Center for Advanced Studies in
Measurement and Assessment

CASMA Research Report

Number 8

Some Perspectives on Inconsistencies Among Measurement Models

Robert L. Brennan[†]

July, 2004

[†]Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 297 Lindquist North, College of Education, University of Iowa, Iowa City, IA 52242 (email: robert-brennan@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Fax: 319-384-0505
Web: www.uiowa.edu/~casma

Contents

Abstract iv

Introduction 1

Some Background and History 1

Classical Test Theory . . . 1
Generalizability Theory . . . 3
Item Response Theory . . . 5
Equating/Linking . . . 8

Inconsistencies Across Models 8

What Constitutes a Replication? . . . 9
What are True Scores? . . . 10
What is Error? . . . 11
How Should Tests be Scored? . . . 14
How Should Scores from Various Tests be Combined? . . . 15

Concluding Comments 16

References 18

Abstract

Almost always the psychometric tasks associated with a large-scale testing or assessment program involve the use of several psychometric models such as classical test theory, generalizability theory, item response theory, and equating/linking models. On a superficial level, these models sometimes seem to provide nothing more than different ways to address the same issues. However, a deeper consideration often reveals inconsistencies or ambiguities that have been considered only occasionally in the literature.

After brief discussions of each of these four models, inconsistencies or ambiguities among them are illustrated in the context of five questions: what constitutes a replication; what are true scores; what is error; how should tests be scored; and how should scores from various tests be combined? These are not an all-inclusive set of relevant questions, and the discussions provided are not intended to be particularly extensive. However, these questions represent fundamental concerns in measurement, and the discussions illustrate some important differences among the models in how they approach measurement issues.

Introduction

Almost always the psychometric tasks associated with a large-scale testing or assessment program involve the use of several psychometric models to address various tasks. On a superficial level, these models sometimes seem to provide nothing more than different ways to address the same issues. A deeper consideration of the models, however, often reveals discontinuities, inconsistencies, or ambiguities that potentially threaten the inferences drawn about the psychometric characteristics of the testing program. For the most part, these problems have been ignored to date—in both the theoretical literature and in actual practice. The purposes of this paper are to: (a) provide a brief overview of various measurement models focusing primarily upon some of the more salient similarities and differences in the model assumptions; and (b) to discuss a few of the inconsistencies among models that have both theoretical and practical implications. The models considered are: (1) classical test theory; (2) generalizability theory; (3) item response theory; and (4) equating/linking models. It might be more appropriate to characterize equating/linking as a methodology rather than a model, but we overlook this terminological issue here for the sake of simplicity.

Some Background and History

It is relatively rare for a paper or book chapter to treat more than one measurement model. Sometimes two are discussed (e.g., Feldt & Brennan, 1989, discuss classical theory and generalizability theory; Bechger, Béguin, Maris, & Verstralen, 2003, discuss classical test theory and item response theory). Almost never is there a discussion of three or more models (but see Nugent & Hankins, 1992). Yet, many if not most large-scale measurement programs employ all of the models discussed here.

Furthermore, almost certainly, just about all states will employ all of these models in their efforts to satisfy the requirements of the No Child Left Behind Act (NCLB, 2002). For example, classical theory is likely to be used to provide typical reliability coefficients (e.g., Coefficient alpha and its associated standard error of measurement). Generalizability theory is likely to be used to provide reliability-like coefficients and *SEMs* for writing assessments and other types of performance assessments. Item response theory is likely to be used to characterize items, select items for a test, and perhaps even score tests. Equating/linking procedures will be necessary to relate scores on a current year's assessment to scores on a previous year's assessment, as well as to link scores across various grades. Some approaches to equating/linking rely on descriptive statistics, solely; other approaches rely on classical test theory or variants of it; still other approaches rely on item response theory.

Classical Test Theory

Extensive treatments of classical test theory are provided by Gulliksen (1950), Lord and Novick (1969), and Feldt and Brennan (1989). The theory is based on the extraordinarily simple, and seemingly self-evident, equation $X = T + E$, where X is observed score, T is true score, and E is error score. The model is deceptively simple, however, because the two terms on the right-hand side are latent, or unobservable. For the model to be useful, therefore, additional definitions and assumptions are required. In most treatments of classical test theory, the first definition stated is that T is the expected value of X over replications of the measurement procedure, which leads to E having an expected value of zero. Then it is often assumed that the covariance of T and E is zero. These are the central assumptions (not the only ones) that lead to the usual results in classical test theory. A crucial aspect of these assumptions is encapsulated in the phrase "replications of the measurement procedure."

As discussed extensively by Brennan (2001a, c), the history of classical test theory is replete with different perspectives on, and even arguments about, "What constitutes a replication of a measurement procedure?" Different answers lead to different conclusions about results such as reliability. As stated by Brennan (2001c):

> Reliability, broadly conceived, involves quantifying the consistencies and/or inconsistencies in examinee scores. It has been stated that, "A person with one watch knows what time it is; a person with two watches is never quite sure." This simple aphorism highlights how easily investigators can be deceived by having information from only one element of a larger set of interest. (p. 7)

In short, the notion of replications is central to a conceptualization of T in classical test theory, and replications are necessary to estimate reliability. These replications may be somewhat contrived (e.g., all possible split halves), but replications in some form are necessary for estimating reliability. Focusing on replications inevitably causes users of scores to be more uncertain about their decisions (and appropriately so) than otherwise would be the case (Brennan, 1998a).

Traditionally, in classical test theory two types of statistics predominate: reliability coefficients (r) and standard errors of measurement (*SEMs*). For a typical examinee these statistics are related by the formula $SEM = S\sqrt{1-r}$, where S is the standard deviation of observed scores, and SEM is sometimes called the "overall" SEM. A reliability coefficient can be defined as the correlation between replications (actual or hypothetical) of a measurement procedure. Consequently, there are as many different values for a reliability coefficient and its corresponding overall SEM as there are definitions of what constitutes a replication. A particular examinee's *conditional SEM* is the standard deviation of the observed scores for the examinee, which characterizes the uncertainty in the examinee's observed scores. The average of conditional *SEMs* is the overall SEM.

Generalizability Theory

Generalizability theory can be viewed an extension or liberalization of classical test theory through an application of certain analysis of variance (ANOVA) procedures to measurement issues. The defining treatment of generalizability theory was provided by Cronbach, Gleser, Nanda, and Rajaratnam (1972). Brennan (2001b) provides a recent extensive treatment. Brennan (1983, 1992) provides a relatively detailed treatment. Shavelson and Webb (1991) provide an introductory monograph.

In classical test theory, E is a single undifferentiated random error term. As such, any single application of classical test theory cannot distinguish among multiple sources of error. By contrast, when Fisher (1925) introduced ANOVA, he

> revolutionized statistical thinking with the concept of the factorial experiment in which the conditions of observation are classified in several respects. Investigators who adopt Fisher's line of thought must abandon the concept of undifferentiated error. The error formerly seen as amorphous is now attributed to multiple sources, and a suitable experiment can estimate how much variation arises from each controllable source (Cronbach et al., 1972, p. 1).

In short, generalizability theory liberalizes classical theory by employing ANOVA methods that allow an investigator to disentangle multiple sources of error that contribute to the undifferentiated E in classical theory.

In discussing the genesis of generalizability theory, Cronbach (1991) states:

> In 1957 I obtained funds . . . to produce, with Gleser's collaboration, a kind of handbook of measurement theory. ...“Since reliability has been studied thoroughly and is now understood,” I suggested to the team, “let us devote our first few weeks to outlining that section of the handbook, to get a feel for the undertaking.” We learned humility the hard way—the enterprise never got past that topic. Not until 1972 did the book appear . . . that exhausted our findings on reliability reinterpreted as generalizability. Even then, we did not exhaust the topic.
>
> When we tried initially to summarize prominent, seemingly transparent, convincingly argued papers on test reliability, the messages conflicted. (pp. 391–392

To resolve these conflicts, Cronbach and his colleagues devised a rich conceptual framework and married it to analysis of random-effects variance components. The net effect is “a tapestry that interweaves ideas from at least two dozen authors” (Cronbach, 1991, p. 394).

Although classical test theory and ANOVA can be viewed as the parents of generalizability theory, the child is both more and less than the simple conjunction of its parents, and appreciating generalizability theory requires an understanding of more than its lineage. For example, although generalizability theory liberalizes classical test theory, not all aspects of classical theory, as explicated by Feldt and Brennan (1989), are incorporated in generalizability theory. Also, not all of ANOVA is relevant to generalizability theory; indeed, some perspectives on ANOVA are inconsistent with generalizability theory. In addition, the ANOVA issues emphasized in generalizability theory are different from those that predominate in many experimental design and ANOVA texts. In particular, generalizability theory concentrates on variance components and their estimation.

Perhaps the most important aspect and unique feature of generalizability theory is its conceptual framework. Among the concepts are *universes of admissible observations* and G (*generalizability*) studies, as well as *universes of generalization* and D (*decision*) studies. Basically, a universe of admissible observations specifies all the facets that are of potential interest to an investigator for some purpose, and a G study is conducted to estimate variance components associated with this universe. These G study variance components are for single conditions of the facets, and they can be used in turn to estimate variance components for multiple conditions of facets (e.g., the specific number of items in a test, the number of raters for an assessment, etc.) that characterize a D study and a universe of generalization—the universe to which an investigator wants to generalize based on the scores for a specific instance of a measurement procedure. Another way of viewing a universe of generalization is to say that it is the set of all replications of the measurement procedure. A single G study can provide a basis for estimating results for a number of D studies and universes of generalization that can differ with respect to which facets are fixed and which are random, the sample sizes for facets, and the structure of the D study.

For example, a G study might be conducted for a writing assessment in which the facets in the universe of generalization are essay prompts (t) and raters (r). If a sample of persons (p) is administered n_t prompts each of which is evaluated by n_r raters, the G study design can be characterized as $p \times t \times r$, and ANOVA procedures can be used to estimate the seven variance components associated with this design: $\sigma^2(p)$, $\sigma^2(t)$, $\sigma^2(r)$, $\sigma^2(pt)$, $\sigma^2(pr)$, $\sigma^2(tr)$, and $\sigma^2(ptr)$. Then, reliability-like coefficients and error variances can be estimated for various D study designs and universes of generalization, such as the following.

1. The D study design is $p \times T \times R$ with decisions based on examinee mean scores over $n'_t = 3$ prompts and $n'_r = 2$ raters, under the assumption that prompts and raters are random effects. This means that generalization is intended to a wider universe of prompts and raters.

2. Same D study design as in Example 1, but n'_t and n'_r are different from 3 and 2, respectively.

3. Same D study design as in Example 1, but it is assumed that prompts are fixed effects and raters are random effects. This means that generalization is intended to a wider universe of raters than the n'_r raters used in the D study. Stated differently, every instance of the measurement procedure in

the universe of generalization would involve the same set of n'_t prompts but a different set of n'_r raters.

4. The D study design is $p \times (R\!:\!T)$, where the colon is read "nested within." Decisions are based on examinee mean scores over $n'_t = 3$ prompts with each prompt evaluated by a different set of $n'_r = 2$ raters, under the assumption that prompts and raters are random effects.

5. Same D study design as in Example 4, but it is assumed that prompts are fixed effects and raters are random effects.

These examples illustrate the flexibility of generalizability theory.

Generalizability theory can distinguish among several types of error variances, including, most importantly, relative error variance and absolute error variance. Relative error variance, $\sigma^2(\delta)$, is appropriate when decisions about objects of measurement (usually persons) are based on rank ordering. By contrast, absolute error variance, $\sigma^2(\Delta)$, is appropriate when the errors of interest are the actual differences between observed scores and universe scores (analogous to true scores in classical test theory). Various reliability-like coefficients are typically employed in generalizability theory including generalizability coefficients that employ $\sigma^2(\delta)$ and phi coefficients that use $\sigma^2(\Delta)$. More recently, other measures of precision that are somewhat reminiscent of indices in the physical sciences have been proposed by Kane (1996).

The foregoing description of generalizability theory is more correctly a description of univariate generalizability theory in the sense that each object of measurement has only one universe score. In multivariate generalizability theory, each examinee has multiple universe scores, each of which corresponds to a level of a fixed facet. In fact, a univariate generalizability analysis with a fixed facet is really a simplified version of a more informative multivariate analysis. A commonly occurring multivariate example is a test organized according to a table of specifications (see Brennan, 2001b, sect. 9.1, for an overview). In this example, each category in the table constitutes a level of a fixed facet, and within each category there is a simple persons-crossed-with-items ($p \times i$) design. The actual scores used to make decisions are a weighted composite of the category scores. Often, the weights are proportional to the numbers of items in the various categories, but the theory makes no such restriction. Multivariate generalizability theory is extraordinarily flexible because it can faithfully model so many different types of measurement procedures.

Item Response Theory

In item response theory, examinee responses are modeled at the item level, whereas, for the most part, classical test theory and generalizability theory focus on test scores over items. There are numerous item response theory models; some for dichotomously-scored items and others for polytomous items. These models express the probability of an examinee's response to an item as a function of an underlying latent or proficiency variable usually denoted θ and defined over the range $-\infty < \theta < \infty$. Lord (1980) provides an authoritative treatment of item response theory. A particularly readable treatment is provided by Hambleton, Swaminathan, and Rogers (1991).

For dichotomously-scored items, the three-parameter logistic (3PL) model is very frequently discussed. For this model, it is assumed that the probability (P) that an examinee with ability θ will get item j correct is

$$P_j(\theta) \equiv P(x_j = 1|\theta; a_j, b_j, c_j) = c_j + (1 - c_j)/\left\{1 + \exp\left[-Da_j(\theta - b_j)\right]\right\}, \quad (1)$$

where x_j is the response to the item, b_j is the item “difficulty” parameter, a_j is the item “discrimination” parameter, c_j is the lower asymptote or “pseudo guessing” parameter, and $D = 1.7$. Equation 1 is sometimes called an item characteristic curve, ICC, or item response function. The b_j parameter is the value of θ at which $P_j(\theta) = .5(1 - c_j)$. The a_j parameter is proportional to the slope at the point $\theta = b_j$, which is the inflection point of the ICC. Typically, the c_j parameter is close to the probability of getting the item correct by random guessing.

There are two frequently cited special cases of Equation 1. Setting $c_j = 0$ gives the two parameter logistic model (2PL). Setting $c_j = 0$, $a_j = 1$, and $D = 1$ gives the one parameter logistic model (1PL), or the Rasch model.[1] Also, each of these logistic models (in particular, the 2PL model) is sometimes viewed as an approximation to a corresponding normal ogive model (see Lord & Novick, 1968, p. 399).

In addition to Equation 1, there are two crucial assumptions in item response theory: unidimensionality and local independence. The assumption of unidimensionality means that examinee ability or proficiency can be described completely by a single latent variable, denoted θ here. The assumption of local (or conditional) independence means that, for any examinee (or, equivalently, any population of examinees with the same θ), examinee responses to items are statistically independent. The local independence assumption means that there are no dependencies among the items other than those that are attributable to θ. Strictly speaking, this characterization of item response theory is a description of *unidimensional* item response theory. Although multidimensional models are less frequently discussed, they are sometimes used in simulations, and they are occasionally used in the National Assessment of Educational Progress (NAEP).

The item response theory models discussed above are for dichotomous data. There are a number of models that have been proposed for polytomous data. These models use more complicated expressions than Equation 1, but the assumptions of unidimensionality and local independence are still required.

For a test with n items, a test characteristic curve (TCC) in the total-score

[1] Proponents of the Rasch model sometimes quarrel with this characterization of the model by noting that the Rasch model can be developed from different principles that do not require it to be viewed as a simplifying case of Equation 1 (see, for example, Wright & Stone, 1979, chap 1).

metric is simply the sum of the n ICCs:

$$\tau_{irt}|\theta = \sum_{j=1}^{n} P_j(\theta), \tag{2}$$

where τ_{irt} is the true score in item response theory, which is a non-linear transformation of θ. Equation 2 provides the expected observed score (in the total score metric) for an examinee with ability θ.

In item response theory the concept of "information" is used to describe items and tests, and to obtain certain standard errors. For dichotomously-scored items, the information provided by item j is given by the item information function:

$$I_j(\theta) = \frac{[P_j'(\theta)]^2}{P_j(\theta)[1 - P_j(\theta)]}, \tag{3}$$

where $P_j'(\theta)$ is the first derivative of $P_j(\theta)$ with respect to θ. The test information function is simply the sum of the item information functions:

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta). \tag{4}$$

Given the form of Equation 4, it is clear that items contribute independently to the test information function, which is not true in classical test theory. For example, in classical test theory, item discrimination indices are dependent on the characteristics of the other items in the test.

The precision with which abilities are estimated (under maximum likelihood) is related to the test information function in the following manner

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}, \tag{5}$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ. For any given value of θ, Equation 5 provides what is called the conditional standard error of estimation, or conditional *SEE*. This serves a role similar to that of the conditional *SEM* in classical test theory.

θ is usually scaled to have a mean of 0 and a standard deviation of 1. Any monotonic transformation of θ could serve equally well (see Lord, 1980, pp. 84–88). However, different non-linear transformations of θ lead to different test information functions and, hence, different *SEE* functions.

Item response theory involves certain indeterminacies. Referring to Equation 1, if we replace θ by $\theta^* = \alpha\theta + \beta$, b_j by $b_j^* = \alpha b_j + \beta$, and a_j by a_j/α, then $P_j(\theta)$ is unchanged. This means that the origin and unit for measuring ability are purely arbitrary. So, for example, if we determine the b_j for a set of items based on one group of examinees and then independently for a second group, we would not expect the two sets of b_j to be identical. Rather, they would be related by a linear transformation—the same linear transformation that relates θ for the two groups. That is the basis for item response theory equating of two forms of the same test.

Equating/Linking

Equating is a process for transforming scores on one form of a test to the scale of another form, where both forms are constructed according to the same content and statistical specifications. Methodology for equating is discussed extensively by Kolen and Brennan (2004). When the forms differ in their content and/or statistical specifications, the transformation is referred to as a linking. In addition, the word “linking” is used for a relationship between scores on *different* tests. Linking has been discussed by Mislevy (1992), Linn (1993), Feuer (1999), and Kolen and Brennan (2004), among others. There are numerous types of linking, a variety of perspectives on the matter, and a large number of relevant publications. The common denominator between equating and linking is that the end result is a transformation that relates scores on two or more forms or tests, but the strength of the relationship is very much dependent on the similarity of the forms or tests and the conditions under which they are administered.

Equating/linking is an important issue for practically every testing program. For example, K-12 testing programs use equating procedures to transform scores on new forms of their tests to some original scale. Linking procedures are used to relate scores for tests developed for different grades. For many states, the NCLB Act will necessitate the development of many more tests than in the past (e.g., every-student testing in math, reading, and eventually science in grades 3-8), which will lead to a substantial increase in the amount and difficulty of equating/linking.

The methodologies used in equating and linking are an eclectic set. Some procedures (e.g., observed-score equipercentile equating) don’t even formally recognize the existence of true scores, which are central to classical test theory, generalizability theory, and item response theory. Other procedures (e.g., item response theory true-score equating) provide transformations in terms of true scores, which are unusable in practice unless one substitutes observed scores for true scores. That is precisely what is usually done, but there is no theoretical justification for doing so. Given these types of discontinuities, it is perhaps remarkable how well equating usually seems to work, but it is also disconcerting that there is often only a tenuous relationship between equating/linking procedures and various elements of the previously discussed models.

Inconsistencies Across Models

The models that have been discussed are quite well developed and widely used, but, in many respects, they are not well integrated. For example, it is not at all unusual for an assumption used in, say, equating test forms to be blind to an assumption used to document the reliability of scores for those forms.

In considering relationships among models, it is important to pay attention to both mathematical and conceptual issues. Mathematics can address syntactical similarities, but semantics and inference require attention to conceptual issues. Of the models discussed above, the two that are probably the most closely integrated are classical test theory and generalizability theory. Both conceptually and mathematically, it can be argued that classical test theory is a special case of generalizability theory. It can also be argued that classical test theory is a special case of item response theory, but such arguments are sometimes more mathematical than conceptual, even though some of the words used in both theories (e.g., difficulty and discrimination) are the same.

Perhaps the most obvious inconsistency among these models is that item response theory pays particular attention to items, whereas the other models are largely test-score based. However, this obvious difference is not nearly as important as differences in what constitutes a replication, differences in definitions of true scores, differences in the conceptualization and estimation of error, differences in how tests are scored (i.e., how true scores or values of a latent variable are estimated, and different perspectives about how scores from various tests should be combined. These five issues are discussed next.

What Constitutes a Replication?

Reliability (either a coefficient or an *SEM*) involves quantifying the consistencies and/or inconsistencies in examinee scores over replications. It follows that grasping the concept of reliability and its estimates involves grappling with the question: “What constitutes a replication of a measurement procedure?” (See Brennan, 2001a.)

Generalizability theory is especially well-suited for providing a detailed specification of replications. However, we do not need to invoke the full conceptual framework of generalizability theory to capture one very important distinction—namely, the notion of replications is operationalized in part by specifying which sets of conditions of measurement (items, occasions, tasks, raters, etc.) are *fixed* for all replications and which are *random* (i.e., variable) over replications. In generalizability theory, a set of conditions of measurement is called a facet. So, the notion of replications involves specifying which facets are fixed and which are random.

Careful thought about replications requires that an investigator have clear answers to two questions:

1. What are the intended (possibly idealized) replications of the measurement procedure?

2. What are the characteristics of the data actually available, or to be collected, to estimate reliability?

It is particularly important to note that if a facet is intended to be random (Question 1), but it is effectively fixed—e.g., only one instance—in a particular set of data (Question 2), then any reliability coefficient computed using the data will likely overstate matters, and error variance will be understated.

Many apparent conflicts in reliability results can be explained by careful consideration of these matters. For example, conventional wisdom holds that group means are more reliable than scores for individuals. Brennan (1995) has shown that this conclusion is not necessarily true, and he has explained why it can be false in terms of which facets are fixed and which are random for various group-mean reliability coefficients. Also, Brennan (2001b, pp. 127–129) has shown how differences in the magnitudes of traditional coefficients of reliability are explainable in terms of fixed versus random facets.

Historically, in item response theory, the terms “fixed” and “random” have not been widely used. However, these notions play a role in the theory. Specifically, in typical treatments of item response theory, the n items are *fixed*, or, more correctly, the parameters of the items are fixed. That is, a replication would involve a set of n items with *identically* the same item parameters. This notion of a replication is much more restrictive than that of classically parallel tests, and dramatically more restrictive than that of randomly parallel tests in generalizability theory. In effect, there is no assumed sampling of content in the usual item response theory models, whereas other models permit content sampling. One clear implication is that, all other things being equal, *SEMs* (or, more correctly, *SEEs*) in item response theory will be smaller than *SEMs* in classical test theory or generalizability theory solely because of model assumptions (see, Lee, Brennan, & Kolen, 2000, pp. 14–16).

What are True Scores?

Theories of measurement make repeated reference to true scores (or scores on some latent trait). As noted above, the most obvious example is the classical test theory model, but generalizability theory and item response theory have their own versions of these concepts. Since true scores are unobservable, they must be defined for these theories to have any utility. The manner in which these entities are defined can make a very big difference

For example, as noted by Brennan (2001c),

> Lord and Novick (1968) go to considerable length distinguishing between the “expected-value” and “platonic” notions of true scores. Basically, the “expected-value” notion defines true score as the expected value of observed scores, whereas the “platonic” notion refers to a known-to-be-true attribute of an object. As Lord and Novick demonstrate, the two notions are not the same, and they can lead to considerably different results. Classical theory and generalizability theory employ the expected-value notion, but many public statements about true scores have a much more platonic flavor—i.e, true score is frequently discussed as if it were Truth. (p. 7)

Classical test theory and generalizability theory employ the expected-value notion of true score. By contrast, when item response theory is used with dichotomously-scored items, some of the arguments among proponents of the 1PL and 2PL models vis-a-vis the 3PL model are essentially arguments about what shall be considered true score. The 3PL model with its lower asymptote is reasonably consistent with defining true score as an expected value, because it acknowledges that a low-ability examinee has a positive probability of a correct response. By contrast, the 1PL and 2PL models require that low ability examinees have a probability-of-correct response approaching zero. It appears that these latter models are based on defining true score in the more platonic sense of "knowing" the answer to an item, as opposed to getting it correct.[2]

Recall, as well, that in traditional treatments of item response theory, the n items in an analysis are fixed, which means that true scores given by the test characteristic curve in Equation 2 are for the fixed set of items. By contrast in classical test theory, true score is defined as the expected value of observed scores over forms that are "similar" in some sense, and in generalizability theory, true score (called "universe score") is the expected value over randomly parallel forms. These differences, which are often unacknowledged, have important theoretical and practical implications, especially for item response theory vis-a-vis the other two theories.

What is Error?

In his introduction to *Generalizability Theory* Brennan (2001) states:

> The pursuit of scientific endeavors necessitates careful attention to measurement procedures, the purpose of which is to acquire information about certain attributes or characteristics of objects. However, the information obtained from any measurement procedure is fallible to some degree. This is evident even for a seemingly uncontroversial measurement procedure such as one used to associate a numerical value (measurement) with the length of an object. Clearly, the measurements obtained may vary depending on numerous conditions of measurement, such as the ruler used, the person who records the measurement, lighting conditions, and the like.
>
> Although all measurements are fallible to some extent, scientists seek ways to increase the precision of measurement. To do so, they frequently average measurements over some subset of predefined conditions of measurement. This average measurement serves as an estimate of the "ideal" measurement that would be obtained (hypothetically) by averaging over all predefined conditions of measurement. A substantive question then becomes, "How many instances of which conditions of measurement are needed for acceptably precise measurement?" For example, if prior research has demonstrated that the choice of ruler has little influence on measurements of the length of certain objects, but considerable variability is associated with the persons who record measurements, then it is sensible to average measurements over many persons but few rulers.
>
> Another way that scientists sometimes increase the precision of measurement is to fix one or more conditions of measurement. For example, a specific ruler might be used to obtain all measurements of

[2] Of course, these models might be adopted merely as approximations to the 3PL model or some other model.

> the length of an object. However, the choice of a specific ruler for all measurements involves a restriction on the set of measurement conditions to which generalization is intended. In other words, fixing a condition of measurement reduces error and increases the precision of measurements, but it does so at the expense of narrowing interpretations of measurements. (p. 1)

Differences in the underlying notions of true score are a primary contributing factor to different conceptions and estimates of error. In classical theory and generalizability theory, error does not mean mistake, and it does not mean "model misfit" in the usual sense of that term. Rather, important aspects of error are defined directly or indirectly by the investigator. This is eminently obvious in generalizability theory which requires that the investigator explicitly define both true score (i.e., universe score) and the type of error under consideration. In classical theory, the investigator effectively defines error through specifying a data collection design. That is why traditional coefficients of internal consistency, stability, and stability and equivalence typically lead to different estimates of error variance. That is, the error variances associated with these coefficients are *not* different estimates of the same quantity; rather, they are estimates of *different* quantities. It is often unrecognized, but nonetheless true, that the investigator through various overt or hidden choices is actively involved in deciding what shall be considered as error.

There is no E term per se in item response theory, but there are different notions of error that are often discussed in the model. For example, the extent to which the model does not fit the data is a type of error. Rarely, however, is model misfit reported as a quantified amount; rather, various methodologies are employed to assess model (mis)fit. Also, the conditional *SEEs* given by Equation 5 are usually used in much the same way that conditional *SEMs* is used in classical test theory or generalizability theory. Very often, however, the two statistics are not comparable for any one of three reasons.

First, given the definition of the test information function in Equation 4, the *SEE* in Equation 5 applies only to maximum likelihood estimates (*MLEs*) of θ. There are other item-response-theory estimators of θ; and, in many practical circumstances, number-correct scores (or transformations of them) are used to make decisions, even if item response theory is used for other purposes in the testing program. Under such circumstances, Equations 4 and 5 need to be modified.

Second, neglecting the difference between an *SEE* and an *SEM*, it is often stated that, under item response theory assumptions, conditional *SEMs* are *larger* at the extremes of the score scale than in the middle. By contrast, in classical test theory and generalizability theory, almost always conditional *SEMs* are *smaller* at the extremes than in the middle. This is a dramatic difference that is sometimes used as an argument against the credibility of conditional *SEMs* in classical test theory and generalizability theory. However, this difference is almost always an artifact arising from the choice of the θ scale (see Brennan, 1998b, pp. 326–328), and, as noted above, there is no theoretical virtue to any particular choice of scale.

Third, even if we view the conditional *SEEs* given by Equation 5 as conditional *SEMs*, they do not distinguish among multiple sources of error. In fact, given the unidimensionality assumption of item response theory, there is no obvious role for multiple sources of error.

Recognizing this problem, Bock, Brennan, and Muraki (2002) have suggested an ad hoc approach for incorporating multiple sources of error in an item response theory analysis for a test consisting of items scored by multiple raters. This is a simple matter in generalizability theory, but as Bock et al. (2002) note:

> Regrettably, a similar straightforward approach to estimation of proficiency from multiple ratings does not exist in present item response theory (IRT). An essential assumption of IRT is that the scores in the examinee's response . . . are conditionally independent, given the examinee's level of proficiency. This is not true of multiple ratings of a response to a given item: They provide additional information . . . only to the extent that they attenuate rater error.
>
> Although treating multiple ratings as if they were separate items in an IRT analysis would not in general bias estimation of examinee proficiency, the standard error of estimate would be biased downward. (p. 365)

The ad hoc solution proposed by Bock et al. (2002) involves a modification of information functions based on results from a generalizability analysis. The net effect is to adjust the *SEEs* so that they incorporate error attributable to both items and raters; that is, in a sense, the procedure induces more random error into the model.

There are at least two other classes of approaches that have been proposed for introducing more randomness along the items dimension into item response theory models. Both approaches attempt to relax the “fixed items” assumption of traditional item response theory analyses. The first approach, discussed by Kolen and Harris (1987), uses both multivariate generalizability theory and item response theory to model tests developed according to a table of specifications. In effect, in their approach the items in a particular test form are viewed as a sample from a stratified universe of items. In the second approach, prior distributions are placed on item parameters (see, for example, Glas & van der Linden, 2003). Then, sampled values of the item parameters are viewed as realizations of a random vector. This approach is being actively discussed for computerized adaptive testing in the context of using item forms or templates. There are also other Bayesian approaches that effectively introduce randomness into item response theory models, although their primary purpose is usually to improve parameter estimation, especially the estimation of ability (see, for example, Mislevy 1993).

To this point, the notion of error has been tied to the extent to which an examinee's observed or estimated score is, in some sense, a “good” estimate of his/her personal parameter (true score, universe score, latent ability or proficiency, etc.). Loosely speaking, this involves generalization from a sample of behavior to some "universe" of behavior. By contrast, traditional treatments of statistics emphasize generalizing from a sample of persons to a population of persons, with the persons' scores typically treated as fixed. So, errors are with respect to sampling of persons. This is also the focus of most statistics used to quantify errors in equating/linking (see Kolen & Brennan, 2004). Indeed, this discontinuity between the treatments of error in many measurement models and the treatments in the equating/linking literature is remarkable. It also seems problematic, because important decisions about examinees are frequently based on equated or linked scores, which are certainly fallible to some extent, and the degree of fallibility must depend in some sense on both the sampling of persons and the sampling of behavior.

On balance, then, different measurement theories are quite different with respect to their conceptions of error, how to quantify it, and how to explain it. Since error is so fundamental in measurement, such inconsistencies among the theories cast considerable doubt on their interchangeability at the current time. For the most part, the different theories do not provide alternatives to answering the same questions about error; rather, they more frequently provide answers to *different* questions about error.

How Should Tests be Scored?

Deciding how to score a test can be considerably more complicated than it may appear, and the actual decision made is sometimes based on the psychometric model that is adopted. Traditionally, for a test consisting of multiple-choice items, the score that is used is simply the number of items the examinee got correct, which is usually called the examinee's raw score. However, there are some testing programs (e.g., the *SAT*) that use a so-called "formula score" that adjusts the raw score for the possibility of getting items correct by guessing. There is a long-standing debate about the practical and psychometric benefits of such corrections for guessing, but that debate pales by comparison with arguments concerning the differential weighting of items that is often discussed in item response theory.

For example, for the 2PL model, a so-called "sufficient statistic" for θ is the sum of the a_j terms for the items that an examinee gets correct.[3] Clearly, then, if the sufficient statistic is used to estimate θ, different items contribute differentially to the examinee's score. Matters are even more complicated for the 3PL model, because a sufficient statistic does not even exist, which leads psychometricians to use other methods for estimating θ. These methods are well-described in a recent book edited by Thissen and Wainer (2001), and will not be considered further here, except to note the obvious—namely, procedures for obtaining examinee scores in item response theory can be very complicated

[3] A sufficient statistic contains all the information in the data for estimating some unknown parameter, here θ.

and counter-intuitive for lay persons. In particular, it is quite possible for examinees with lower/higher raw scores to get higher/lower estimates of θ. These problems are a principal reason that number-right or "summed" scores are being actively considered by many researchers in item response theory (see, again, Thissen & Wainer, 2001).

When items are not scored dichotomously, other complexities arise. The quintessential example is an essay prompt for which a small number of scores are defined according to a rubric. Almost always these scores are consecutive integers (e.g., 0, 1, 2, 3, 4, 5, 6). Arguments often arise, even among content-matter experts, about the appropriateness of the rubric's statements about particular scores. Also, complicated questions can arise about the relative values of the score points. For example, is an essay scored 4 really twice as good as an essay scored 2? In a sense, the answer is "yes" for traditional scoring procedures, but not necessarily for scoring procedures that might be used in item response theory.

How Should Scores from Various Tests be Combined?

In many environments, decisions are based on examinee performance on multiple tests or tests divided into multiple parts (e.g., testlets). Indeed, this is a virtual requirement under the NCLB legislation. There is substantial debate among both lay persons and measurement specialists about how scores from multiple tests should be combined. Sometimes the debate focuses on the perceived or judged relative "importance" of the assessments. For example, suppose an expert or agency argued that scores on an essay test are twice as important as scores on a multiple-choice test of writing skills. What does that mean? Let us suppose that there is one essay scored on a scale of 0-6, and there are 27 multiple-choice questions. One answer to the "twice as important" question is obtained by multiplying essay scores by 9. That is, letting X be essay score and Y be multiple-choice score, the composite score is defined as

$$C = (9)X + (1)Y = 9X + Y$$

Then, the maximum multiple-choice score will be 27, the maximum essay score will be twice as large (namely, 54), and the maximum composite score will be 81. Alternatively, if $\overline{X}$ and $\overline{Y}$ are the corresponding proportion correct scores, then the composite mean score is

$$\overline{C} = .9X + .1Y,$$

and the weights are .9 for the essay and .1 for the multiple-choice test, which sum to 1 (the usual convention).

These types of weights are sometimes called "a priori" or "nominal" weights. They may be quite justifiable and have considerable "face validity," but they may not produce results that are otherwise acceptable. For example, scores for a single essay prompt are not likely to be very reliable, while scores for a 27-item multiple-choice test might well have moderate reliability. If so, the reliability of the composite scores (C or $\overline{C}$) will be dominated by the reliability of the essay scores, which is the less reliable component of the composite. Hence, composite scores will have relatively low reliability.

Furthermore, if we focus on the variability of the composite scores, it is unlikely that the nominal weights or .9 (for the essay) and .1 (for the multiple-choice test) will be reflected in the relative contribution of the essay and multiple-choice scores to the variability in the composite scores. These relative contributions are usually called "effective" weights, as distinct from the a priori or nominal weights (see, for example, Brennan, 2001b, pp. 305–307, who discusses these matters in the context of generalizability theory). In other words, it is entirely possible that the variability among examinees' composite scores will be influenced more by the multiple-choice scores than the essay scores.

The foregoing description of nominal and effective weights has been couched in terms of classical test theory or generalizability theory. Similar issues arise in item response theory, but the methodologies for addressing these issues are quite different and may lead to substantially different results.

Furthermore, composite scores present substantial challenges for equating. The essential problem is this: should composite scale scores be obtained through equating the composite directly or through a weighed sum of the equated scores for the component parts. In either case, issues of weighting arise, and the two procedures are not likely to give the same results.

Concluding Comments

This paper has considered some inconsistencies or ambiguities among four psychometric models (classical test theory, generalizability theory, item response theory, and equating/linking models). These inconsistencies or ambiguities have been illustrated in the context of five questions: what constitutes a replication; what are true scores; what is error; how should tests be scored; and how should scores from various tests be combined? These are not an all-inclusive set of relevant questions, and the discussions provided are not particularly extensive. However, these questions are fundamental to measurement itself, and the discussion of them illustrates some important differences among the models in how they approach measurement issues.

Given the discontinuities and ambiguities across measurement models, it is natural to ask which model provides the correct or right answer to the questions posed. For the most part, there is no right answer, and investigators searching for that "Holy Grail" will be forever disappointed. The models are just that—models, not reality; each of them has its own set of assumptions, and the assumptions do not mesh perfectly across models. Practitioners sometimes do not realize that model assumptions are not always chosen because they are thought to reflect reality. Rather, assumptions are often chosen because they seem to be natural ways, in the context of a particular model, to solve or at least simplify otherwise intractable estimation issues.

Arguments over assumptions often arise in discussions about various models, particularly discussions of classical test theory vis-a-vis item response theory.

Classical test theory uses a very small number of simple definitions and relatively weak assumptions. “Weak” in this context does not mean wrong; it merely means not strong and/or not demanding assumptions about distributional form. By contrast, item response theory makes much stronger assumptions. Strong assumptions generally lead to “strong” results, but of course the strength of the results depends on the credibility of the assumptions.[4]

The weak assumptions in classical test theory permit the derivation of a remarkably large number of very useful results. Still, classical test theory cannot be used to draw all of the inferences that some decision makers want to make. By contrast, many of the most intractable problems in measurement become trivial if the assumptions of item response theory hold. However, the item response theory assumptions are so strong that they are likely false in almost all situations. Consequently, it is important to consider the extent to which item response theory results are robust with respect to assumption violations, and/or the extent to which a particular application of item response theory challenges its assumptions. For example, in typical equating contexts with intact test forms, item response theory assumptions are not severely challenged, and item response theory has been shown to work very well for equating in numerous testing programs. (These are contexts in which classical procedures and item response theory procedures tend to give very similar results.) In vertical scaling, however, the item response theory assumptions are severely challenged, and we are usually much less convinced that the model is working as well as we would like. (Different procedures for vertical scaling tend to give different results, as illustrated by Kolen & Brennan, 2004, chap. 9.)

It is the contention of this author that the major measurement models are well-developed and generally internally consistent, but the field of measurement as a whole is not nearly as well integrated as one would ideally like. Too frequently, apparent similarity across models in terminology, notation, and/or concepts masks differences that have important theoretical import and/or real practical consequences. There is still much work to be done.

[4] Recently, Holland and Hoskens, 2003, have provided an ingenious integration of aspects of classical test theory and item response theory, but their work does not resolve all discontinuities between the two theories.

References

Bechger, T., Béguin, A., Maris, G., & Verstralen, H. (March, 2003). *Combining classical test theory and item response theory.* Measurement and Research Department Reports No. 2003-4. Arnhem, Netherlands: CITO, National Institute for Educational Measurement.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*, 364–375.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: ACT, Inc.

Brennan, R. L. (1992). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: ACT, Inc.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 14*, 385–396.

Brennan, R. L. (1998a). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice*, *17*(1), 5–9, 30.

Brennan, R. L. (1998b). 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.* Springer-Verlag.

Brennan, R. L. (2001c). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*(4), 6-18.

Brennan, R. L. (in preparation). *Educational measurement* (2nd ed.). Greenwood Press.

Cronbach, L. J. (1991). Methodological studies—A personal retrospective. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 385–400). Hillsdale, NJ: Erlbaum.

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: Wiley.

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

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.) (1999). *Uncommon measures.* Washington, DC: National Academy of Sciences.

Fisher, R. A. (1925). *Statistical methods for research workers.* London: Oliver & Bond.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*, 247–261.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley. [Reprinted by Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.]

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68*, 123–149.

Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education*, *9*, 355–379.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* (2nd ed.). New York: Springer-Verlag.

Kolen, M. J. & Harris, D. J. (April, 1987). *A multivariate test theory model based on item response theory and generalizability theory.* A paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*, 1–20.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83–102.

Lord, F. M. (1980). *Applications of item response theory to practical testing* *problems.* Hillsdale, NJ: Erlbaum.

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

Mislevy, R. J. (1993). Some formulas for use with Bayesian ability estimates. *Educational and Psychological Measurement, 52*, 315–328.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002)

Nugent, W. R., & Hankins, J. A. (1992). A comparison of classical, item response, and generalizability theories of measurement. In D. F. Gillespie and C. Glisson (Ed.), *Quantitative methods in social work: State of the art* (pp. 11–39). Binghamton, NY: Haworth Press.

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

Thissen, D., & Wainer, H. (2001). *Test Scoring.* Mahwah, NJ: Erlbaum.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press.