT methods has the opposite, regular U shape. That is, there is less error in the center of the latent continuum near the point of maximum test information and more error for extreme values on the latent continuum. Thus, local measures of precision need to be considered in observed and latent variable models. Moreover, it is clear that a *conditional raw-score standard error of measurement* (*CRS-SEM*) should be used for fair decision making based on raw scores and that a *conditional scale-score standard error of measurement* (*CSS-SEM*) should be reported if raw scores are transformed via linear or nonlinear transformations to some other practically meaningful scale such as the percentiles, grade point equivalent, or stanine scales.

Although in the 1989 chapter by Feldt and Brennan, CRS-SEM only received a two-page treatment nested within a section on “special” issues in reliability and CSS-SEM was not discussed in much detail, during the past decade, researchers in the field of

measurement have produced a series of papers that meticulously investigated different approaches to estimating local or conditional standard errors for scoring models on different scales and the behavior of these approaches in different calibration situations (e.g., Brennan, 1998b; Brennan & Lee, 1999; Feldt, 1996; Feldt & Qualls, 1996, 1998; Kolen, Hanson, & Brennan, 1992; Kolen, Zeng, & Hanson, 1996; Lee, 2000; Qualls-Payne, 1992; see also May & Nicewander, 1994). In general, most methods produce similar results that lead only to slight differences in confidence interval width if the conditional standard errors are used for their construction. As usual, CTT methods are comparatively easier to compute and do not rely as heavily on larger sample sizes for stable parameter estimation.

From earlier discussions, it should be clear that the explicit treatment of specific error structures in scoring models has been one of the most important contributions in the past decade. Within an observed-score context of conditional standard error, this has most notably resulted in a synthesis of conditional standard error estimation approaches for *g*-theory designs and estimations that include CTT scenarios as special cases (Brennan, 1998b). Within a latent trait framework, the dependency of responses for items presented with the same stimulus in testlets has driven researchers to develop a Bayesian estimation framework for dichotomous and polytomous items on the same test scored with IRT models (Bradlow et al., 1999; Wainer, Bradlow, & Du, 2000; Wainer & Thissen, 1996; Wainer & Wang, 2001; Wang et al., 2002; see also Sireci, Thissen, & Wainer, 1991, for reliability estimation as well as Lee & Frisbie, 1999, for a *g*-theory approach). These studies have found that incorporating testlet effects into an IRT model or *g*-theory model always improved estimation accuracy by incorporating within-testlet response pattern information into parameter estimates and is necessary if strong testlet effects are present to prevent biased ability estimates and thus incorrect decisions. This conclusion was further supported in a direct comparison of CRS-SEM estimates with models that accounted for testlet effects producing more accurate CRS-SEM under all conditions, even though *g*-theory estimation, as an alternative to IRT testlet models, worked well under mild testlet dependencies (Lee, 2000; see also Lee & Frisbie, 1999). Again, the message is that for larger sample sizes, it is particularly important to assess whether model assumptions are likely to hold, but for both smaller and larger sample sizes, conditional standard errors should be computed and used for decision making. It appears that the particular method for computing CRS-SEM or CSS-SEM does not

matter much for most practical decisions and that the one that is simplest to implement should be chosen.

4.3.6. Relationships Between Error Estimates in Different Scoring Frameworks

We want to close the discussions about reliability and measurement error with a section on relationships between observed-score and IRT models as some concepts are often confused. As we have just seen, the notion of a local measure of precision, which is captured by the information function in IRT, also exists in CTT through conditional standard errors for raw and scale scores. Moreover, it is similarly possible to compute information functions in CTT (Feldt & Brennan, 1989; Mellenbergh, 1996) as well as unconditional standard errors and reliability coefficients in IRT (e.g., Samejima, 1994). In particular, the IRT equivalent to the unconditional standard error in CTT is the expectation of the asymptotic conditional standard error:

$$\mathrm{SEM} = \sigma_\varepsilon = \int_{-\infty}^{\infty} [I(\theta)]^{-1/2} f(\theta) d\theta.$$

For practical estimation purposes, the information function in the above equation is replaced by the estimated test information function, and the ability distribution can be empirically estimated if conditional unbiasedness of # holds; otherwise, test information functions adjusted for bias should be used (Samejima, 1994). The reliability coefficient can now be predicted from a single administration of a test using the observed variation in # 9 and the estimated standard error as described above (in the formula, SEM indicates standard error):

$$\hat{\rho}_{\hat{\theta}_1,\hat{\theta}_2} = \frac{\hat{\mathrm{Var}}(\hat{\theta}) - \hat{\mathrm{SEM}}}{\hat{\mathrm{Var}}(\hat{\theta})} = \frac{\hat{\mathrm{Var}}(\theta)}{\hat{\mathrm{Var}}(\hat{\theta})}.$$

The relationship between the multiple-occasion estimators of the reliability coefficient in CTT and IRT models has been investigated for some time, and some authors even go so far as to declare the reliability coefficient redundant (Samejima, 1994, p. 243). This statement seems a bit extreme because the appropriateness of an IRT estimate of reliability depends on the accuracy of the fitted model (see Meijer, Sijtsma, & Molenaar, 1995, p. 334, for this argument in a nonparametric context), and fitting a more complex IRT model may require more data than are available at a given moment. In addition, even though in IRT, standard errors are larger at the extreme ends of the scale (Lord, 1980), this is dependent on the choice of transformation from the true score to the latent trait scale, and dramatic differences between conditional standard errors can be observed for different choices of transformation (see Brennan, 1998b).

As another similarity between CTT and IRT models, recall that the reliability coefficient in CTT is the ratio of true-score variance to total observed variance or the ratio of signal to signal plus noise. Put differently, the signal-to-noise ratio equals the correlation coefficient divided by 1 minus the correlation coefficient. Therefore, a local reliability coefficient can be defined as a function of the item information function, which is itself proportional to the local signal-to-noise ratio (Nicewander, 1993).

Conditional standard errors for absolute decisions (and thus dependability coefficients) or relative decisions (and thus generalizability coefficients) can also be formulated in *g*-theory (Brennan, 1998b, 2001). In *g*-theory, the class of model specifications, albeit all generalized linear models (GLIMs), has been enlarged, but typically larger sample sizes are required for accurate estimation of variance components. In IRT, the class of GLIMs uses different link functions, but choices have to be made now between logit and probit models, the number of parameters in the model, and whether to choose a parametric or nonparametric formulation. In the latter case, reliability estimation is not even common practice, and even though a reliability coefficient that is related to a scalability coefficient can be estimated in Mokken's nonparametric alternatives to the Rasch model, their complementary uses remain unclear (Meijer et al., 1995; Meijer, Sijtsma, & Smid, 1990).

Finally, it needs to be highlighted that one of the advantages of reliability estimation in CTT is the relative simplicity of the model, whose only major alternatives consisted of different assumptions about its unobserved components. Claims that CTT is merely a special case of IRT (Nicewander, 1993) or FA seem to be overstatements and

seem to ignore the difference between score-level and item-level modeling, as well as between a latent variable and a more general unobserved variable such as the true score in CTT. To the contrary to the overstatement, it can be argued that IRT is a first-order approximation to CTT. The overstatement also ignores the role that parameter estimation strategy has in defining a psychometric model. Put simply, the liberalization of CTT into *g*-theory—along with its reformulation and extension, in latent variable terms, in FA and SEM—and the advent of IRT have come at the price of stronger requirements on the data, which have affected reliability estimation. For larger sample sizes, we can definitely investigate more complex assessment scenarios through *g*-theory, as well as more complex dependency structures through FA and SEM, and achieve invariance properties for adaptive testing in IRT (see Rupp, 2003; Rupp & Zumbo, 2003, in press, on quantifying a lack of invariance in IRT models), but for smaller sample sizes, these advances are often of limited benefit to the practitioner. In addition, no matter how sophisticated the model statement and estimation routines have become, a test of poor validity and thus poor conceptual reliability will always remain unaltered. This brings us to our final section.

4.4. Validity and the Practice of Ralidation

Validity theory aids us in the inference from the true score or latent variable score to the construct of interest. In fact, one of the current themes in validity theory is that construct validity is the totality of validity theory and that its discussion is comprehensive, integrative, and evidence based. In this sense, construct validity refers to the degree to which inferences can be made legitimately from the observed scores to the theoretical constructs about which these observations are supposed to contain information. In short, construct validity involves generalizing from our behavioral or social observations to the *concept* of our behavioral or social observations. The practice of validation aims to ascertain the extent to which an interpretation of a test is *conceptually* and *empirically* warranted and should be aimed at making explicit hidden ethical and social values that influence that process (Messick, 1995).

It is hard not to address validity issues when one is discussing errors of measurement. Yet the developments in validity theory have not been as dramatic over the past 15 years as have been the developments in reliability estimation and measurement model

development. For a cursory overview, several papers are available that describe important current developments in validity theory (Hubley & Zumbo, 1996; Johnson & Plake, 1998; Kane, 2001). In brief, the recent history of validity theory is perhaps best captured by the following observations.

- As one can see in Zumbo's (1998) volume, there is a move to consider the *consequences* of inferences from test scores. That is, along with the elevation of construct validity to an overall validity framework for evaluating test interpretation and use came the consideration of the role of ethical and social consequences as validity evidence contributing to score meaning. This movement has been met with some resistance. In the end, Messick (1998) made the point most succinctly when he stated that one should not be simply concernedwith the obvious and gross negative consequences of score interpretation, but rather one should consider the more subtle and systemic consequences of “normal” test use. The matter and role of consequences still remains controversial today and will regain momentum in the current climate of large-scale test results affecting educational financing and staffing in the United States and Canada.
- Although it was initially set aside in the move to elevate construct validity, criterion-based evidence is gaining momentum again in part due to the work of Sireci (1998).
- Of all the threats to valid inferences from testscores, test translation is growing in awareness dueto the number of international efforts in testing and measurement (see Hambleton & Patsula, 1998).
- The use of cognitive models as an alternative to traditional test validation is gaining a great deal of momentum. One of the limitations of traditional quantitative test validation practices (e.g., factor-analytic methods, validity coefficients, and multitrait multi-method approaches) is that they are descriptive ratherthan explanatory. In other words, they are statistical and not psychological. Models for cognitively diagnostic assessment, particularly the work of Susan Embretson and Kikumi Tatsuoka, has expanded theevidential basis for test validation as well as the nomo-thetic span of the nomological network. The basicidea is that if one could understand why an individual responded a certain way to an item, then that would goa long way toward bridging the inferential gap betweentest scores and constructs.

Given that cognitive models present one of the most exciting new developments with implications for validity theory, the next section discusses them in more detail.

4.4.1. Cognitive Models for a Stronger Evidentiary Bases of Test Validation

It is informative to start this discussion by addressing the use of the term *cognitive psychology* in the literature on cognitive models. For many assessment situations, researchers use the word *cognition* to loosely refer to any process that is somehow grounded in our minds and therefore eventually our brains. Yet there is little doubt that measurement specialists are not interested in the biological or neuroscientific bases of cognitive processes for typical cognitively diagnostic assessments, so that we really often mean a “soft” form of cognitive psychology in a measurement context.

Some will undoubtedly argue that it is not the job of a psychometric data modeler to worry about what is done with the numerical estimates once they are handed down, but it is exactly this neglect of meaningful inferences at the expense of sophisticated estimation techniques that has often eradicated the psychology in “psycho”-metrics. The realization that it is time to put psychology back into the equation so that investigators who desire “reliable” tests are assured by modelers that their data do indeed provide evidence for dependable meaningful inferences. To appreciate the relevance and importance of cognitive models, one has to understand that we have not made many significant advances toward explicit validation of the inferences drawn from test scores through mathematical models. This holds true despite the injection of a latent continuum that allows modelers to extract information from test data more flexibly and accurately at the item level in IRT. For example, Junker (1999) suggests that

> despite the persistence of the “latent trait” terminology in their work, few psychometricians today believe that the latent continuous proficiency variable in an IRT model has any deep reality as a “trait”; but as a vehicle for effectively summarizing, ranking, and selecting based on performance in a domain, latent proficiency can be quite useful. (p. 10)

The main goal of modeling test data should always be to make valid inferences about the examinees, but the validity of these inferences cannot be mechanically increased by inducing latent constructs into the data structure.

Cognitive models seek to explicitly represent the cognitive processes that examinees engage in when responding to items through parameters in mathematical models, which typically consist of augmented IRT models, classification algorithms based on regular IRT models, or Bayesian inference networks that have IRT models as a central component. One approach to cognitively diagnostic assessment is the *rule-space methodology* that attempts to classify examinees into distinct attribute states based on observed item response data, an appropriate IRT model, and the attribute specification for the items (Tatsuoka, 1983, 1991, 1995, 1996; Tatsuoka & Tatsuoka, 1987). Despite a lack of consensus in the literature about what is meant exactly by an *attribute* and the sensitivity of the classification to the appropriateness of the chosen IRT model, this approach forces test developers to specify prerequisite cognitive characteristics of examinees—ideally before designing a test (Gierl, Leighton, & Hunka, 2000). Other approaches based on item attribute incidence or *Q*-matrices have been developed (e.g., DiBello, Stout, & Roussos, 1995), but their main weakness to date remains the vagueness and lack of guidance in attribute specification (e.g., Junker & Sijtsma, 2001). Developments in cognitive models have often taken place primarily in educational achievement and psychoeducational assessment contexts, though. An exception was Zumbo, Pope, Watson, and Hubley (1997) in personality assessment, in which they studied the relation of the abstractness and concreteness of items to the psychometric properties of a personality measure. Other advances are currently made in the development of simulation-based assessment software that emphasizes a deeper and richer understanding of the cognitive processes required for performing certain tasks in which data are analyzed through Bayesian networks (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999).

More sophisticated models for cognitive assessment do not come without a price. One of the components of this price is again sample size because more complex IRT models, cognitive state models, or Bayesian inference networks typically require a larger number of parameters to be estimated. More important, however, the more useful models for cognitively diagnostic assessment are built on a solid understanding of the cognitive processes underlying the tasks that are being assessed. As an excellent

example, consider the work by Embretson (1998), who used the cognitive process analysis of the Raven's Advanced Progressive Matrix test by Carpenter, Just, and Shell (1990) to model examinees' responses, extract diagnostic information, and generate similar items. Comprehensive models of cognitive abilities are still relatively rare, and even though advances have been made, it is necessary to note that their most important cornerstone, the analysis of cognitive processes, is still their weakest element.

The issue is less a lack of models for new kinds of test data but rather a lack of awareness in the applied world that these models exist along with a mismatch of assessment instruments and modeling practice. In other words, if test developers are interested in providing examinees and institutions with richer profiles of abilities and developmental progress, the nature of the assessment methods has to change to provide richer data sets from which relevant information can be more meaningfully extracted. What is meant by *more meaningful* will, of course, in the end depend on the use of the assessment data, but in general, authorities in the field are nowadays beginning to agree that we need more than simple test responses scored 0 and 1 to validate the inferences that are being made from the test data. As Embretson's (1998) work demonstrates, the key to useful cognitive models is that they need to be explanatory and not just another set of descriptive models in cognitive terms rather than mathematical terms (Zumbo & MacMillan, 1999). Put differently, a change of terminology is insufficient to claim true advances in gathering more meaningful and weighty validity evidence.

A similar push for explanatory power has also taken place in the area of differential item functioning, where attitudinal, background, and cognitive variables are used to account for differential achievement profiles to investigate the inferential comparability of scores across populations (Klieme & Baumert, 2001; Watermann & Klieme, 2002). The developments that are currently taking place serve in part as a consciousness-raising device to help test developers and users to reflect more closely on how valid their inferences from test data really are and how these inferences can be improved. This continues the path toward a comprehensive and unified validation process of assessment instruments that has been eloquently laid out by Messick (1989, 1995).

4.4.2. Implications of Cognitive Models for Modeling Novel Dependency Structures

In traditional psychometric models, dependencies among item responses over and above what can be accounted for by the unobserved variables have been a dreaded feature of test data, and every effort has always been made to eliminate this dependency through test design or modeling efforts. This may be the wrong lens that is applied to the data, and it appears that cognitively diagnostic assessments—along with models for testlet structures and more complicated error dependencies—are the new figures that are slowly taking shape under a new perspective on items and responding to items. We have begun to shift our thinking back to the individual examinees because we are starting to realize that the goal of any assessment, be it strictly cognitively diagnostic or not, is to arrive at better inferences about examinees' abilities. Furthermore, item difficulty and discrimination are properties of the examinees that respond to the items because the items are windows into the minds of the examinees and are not qualities inherent in items independent of populations of examinees.

All of this is to say that the current push toward cognitively diagnostic assessment seems to be more than just an extension of currently existing models and statistical methodologies to richer domains. In fact, it is our chance to clean our windows into the minds of examinees and to refocus our lenses toward the examinees as the unit of investigation that matters most. From a mathematical perspective, this means looking for different types of information in data structures that may posit new challenges to the modelers. In particular, if cognitive processes are highly interrelated in complex neural networks at a biological-chemical level, then we can expect that item responses are probably also interrelated to a much higher degree than gives us comfort. Indeed, what we need is an extension of the models that are currently used in covariance structure analysis because the future seems to lie in accepting covariation and interrelationships rather than dreading them.

This can be seen not only in the models and scenarios discussed so far but also by looking at the variety of item types that can be found in new tests across multiple

disciplines (Zenisky & Sireci, 2002). As these authors show, traditional test formats have been augmented with a whole new battery of items that require the test taker to engage in more sophisticated complex cognitive processes. We certainly have choices when scoring these item types as we really also have when dealing with the items that are used in cognitively diagnostic assessments. We could theoretically score them all 0–1 or on a simple graded scale and apply traditional models in CTT, *g*-theory, or IRT to the responses. We might find, however, that dependency structures in the data sets might compromise our simple analyses because the items are not isolated items anymore. Indeed, to use such items more successfully, it would make much more sense to focus on the interdependencies and go from there.

It should also be noted that the nature of dependencies that are deliberately build into more complex item types has crept up with traditional tests as well. For example, researchers have been busy investigating the data structure for CTT models in terms of the degree of test parallelism. As one dimension of complexity, researchers have defined parallel, essentially parallel, #-equivalent, essentially #-equivalent, and congeneric tests; as a second dimension, they consider uncorrelated and correlated errors; and as a third, they investigate sampling type (i.e., Type 1, Type 2, Type 12). With all these considerations at hand, psychometricians have been busy trying to find the best estimators of quantities such as the reliability coefficient or conditional standard errors for different data structures. Nevertheless, we are faced these days with data structures that do not adhere to any of the criteria above (e.g., spherical covariance matrices; see Barchard & Hakstian, 1997; Hakstian & Barchard, 2000), which compel us to search for better descriptions of the data structure at hand.

4.5. Conclusion

The emphasis in this chapter has been on measurement error, reliability, and validity through the lens of scoring data from tests within a particular scoring framework. We have highlighted on several occasions the distinctions between observed variable frameworks (i.e., CTT and *g*-theory) and latent variable frameworks (i.e., EFA, CFA, SEM, and IRT). We believe it is important to understand that the use of a particular scoring model always remains the choice of the data analyst and is not necessitated by the data. More often than not, the choice of a particular scoring model is the result of

personal beliefs, training, and working conventions (Rupp & Zumbo, 2003). Yet it has severe consequences for how we define, quantify, and use measurement error and the decisions that we base thereon. Choosing a scoring model is an *empirical commitment* that demands the data analyst take responsibility for the consequences imparted on the examinees by this choice.

To underscore this responsibility one last time, consider for a moment a few issues that can arise with some popular scoring models. When working within a latent variable framework, it is certainly irresponsible to blindly fit IRT models to any kind of data—even if the models formally match the type of scores given (e.g., dichotomous, polytomous)—without ensuring that *sufficiently large* and *representative* calibration samples are available so that *stable* and *representative* parameter estimates can be obtained. If the parameters are not well estimated, decisions will be biased. In addition, if the intention is to use a one-shot calibration at one point in time with one set of examinees, it is logically inconsistent to justify the use of an IRT model because model parameters possess the feature of *invariance*. Invariance refers to the identity of item and examinee parameters from *repeated* calibration for *perfect* model fit and is not needed in this case. Hence, it should not be cited as the *primary* reason for using such a model.

Another example comes from the area of cognitively diagnostic assessment. Without any detailed data collected on examinees and any detailed attempts to develop realistic processing models, truly cognitively diagnostic assessment is not possible. In addition, an augmented IRT model for cognitive assessment needs to be *judiciously* chosen based on the cognitive theory underlying the test response processes and not simply because it is an interesting extension of basic IRT models (for excellent examples, see Embretson, 1998; Maris, 1995).

In the area of observed-score modeling, it is equally irresponsible to use the unconditional raw-score standard error when a large body of evidence has shown for years that CRS-SEM varies along the score continuum. Similarly, CSS-SEM needs to be computed separately if scores are transformed to scales such as stanine, percentile, or grade-equivalent scales as it also varies and is generally not equal to the CRS-SEM. Using inappropriate measures of error can lead to incorrect and unfair decisions for some, if not most, students. On a more subtle level, most observed-score scoring methods rely on assumptions about the score matrix such as parallelism, essential

equivalence, or congenerity. In some cases, failing to adjust reliability coefficients or other measures of error to the right model can lead to biased statements about a test, overconfidence in test use, and unfair decisions about examinees. Moreover, factor-analytic procedures and software can nowadays easily be used to test for these assumptions and to produce appropriate error estimates for larger sample sizes.

It is the responsibility of mathematically trained psychometricians to inform those who are less versed in the theory about the consequences of their decisions to ensure that examinees are assessed fairly. Because models (which, in part, include the parameter estimation strategy) are empirical commitments, it is measurement specialists who need to take partial responsibility for the decisions that are being made with the models they provide to others. Everyone knows that a useful and essential tool such as an automobile, a chainsaw, or a statistical model can be very dangerous if put into the hands of people who do not have sufficient training and handling experience or lack the willingness to be responsible users.

All of this is not to say that decision-making disasters will immediately occur if the above things are not adhered to in the fullest. However, it can also be too tempting to take that exact fact to be less stringent and less careful about our practices, and we believe it is important that we all in the psychometric community work together to ensure fair and sound decision making. Technological advances have opened up doors for us to do more sophisticated and complex simulation work, analyze richer and more nested data structures than ever before, and synthesize findings across analyses. At the same time, it is important to remember that examinees are typically not interested in the particular scoring models used for obtaining their score but rather in a fair assessment, which simply translates to fair decisions based on their responses. The term *fair* is of course heavily value laden and can take on different shades of meaning for different examinees, but nevertheless responsible data models consider the consequences of test score interpretation for which they provide the numerical ingredients.

Numerous questions about the reliability of tests have been asked in the past decade, and important advances have been made in the area of estimating conditional standard error for nonlinear scale transformations, estimating bias of reliability coefficient estimates such as coefficient # under simultaneous violations of assumptions, deriving algorithms for maximizing #, deriving tests for #s from different populations, and

establishing relationships between CTT, *g*-theory, IRT, and SEM that show the inter-relatedness of these procedures. In other words, we have been able to make convincing arguments for the unification of measurement models (see McDonald, 1999; Rupp, 2002; Zimmerman & Zumbo, 2001), and we have made convincing arguments for advantages of *g*-theory over CTT, IRT over *g*-theory, IRT over CTT, and SEM over *g*-theory, CTT, and IRT and so on. Important research in this area still needs to happen, and a wealth of unanswered research questions can be found in the concluding sections of the more than 100 articles that we could find in journals over the past 10 years.

We believe that this is fruitful work but that it is at least as important to reflect on our testing practice in the new millennium. Cognitively diagnostic assessments will play an important part, but we believe that they will neither replace traditional assessments entirely in the near future nor answer all of the problems encountered by psychometricians at the moment. But they are the psychometric discipline's way of pointing out that data modelers are ready to face new challenges posed by the need for richer information about examinees, concurrent new item types, redefinitions of the construct of an item itself, and a higher degree of inter-relatedness of responses from a mathematical as well as from a soft cognition perspective. Reliability and validity will always be important in test development. Reliability indices are not irrelevant, as some proclaim, because they serve different purposes than conditional SEM and test information functions, and validity will always be the cornerstone of test development and use, particularly if we move to a more unified test development–data modeling–test use process. Measurement specialists are beginning to talk and reach out to each other more and more across disciplines and cultural boundaries. Content experts, psychometric data analysts, and cognitive psychologists may not always be at the same table yet, but at least they are more often pooling their expertise in the same metaphorical room, and that is certainly a good thing. We are far from a *practical* revolution in testing, but we seem to be at an exciting juncture for pausing and reflecting on what to focus on.

REFERENCES