Psychological Methods
2008, Vol. 13, No. 4, 376–390

1082-989X/08/$12.00 DOI: 10.1037/a0014269

A Proposed Framework for Conducting Data-Based Test Analysis

Kathleen L. Slaney and Michael D. Maraun
Simon Fraser University

The authors argue that the current state of applied data-based test analytic practice is unstructured and unmethodical due in large part to the fact that there is no clearly specified, widely accepted test analytic framework for judging the performances of particular tests in particular contexts. Drawing from the extant test theory literature, they propose a rationale that may be used in data-based test analysis. The components of the proposed test analytic framework are outlined in detail, as are examples of the framework as applied to commonly encountered test evaluative scenarios. A number of potential extensions of the framework are discussed.

Keywords: data-based test analytic framework, test analysis, assessment of psychometric properties of tests, internal test validity, external test validity

Testing and measurement in psychological science has a long and rich history, the roots of which can be traced back to the very origins of the discipline. Anyone engaged in empirical research in which quantitative measures are employed will need to become acquainted, at least to some extent, with measurement issues. However, the current state of applied (i.e., data-based) test analytic practice more than hints at the fact that there does not exist among researchers a clearly defined and established set of conventions as to what, exactly, a test analysis is to consist or on how the test analyst is to proceed. For instance, on what grounds is one justified in compositing across a set of test items to produce a test score? At what point in a test analysis does the evaluation of measurement precision occur? Or of validity? And what is the relationship between the two, and how does it bear, if at all, on the coherency of particular test analytic practices? On what grounds does one justify the choice of a particular statistical model in a given test analysis? Answers to these and similar questions are not, as it turns out, as easy to find as one might think, and, hence, it is no wonder that the applied test analytic literature consists in the application of a host of different procedures and techniques with seemingly little common rationale on which their use is based.[1]

Since the publication of the recommendations of the APA Committee on Test Standards (American Psychological Association [APA], 1954) and Cronbach and Meehl's (1955) subsequent paper, a great deal of validity theory has been generated and a number of sophisticated and complex validation frameworks have been produced. Cronbach himself elaborated substantially on many of the ideas put forth in the 1955 paper. He was one of the first to emphasize that validation is a function of the particular uses to which tests are put (e.g., pragmatic, operationist, scientific; Cronbach, 1988) and that validation of test scores calls for the integration of many different types of evidence across both context and time (Cronbach, 1971).

Others have written extensively on the issue of test validity. Elaborating and extending many of the ideas expounded by Cronbach, Messick (1980, 1988, 1989, 1995, 1998) argued that test validation is an overall evaluative judgment about the adequacy and appropriateness of particular inferences from test scores that should be based on both the existing evidence for and potential consequences of such

Kathleen L. Slaney and Michael D. Maraun, Department of Psychology, Simon Fraser University, Burnaby, British Columbia, Canada.

Portions of this article appear in Kathleen L. Slaney's unpublished Ph.D. thesis. This research was partially supported by a Social Sciences and Humanities Research Council Doctoral Fellowship held by Kathleen L. Slaney during her doctoral studies.

Correspondence concerning this article should be addressed to Kathleen L. Slaney, Department of Psychology, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada. E-mail: klslaney@sfu.ca

[1] A detailed empirical examination of current test analytic practices will not be given here. The interested reader is referred to Hogan and Agnello (2004); Meier and Davis (1990); Vacha-Haase, Ness, Nilsson, and Reetz (1999); and Whittington (1998) regarding issues pertaining to the legitimacy of particular reporting practices, to Green, Lissitz, & Mulaik (1977) and Hattie (1984, 1985) as regards practices having to do with the assessment of unidimensionality of tests and items, and to Blinkhorn (1997) for a discussion of certain of the problems that might accompany the employment of complex statistical models in the evaluation of test data.

inferences. He also emphasized that validation will often, and should, involve multiple sources of evidence, including but not limited to considerations of content, interrelationships both among items responses and between test scores and external variables, processes underlying item responses and task performance, experimental manipulations of test performance, and the value implications and social consequences of using and interpreting test scores in particular ways (Messick, 1989).

Kane (1992, 2006) has proposed an "argument-based approach" to validity, which is similar to what Cronbach (1989) described as the strong program of construct validation. In the argument-based approach, interpretative argument is adopted as a framework for collecting and presenting validity evidence. Kane has identified several major types of inferences that commonly appear in interpretative arguments, each associated with particular assumptions and each supported by different types of evidence. Other validation frameworks have been proposed by, among others, Shepard (1993); Mislevy, Steinberg, and Almond (2003); and Borsboom, Mellenbergh, and van Heerden (2004).

Cronbach's influence, and that of those who have extended and refined his ideas, can be seen clearly in the validity guidelines specified in the most recent version of *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], APA, & National Council on Measurement Education, 1999). For instance, the creators of the document assumed not only scientific contexts of test development and use but pragmatic or otherwise utilitarian uses. In addition, closely mirroring Messick (1989), *Standards* provides a summary of different sources of evidence (i.e., those based, respectively, on content, response processes, internal structure, relations to other variables, and the consequences of testing) that emphasize different lines of validity evidence as opposed to different types of validity. Furthermore, validation is characterized as an ongoing program of research in which sound arguments are made for particular interpretations or uses of test scores.

Although *Standards*, and the body of validity literature on which it is largely based, certainly does provide a description of the broad array of issues relevant to testing, the guidelines it provides pertain more to the broader context of test validation and less to the singular, data-based analyses of the psychometric properties of tests used for some broader research aim. Moreover, aside from emphasizing that a sound validity argument requires integration of these various components of test validation, *Standards* provides few details about are how they are related to one another and about how the researcher who wants to demonstrate the psychometric soundness of the measures he or she employs should proceed in order to provide evidence to support particular measurement claims. In fact, very little of the validation literature in general has been dedicated to developing explicit data-based test analytic frameworks that tie together in a coherent manner the different strands of test evaluation in order to give guidance to the applied test analyst. It is our aim in the current work to provide such a framework. Specifically, we propose a step-by-step approach to data-based assessment of test performance, a "way of doing business," if you will, wherein the "business" is evaluating the performances of tests in particular research contexts.

A Proposed Framework for Test Analysis

We will admit up front that the expression "test analysis" carries with it a fair degree of ambiguity, and, hence, the question as to what exactly a test analytic framework is to consist may well depend on whom you ask. Whereas the broader task of test validation, as is described in *Standards*, requires, at least potentially, the consideration of evidence bearing on everything from content to consequences of tests, as well as on the integration of all of the evidence that has to date been brought to bear on the validity of scores from applications of a given test, here we elaborate a framework that is appropriate primarily for data-based test analysis. In such analysis, the aim is to determine whether the responses to the items of a test that has been administered are related in ways they should be in order to justify (a) the forming of item composites (i.e., test scores) and (b) the entering of such composites into investigations concerned either with other aspects of validity (e.g., consequences of test use) or with the broader evaluation of a test over various contexts of use.

The logic underlying the framework described below is grounded in the recognition that the validity and precision of a test score bear a certain relation to one another and that a proper evaluation of a test's performance in a given instance of use should reflect this relationship. Specifically, the logic is based on the notion that the various aspects of the "external" validity of scores on a test of some particular attribute can be meaningfully explored if and only if there is evidence that the scores are adequately precise measures of the common attribute for which the test items have been designed as indicators.[2] But this requires that justification exists for compositing responses to the items of the test into a single composite (i.e., that the item responses measure in common a single primary attribute). We contend, thus, that establishing what we herein call *external test validity* is dependent on first providing evidence that supports the validity of the internal relations among item responses, or

[2] In the interests of generality, we use the term *attribute* to refer to whatever property, process, characteristic, theoretical construct, and so on, a given test is presumed to measure.

the *internal test validity*,[3] and that a test score formed as result of such evidence shows an acceptable degree of measurement precision (i.e., as determined by the conventions set forth in a given research domain).

Thus, the framework proposed here is not intended to address all of the issues relevant to test validity. Rather, it speaks to a relatively specific arena of test evaluation, namely, that pertaining to the appropriateness of choosing particular psychometric models and scoring rules in order to form suitably "reliable" and "valid" test scores that may the be meaningfully used in further analyses. As such, it specifies guidelines for evaluating test performances on the basis of a subset of the different sources of validity evidence described in *Standards* and other validation frameworks, specifically, those bearing on internal test validity, precision of measurement, contextual validity,[4] and external test validity. The framework does not, conversely, provide an explicit rationale for conducting analyses of content, process, or consequences of test use. However, the absence of these aspects of validation should not be taken as an indication that we place no importance on these issues. To the contrary, our interest is in providing a rationale according to which the psychometric soundness of scores from a test whose content validity (and, perhaps, also other components of validity) has been well established.

The components of the framework for conducting data-based test analyses, presented in the order in which they should be addressed, are as follows:[5]

1. The formal structure of the test is specified.
2. An appropriate translation of the formal structure into a set of quantitative requirements for the joint distribution of item responses is made; the resulting translation is called the *quantitative characterization* of the formal structure.
3. The conformity of the joint distribution of item responses (in a focal population) to the chosen model is assessed.
4. Conditional on the conformity of the joint distribution of item responses to the model, an optimal, model-implied compositing rule is derived and employed in order to scale the respondents with respect to the attribute purportedly measured by the items of the test.
5. The reliability (or, more generally, the precision) of the resulting composite is estimated.
6. Conditional on the composite possessing an adequate degree of precision, the composite is entered into external validation studies, or, more generally, into tests of theory-generated hypotheses about the attribute for which the composite is taken to be an indicator.

A test, then, may rightly, but provisionally, be judged to be performing adequately in a focal population of respondents if it (a) conforms to its formal structure (i.e., the theory describing the relationships between the responses to items of a test and the attribute purportedly measured by those items) and the measurement precision of the resulting composite is adequate and (b) has behaved, to date, "as it should behave," or as predicted by the "nomothetic span" of the test (i.e., the theory relating responding to the test to both other tests of the same attribute and tests of different attributes; cf. Embretson, 1983).

1. Specification of the Formal Structure of the Test

According to the authors of *Standards*,

> the conceptual framework for a test may imply a single dimension of behavior, or it may posit several components that are each expected to be homogenous, but that are also distinct from each other. . . . The extent to which item interrelationships bear out the presumptions of the framework would be relevant to validity. (AERA et al., 1999, p. 13)

Here, we refer to this conceptual framework as the *formal structure of a test* and define it as a specification of how the items of a test were designed to measure a given attribute of interest, including, at least roughly, how the distribution of the attribute across a population of individuals is to be conceptualized, how the items are linked to the attribute, and whether or not they are viewed as error-laden indicators of the attribute. We call this the formal structure, both to highlight its emphasis on the form of various properties of the test and the attribute it is intended to measure and to distinguish it from the internal structure of the test, the latter of which is relevant but not equal to the formal structure of the test.

A clear specification of the formal structure of a test is the starting point for any data-based test analysis that is to be evaluative in aim. If a test is to be meaningfully judged as performing adequately (or inadequately, for that matter) in a focal population, senses must be assigned antecedently to these and similar evaluative terms, and the specification of the formal structure of a test is the first step in fixing the senses of these terms. Furthermore, a formal structure must

[3] Internal test validity—and external test validity—should be kept distinct from the more general concepts of internal validity and external validity, as defined in Campbell and Stanley (1963).

[4] By "contextual validity" we mean any and all aspects of the generalizability of score interpretations across types of persons, settings, and times (cf. Messick, 1989).

[5] Elements of the framework presented were initially initially sketched out in Maraun, Jackson, Luccock, Belfer, and Chrisjohn (1998) and in Slaney (2006).

be worked out for each test individually and for each application of a test, because, as noted by Cronbach (1989, p. 148), "a test unsuitable in one setting can serve well in another." And, so, what constitutes adequate performance for a given test administered to respondents drawn from a focal population of interest may well be different for another test, or for the same test with either a different target population or a different context of measurement in mind. Analyses for which one cannot specify the formal structure are not founded on clearly stipulated senses of the test "behaving as it should behave" and similar notions; hence, they yield evaluative claims that are at best ambiguous and at worst vacuous.

However, rarely is the formal structure of a test laid out in exact terms, and, thus, it will often need to be deduced both from the information available about particular formal properties of the test (e.g., item response formats) and from the theoretical structure of the test (i.e., the current theory regarding both the attribute for which the test was designed as a measure and the test itself). Furthermore, the components that should properly comprise the formal structure of a test are open to debate; yet, it would seem that, minimally, the formal structure of any test should specify the following five components: (a) the theoretical "distributional form" of the attribute of which the items of the test are thought to be measures[6] (i.e., how the attribute is conceived to vary across individuals in the population under study); (b) the item response format or formats; (c) the number of attributes that the items were designed to measure; (d) the (theoretical) form of the regressions of the items on the attribute; and (e) the error characteristics of the items.

The distribution of the attribute in the population will, for simplicity, be characterized as taking on one of two forms: "continuous" for attributes that can conceivably take on many (perhaps an infinity) of ordered values and "categorical" for attributes that can assume a finite number of unordered values. As regards item type, commonly employed response scale formats are continuous (Co), x-point graded (xPL), and categorical (Ca), with the most commonly encountered special case of the latter being dichotomous (Di). Although, in practice, the number of attributes measured can take on any positive integer, for expository purposes, we will restrict our discussion to tests (or subscales of a test) that are conceptualized as measuring a single attribute. Examples include the 21 items of the Beck Depression Inventory—II (BDI–II; Beck, Steer, & Brown, 1996) as measures of depression, the 34 items of the Well-Being Scale of the Multidimensional Personality Questionnaire (Tellegen, 1981) as measures of well-being, and the 48 items of the Extraversion subscale of the NEO Personality Inventory (Costa & McRae, 1985) as measures of extraversion.

The sense of "regression" as concerns the item/attribute regressions refers to how item responding is conceptualized as varying with the level of the attribute. As with the theoretical distributional form of the attribute, these regressions are nonmathematical (or, more accurately, *pre*-mathematical) because the attribute is not, technically speaking, a random variable but, rather, is an attribute (e.g., property, characteristic, process) that the items were designed to measure. Frequently encountered conceptualizations of the item/attribute regressions are monotone increasing (MI), linear increasing (LI), x-point ordered categorical (xOC), *S*-shaped (S), and inverted *U*-shaped (U). Finally, for the sake of generality, the error characteristics of the items will be allowed to assume two possible values, error free (EF) or error in variables (EIV), even though in modern test analytic practice, the latter is virtually always assumed to be the case.

We acknowledge that providing an unambiguous specification of what and how the items of a given test measure runs somewhat counter to the commonly adopted construct validation approach to test evaluation, the aim of which is to simultaneously validate both tests of and theories about particular constructs (cf. Cronbach & Meehl, 1955; Peak, 1953). However, as Kane (2006) notes, "A system in which theories are evaluated by comparing their predictions to measurements, and the measurements are validated in terms of the theory, clearly has the potential for circularity" (p. 46). We further contend that such a practice lacks the power to pronounce on the quality of a test as a measure of the attribute ("construct") in question. If the aim in a test analysis is to evaluate the performance of a test that is being used as a measure of some particular attribute, the analysis is confirmatory in nature and, thus, requires that one must be able, at least potentially, to work out how the test has been conceptualized as a measure of the attribute under study. For instance, the formal structure of a test consisting of 7-point Likert items designed to be indicators of anxiety proneness will likely differ in important ways from that of a test designed to measure mathematical ability with a set of true/false items. Without this initial specification of the formal structure, the test analyst will have no basis for choosing one measurement model over another and, hence, will be unable to judge the quality of the test's performance within the particular context of measurement at hand.

Now this is not to say that the formal structure of the test exits in some sort of Platonic realm, direct access to which is denied to the researcher, who can then merely guess at

[6] We thank a reviewer of an earlier draft of this article for pointing out the importance of adding this component of the formal structure.

what it is that the test measures. Quite to the contrary, in fact: Presumably, the researcher has chosen the particular test either because it has been designed as a measure of the phenomenon under study or because he or she has some other reason to believe that scores generated from application of the test represent the phenomena in some theoretically or practically relevant way. As such, in identifying the formal structure of the test, he or she will likely appeal to current theory regarding the attribute (e.g., anxiety) for which the test is being used as a measure. Thus, the formal structure, at least for the present purpose, is not something to discover about the test. Rather, its specification constitutes a first step toward choosing, on the basis of careful consideration, an appropriate measurement model in order that the psychometric properties of the test can be properly evaluated. And, although there may be some ambiguity or debate surrounding different components of the formal structure of a given test, if the aim is to evaluate particular uses or interpretations of test scores, one must first be able to confirm that a test is "working" (i.e., that there are good grounds for producing those test scores as representations of the attribute such scores are intended to represent).

For example, the BDI–II is a self-report instrument for measuring severity of depression consisting of 21 items, each of which lists a symptom. The severity of the symptom is rated by the test taker on a 4-point response scale ranging from 0 to 3 (Steer, Ranieri, Kumar, & Beck, 2003, p. 59). On the basis of the scoring rationale given by the authors (see Beck et al., 1996), it may be reasonably presumed that, for each item, low ratings correspond with a lesser degree of depression and higher ratings with a greater degree of depression. As with most measures of clinical phenomena, the BDI–II items are considered to be imperfect indicators of depression. Thus, a reasonable specification of the formal structure for the BDI–II is {Co,4PL,1,4OC,EIV}, that is, a test for which a set of 4-point Likert items is designed as error-laden indicators of a single attribute (depression) that varies considerably in degree over a population of individuals. For a given item, the relationship between the item and the attribute (i.e., the item/attribute "regression") may be best conceived of as four individual ordered category regressions, in which the probability of responding to a given category varies across levels of the attribute (e.g., for a severely depressed individual, the probability of endorsing a "3" is higher than the probability of endorsing a "2," and so on; see Samejima, 1969, 1996, for an approach to modeling such graded item responses).

Obviously, because the formal structure is a linguistic specification of how the items of a test were designed to measure the attribute, it does not have any specific material implications for the joint distribution of item responses (hereafter, simply referred to as "the joint distribution"), the fulfillment of which is needed in order to justify claims about the conformity of test behavior to the specified formal structure. Testable requirements of the joint distribution must be generated through the translation of the components of the formal structure into quantitative (i.e., mathematical) counterparts. In other words, the test analyst must choose, on the basis of the specified formal structure, a measurement model for assessing the internal test validity of the data at hand.

2. Derivation of the Quantitative Characterization

Whereas the formal structure describes the theoretical relationship between an unobservable attribute and the observable indicators that are the test items, the *quantitative characterization* is the deduced, testable empirical consequences of the formal structure for the joint distribution of responses to those items. It is the quantitative embodiment of the formal structure and specifies the properties that must be possessed by the joint distribution so that item responses will be correctly judged as conforming (or not) to that specified formal structure.

To construct a quantitative characterization of a given formal structure, one maps the components of the formal structure into mathematical counterparts. Because the formal structure for most tests can be characterized in terms of the formal structures for sets of mutually disjoint sets of items, each of which is meant to measure a relatively distinct attribute (or facet of a higher order attribute), most of the measurement models of interest will be defined in terms of particular formal translations of "the items measure just one thing" (i.e., will be founded on particular conceptualizations of unidimensionality). In practice, this means choosing an appropriate unidimensional measurement model.[7] When such models are employed as quantitative characterizations of formal structures, the correspondence relations are as follows:

1. The attribute for which the test items were designed to be indicators is represented by a random variate defined on a focal population of interest (e.g., a random latent variate, θ, for which test items are taken to be indictors).

2. The notion that the test items measure a single attribute is paraphrased as the claim that the item responses are unidimensional in a sense that is dependent upon the other

[7] Once again, this does not preclude translations of formal structures that invoke particular multidimensional models. However, compositing items under such scenarios is more complex and, as such, notions of precision and validity may not be as straightforward as in the unidimensional case that we address explicitly here.

components of the formal structure of the test, namely, the item/attribute regressions and the error characteristics of the items (e.g., as defined by local conditional independence in many classes of latent variable models).

3. The item/attribute regressions referred to in the formal structure are paraphrased as analogous to item/θ regressions. In a loose sense, the latter describe the nature of the relationship between the attribute under study (which is itself represented by the random variate θ) and the items of the test (e.g., the *S*-shaped item/θ regressions that are specified in item response models used to model dichotomous indicators of a continuous latent trait).

4. Most measurement models model the situation in which test items are conceptualized as being fallible indicators of an attribute. Thus, it is natural to paraphrase formal structures with the EIV component as latent variable models (in contrast, for example, to component models).

If the aim is to evaluate the psychometric soundness of a test, the test analysis must begin with a choice of a quantitative characterization in which each and every component of specified formal structure is appropriately represented; the translation of each of the components of the formal structure into quantitative counterparts collectively consists in a statistical model that implies particular empirical requirements for a joint distribution of a set of item responses. This "isomorphism" between the formal structure and quantitative characterization is essential because a model-implied empirical result will be relevant only in the case that the chosen quantitative characterization (i.e., model) appropriately represents all of the components of the formal structure at hand. Lacking this match between the specified formal structure and the quantitative characterization, a given empirical result will be irrelevant for confirming (or disconfirming) that the test is performing in the manner it was designed to perform.

However, there is, at least in theory, the possibility of constructing many sound (and many unsound) quantitative translations of a given formal structure, as there are many different measurement models from which to choose, at least one of which will describe, at least reasonably well, a joint distribution of item responses (cf. Roskam & Ellis, 1992). Hence, mere conformity of the joint distribution to some model can have no necessary implications for the judgment of the performance of the test as a measure of the attribute it was designed to measure. Put differently, if "adequate test performance" were to be equated with simply finding a model that happens to describe the joint distribution, tests would, as a matter of course, be judged as performing adequately, and, conversely, there would exist no grounds for indicting a test.

It has become commonplace for researchers conducting test analyses to fit models to test data, typically with the aim of finding the best fitting out of a set of competing models; however, in the context of test evaluation, the "best" model is not simply the model that best describes the data in a statistical (usually loss function) sense. If the model does not consist in an appropriate translation of each of the components of the formal structure into quantitative counterparts, it cannot reasonably be considered the "best" model, quite regardless of what the particular fit function indicates. For example, if the results of a linear factor analysis indicate a good model-to-data fit by some conventional criterion but the item/attribute regressions are best conceptualized as *U*-shaped (i.e., with formal structure {Co,Co,1,U,EIV}), these linear factor analytic results will have no bearing on judgments as to the adequacy of the test (cf. van Schuur and Kiers, 1994). Just as the value of a Pearson product–moment correlation coefficient will have little relevance to claims about the strength of a nonlinear relationship between two variables, the fit of a measurement model that is a poor translation of the components of the formal structure to test data will not provide an answer as to whether the test's items are related in a way that is consistent with theoretical expectations. The onus is, therefore, on the test analyst to choose as a quantitative characterization the model that is the best available match to the particular formal structure in question (i.e., the one that provides the most reasonable representation of all the components of the specified formal structure).

Table 1 shows how a number of formal structures that are likely to be encountered for tests used in the social and behavioral sciences are mapped into particular quantitative characterizations (i.e., into particular measurement models or classes thereof). For instance, for tests whose formal structure specifies that a set of continuous items measure, with error, a single common "continuous" attribute, and in which the item/attribute regressions are conceptualized as monotone increasing (i.e., {Co,Co,1,MI,EIV}), an appropriate mapping induces a class of models known as unidimensional monotone latent variable (UMLV) models (see Holland & Rosenbaum, 1986, for a description of the properties that can be used to check whether such a UMLV model describes a joint distribution).

Table 1 also lists a special subclass of UMLV models that is arguably the most commonly employed set of models in applied test analyses. This is the class of unidimensional, linear common factor (ULCF) models. The mapping relations for formal structures specifying continuous items as fallible measures of a single "continuous" attribute, with linear item/attribute regressions (i.e., {Co,Co,1,LI,EIV}), are as follows: The relevant attribute is represented by a latent variate,θ. In this case, it is a "common factor." The linear factor analytic paraphrase of the notion that the items measure just one attribute in common is the uncorrelated-

Table 1
Examples of Quantitative Characterizations of Common Formal Structures (FS)

FS	FS → quantitative characterization mapping			Model or set of models
{Co,Co,1,MI,EF}	A = Co	→	The attribute is represented by a random variate, Y[a], which is distributed continuously over population P.	Component models (e.g., linear principal-components models)
	I = Co	→	For each item, responding is represented by a random variable X_j, which is continuously distributed over P.	
	D = 1	→	$f(\underline{X}\|Y)=\prod_{j=1}^{k} f(X_j\|Y)$ (i.e., the X_j are statistically independent).	
	R = MI	→	$E(\underline{X}\|Y)=\underline{g}(Y)$, in which $\underline{g}$ is a vector of functions and $\frac{d}{dY}\underline{g}(Y) > 0$.	
	E = EF	→	The conditional covariance matrix, $C(\underline{X}\|Y)=\psi$, is diagonal and positive definite (i.e., the X_j are required to have positive error variances).	
{Co,Co,1,MI,EIV}	A = Co	→	Attribute represented by random variate θ; θ continuously distributed over P.	UMLV models (cf. Holland & Rosenbaum, 1986)
	I = Co	→	Responding for each item represented by random variable X_j; the X_j are continuously distributed over P.	
	D = 1	→	$f(\underline{X}\|\theta)=\prod_{j=1}^{k} f(X_j\|\theta)$(i.e., the X_j are locally independent).	
	R = MI	→	$E(\underline{X}\|\theta)=\underline{g}(\theta)$, in which $\underline{g}$ is a vector of functions, and $\frac{d}{d\theta}\underline{g}(\theta) > \underline{0}$.	
	E = EIV	→	Conditional covariance matrix, $C(\underline{X}\|\theta)=\psi$, diagonal and positive definite.	
{Co,Co,1,LI,EIV}	A = Co	→	Attribute represented by random variate θ; θ continuously distributed over P.	Linear factor analytic models (cf. Jöreskog, 1966, 1969; Spearman, 1904)
	I = Co	→	Responding for each item represented by random variable X_j; the X_j are continuously distributed over P.	
	D = 1	→	The conditional covariance matrix, $C(\underline{X}\|\theta)=\psi$, is a $k \times k$ diagonal matrix (i.e., the X_j are conditionally uncorrelated).	
	R = LI	→	$E(\underline{X}\|\theta)=\underline{\Lambda}\|\theta$, with the elements of $\underline{\Lambda}$ having the same sign.	
	E = EIV	→	Conditional covariance matrix, $C(\underline{X}\|\theta)=\psi$, diagonal and positive definite.	
{Co,Di,1,S,EIV}	A = Co	→	Attribute represented by random variate θ; θ continuously distributed over P.	Binary item response models (e.g., Rasch's 1-parameter logistic model [Rasch, 1960], 2-parameter IRT models [Birnbaum, 1968; Lord, 1952, 1953])
	I = Di	→	Responding for each item represented by random variable X_j; X_j Bernoulli distributed over P.	
	D = 1	→	$P(\underline{X}=\underline{x}\|\theta)=\prod_{j=1}^{k} P(X_j=x_j\|\theta)$, $j=1..k$ (i.e., the X_j are locally independent).	
	R = S	→	The conditional mean function for each item, $E(X_j\|\theta) = P(X_j = 1\|\theta)$, is represented by either $\frac{e^{a_j(\theta-b_j)}}{1+e^{a_j(\theta-b_j)}}$ *or* $\|(a_j(\theta - b_j))$.	
	E = EIV	→	$V(X_j\|\theta) = P(X_j = 1\|\theta)[1 - P(X_j = 1\|\theta)] > 0$.	
{Co,xPL,1,xOC,EIV}	A = Co	→	Attribute represented by random variate θ; θ continuously distributed over P.	Graded response models (e.g., Samejima's graded response model [Samejima, 1969, 1996], Muraki's modified graded response model Muraki, 1990])
	I = xPL	→	Responding for each item represented by random variable X_j; X_j has x-PL category distribution over P, and $\sum_{k=1}^{x} P(X_{jk}\|\theta) = 1$, $k = 1..x$.	

(*table continues*)

Table 1 (*continued*)

FS	FS → quantitative characterization mapping			Model or set of models
	D = 1	→	$P(\underline{X}=\underline{x}\|\theta)=\prod_{j=1}^{p}P(X^*_{jk}=x^*_{jk}\|\theta)$, $j=1..p$, $k=1..x$, in which $P(X^*_{jk}=x^*_{jk}\|\theta)$ is probability of responding in or above the kth category of the jth item.	
	R = xOC	→	For each j, $l = (x-1)$ boundary response regressions, $\frac{e^{a_j(\theta-b_{jk})}}{1+e^{a_j(\theta-b_{jk})}}$.	
	E = EIV	→	$V(X_{jk}\|\theta) = P(X_{jk}=1\|\theta)[1-P(X_{jk}=1\|\theta)]>0$.	
{Ca,Di,1,M,EIV}	A = Ca	→	Attribute represented by random variate θ; θ discretely distributed across m categories over P, with $\sum_{i=1}^{m}P(\theta = i) = 1, i = 1..m$.	Latent class models (cf. Clogg, 1995; Heinen, 1996; Lazarsfeld, 1950; Lazarsfeld & Henry, 1968)
	I = Di	→	Responding for each item represented by random variable X_j; X_j Bernoulli distributed over P.	
	D = 1	→	$P(\underline{X}=\underline{x}\|\theta)=\prod_{j=1}^{k}P(X_j=x_j\|\theta)$, $j=1..k$.	
	R = M	→	Conditional mean function equal to the probability of response for item j, that is, $E(X_j\|\theta = i) = P(X_j = 1\|\theta = i)$.	
	E = EIV	→	$V(X_j\|\theta = i) = P(X_j = 1\|\theta = i)[1 - P(X_j = 1\|\theta = i)] > 0$.	

Note. A = distributional form of the attribute measured by the items of T (Co = continuous; Ca = categorical); I = item response format (Co = continuous; Di = dichotomous; xPL = x-point graded); D = number of attributes items measure in common (1, . . ., p); R = theoretical form of item/attribute regressions (MI = monotone increasing; LI = linear increasing; S = S-shaped; xOC = x-point ordered categorical; M = flat line or not flat line); E = error characteristics of items (EIV = error in variables; EF = error free).

[a] As is conventional, here Y denotes an observed random variables, whereas θ is used to denote a latent, or unobserved, variable.

ness of the item responses conditional on θ (i.e., the off-diagonal elements of the conditional covariance matrix, Ψ, are all zero). The linear item/attribute regressions specified in the formal structure are paraphrased as linear item/θ regressions. That is, for all items, the mean item response conditional on θ is a linear function of θ. The fallibility of the items as indicators of the attribute is modeled according to the usual factor analytic paraphrase (i.e., that the conditional covariance matrix is positive definite). This means that the elements along the diagonal, which represent the "unique" (error) variances, each have a positive sign. Jointly, the components of this quantitative characterization imply that the population covariance matrix of the items, θ, may be factored as follows:

$$\sum = \Lambda\Lambda' + \psi,$$

in which Λ is a vector of factor loadings and ψ, as stated above, is diagonal and positive definite. This consequence is then a requirement that must be satisfied by the joint distribution in order that a test whose formal structure has been specified to be {Co,Co,1,LI,EIV} may be judged as, in fact, conforming to its formal structure. Table 1 lists classes of quantitative characterizations appropriate for a number of other formal structures that are likely to be encountered in practice.[8]

3. Test of the Conformity of Data to Model

Up to this point we have specified that a test analysis must begin with a specification of the formal structure of the test being evaluated. Then, in order that the conformity of test behavior to the specified formal structure may be established, one must find a quantitative characterization that gives an adequate description of the relations between the items and the attribute they are presumed to measure and among the items themselves. However, putting precise boundaries on what is meant by "adequate" is by no means a simple issue. Rather, it is one that has generated discus-

[8] For a further example of how one might choose an "appropriate" quantitative characterization of a given formal structure, see Mellenbergh's (1994) characterization of a generalized linear item response theory that subsumes many of the better known and employed psychometric models. See also Thissen and Steinberg (1986) for a taxonomy of item response models that may be employed in the analysis of categorical item response data and McDonald (1999) for description of both common factor models and item response models.

sions within the psychometric literature on topics ranging from comparisons of different indices of model fit (cf. Barrett, 2007, and associated commentaries in Volume 42 of *Personality and Individual Differences* for a recent discussion pertaining to structural equation models) to model selection and model equivalence (cf. Lee & Hershberger, 1990; Raykov & Penev, 1999; Stelzl, 1986).

Because here we are concerned primarily with presenting a rationale for analyzing test data, a review of the complex (and considerable) literature on such issues will not be attempted. Rather, as is emphasized by Kane (2006), we wish to underscore the point that if the (latent) variates implied by various measurement models are to function as representations of attributes measured by tests, and the manifest variates (i.e., the item responses) are thought to be linked to each other and to the latent variate in particular ways, "Empirical evaluations of how well the model fits the data . . . provide an empirical check on the validity of the indicators of the latent variables" (p. 43). To this end, test analysts will need to choose an index (or a set of indices) of model fit in order to adequately adjudge whether a given measurement model describes a set of item responses (and, by extension, whether the test behavior conforms to the specified formal structure). For the present purpose, suffice it to say that a given test can justifiably be said to conform to its formal structure in a focal population if (a) the chosen quantitative characterization is an appropriate paraphrase of the formal structure and (b) some chosen fit function is suitably "small" such that it can be reasonably concluded that the quantitative characterization provides a good description of the test data (see Thissen, Steinberg, Pyszczynski, and Greenberg, 1983, for an example that describes a procedure for fitting a unidimensional linear common factor model to a set of 9-point Likert items).

However, regardless of which particular fit function is employed, if one is to justifiably claim that test behavior is (or is not) in keeping with the formal structure of the test, there must be some formal assessment of whether the joint distribution conforms to an appropriately chosen quantitative characterization. In the absence of this, or some other sound justification for judging a test's conformity to its formal structure, any further steps taken in an evaluative test analysis are inherently ambiguous, as the analyst has not established that the test items are indicators of a single attribute, nor that they relate to this attribute in the manner described by the specified formal structure. Hence, there would be no grounds for forming an item composite or for assessing the precision and external test validity of this composite once formed (or other components of validity that require that test scores be conceptualized as indicators of attributes).

4. Derivation of an Optimal, Model-Implied Composite

Typically, test constructors and users alike are not satisfied with measuring a given attribute of interest with a single item. Rather, they want to have multiple indicators of the attribute, indicators that they may ultimately composite in some manner in order to produce a more reliable measure than is given by any single indicator. However, Standard 1.12 in *Standards* clearly states that "where composite scores are developed, the basis and rationale for arriving at the composites should be given" (AERA et al., p. 20). Hence, composites can legitimately be produced only if there exists both a rational argument for doing so (i.e., that the items of the test are presumed to measure a given attribute in common) and evidence that the rationale is sound (i.e., that the item responses can be legitimately represented by some appropriate unidimensional model). If both the rationale and evidence exist, the resulting compositing rule assigns to each individual in a focal population a single real number and thereby scales this population of individuals with respect to the random variate that represents the attribute in question.

Very often, the tests employed in the social and behavioral sciences come with "off-the-shelf" compositing rules, and the undisputed champion of such rules is the unweighted sum. However, it cannot be decided by fiat that a set of items do, in fact, measure a single attribute in common. The issue of a test's compositability in some population is open to question, and claims of compositability must be justifiable. If a test can be shown to be legitimately compositable, the issue becomes which composite to use. The logic is as follows:

1. The formal structure of given test specifies that the items of the test are indicators of, or measure in common, a single attribute.

2. The formal structure is mapped into an appropriate quantitative characterization (i.e., an appropriate unidimensional model is chosen).

3. If the joint distribution of item responses satisfies the requirements imposed by the chosen quantitative characterization, the performance of the test is said to be in keeping with its formal structure. In particular, satisfaction by the joint distribution of the requirements imposed by a given unidimensional model means that the item responses are unidimensional in some particular sense. This, in turn, is taken as meaning that the items measure just one thing in common, arguably the attribute purportedly measured by the test.

4. However, the mapping of the formal structure into a quantitative characterization paraphrases the attribute as a random variate of some sort, most often as a latent variate.

5. Thus, the task of scaling individuals in a focal population with respect to the attribute is paraphrased as the task

of deriving an optimal estimator (predictor)[9] of the random variate representation of the attribute (i.e., the latent variate).

6. When the joint distribution satisfies the requirements of a unidimensional quantitative characterization (i.e., a unidimensional latent variate model), one is justified in estimating the single latent variate. That is, the condition under which it makes sense to develop a single optimal estimator of an unobservable latent variate that is related to the items in a manner described by the formal structure is when the joint distribution satisfies the requirements of an appropriately chosen unidimensional measurement model. This estimator will, of course, be a composite of the item responses and will be created according to some particular compositing rule.

7. The particular form of the optimal compositing rule will be determined jointly by characteristics of the model (e.g., number of free parameters), commitment to a particular definition of "optimal" (e.g., maximum likelihood vs. generalized least squares estimation methods), and pragmatic considerations (e.g., ease of calculability).

For instance, a test consisting of continuous items that are taken to be error-laden indicators of a single "continuously distributed" attribute, and for which the item/attribute regressions are considered to be linear (i.e., a test with formal structure {Co,Co,1,LI,EIV}), is appropriately paraphrased by a ULCF model (see Thissen et al., 1983). If the joint distribution of item responses satisfies the requirements imposed by this quantitative characterization, the items are unidimensional in the linear factor analytic sense, and the common factor is taken as a representation of the attribute the items were designed to measure. Under this condition, one is justified in deriving an estimator of the common factor, a random variate. And, according to the correspondence between the formal structure and this quantitative characterization, estimation of this common factor (by one of a number of different possible methods; cf. McDonald & Burr, 1967) is the operational counterpart of scaling individuals with respect to the attribute for which the items of the test were designed as measures.

Regardless of which optimality criteria (technical, pragmatic, or a combination thereof) are adopted in a given test analysis, the point is that there is no globally correct compositing rule. Any legitimate compositing rule is tied to the union of a particular quantitative characterization (i.e., model), statistical principle, and pragmatic considerations, and there exists latitude in regard to the choice of each. The possibility of virtual exchangeability between unweighted and weighted composites (cf. Grayson, 1988; Wainer, 1976) should not be taken as justification for the perfunctory practice of choosing the former as the default choice. The preference of a particular composite over all others should be the result of a careful consideration of optimality and practicality trade-offs.

5. Estimation of the Reliability of the Composite

In *Standards* (AERA et al., 1999) it is stated that regardless of the form of the items a measure comprises, providing information about measurement error for test data is essential to the proper evaluation of the measure in question. According to the standards listed there, reliability information may be reported as variances or standard deviations of measurement errors, as any of a set of potential indexes (e.g., classical reliability and/or generalizability coefficients), or as test information functions. Furthermore, it is noted that the reporting of reliability coefficients in the absence of the details concerning the methods used in the estimation of such coefficients, sample characteristics, and measurement procedures "constitutes inadequate documentation" (AERA, 1999, p. 31). Although it is recognized that unreliable scores may still be useful in certain contexts of test use,

> the level of a score's unreliability places limits on its unique contribution to validity for all purposes. . .To the extent that scores reflect random errors of measurement, their potential for accurate prediction of criteria, for beneficial examinee diagnosis, and for wise decision making is limited. (p. 31)

The scores whose reliabilities are of interest to the current discussion are those of legitimately formed composites of item responses.

Although the applied researcher typically adopts the expression "reliability" to refer to the measurement precision of test scores generally, the distinction between the definitions of measurement precision born out of classical and modern test theories respectively has long been recognized by psychometricians. Here, we speak generally about the measurement precision of a composite of test items, which

[9] The latent variable models considered herein, and throughout psychometrics, are random latent variable models (i.e., those in which the latent variable has a distribution). Such models are to be contrasted with those in which each person has a "person parameter" to be estimated. However, as Holland (1990) pointed out, there is no sense to the notion of maximum likelihood estimation (or any other type of estimation) of θ in the random models, as θ is not a set of person parameters but, rather, a random variate. Hence, in random latent variable models, θ may be predicted but not estimated. However, maximum likelihood terminology is used here in order to maintain consistency with standard treatments. It should be noted that an additional complication arises in cases in which the chosen quantitative characterization is an indeterminate latent variable model. In such models, there exist an infinity of random variates, U_i, each of which satisfies the requirements for latent variablehood (cf. Guttman, 1955, Schonemann & Wang, 1972, and Steiger & Schonemann, 1978, for discussions of the indeterminacy of factor score matrices); hence, the question of what exactly is being predicted is left ambiguous. Although these issues will undoubtedly ultimately need to be resolved by psychometricians, they do not bear on the logical coherence of the framework proposed herein.

is conceptualized as follows: In an analysis of the performance of a test in a focal population, if a sound (unidimensional) quantitative characterization of the formal structure is shown to describe the joint distribution, one is justified in compositing the item responses, with an optimal, model-implied composite, the latter of which is some function of the items of the test. The precision of the composite will, additionally, be a function of the latent variate (or whichever random variate is a representation of the attribute in the particular quantitative characterization at hand) and can be defined in terms of either reliability or information (see Mellenbergh, 1996). Regardless of whether one chooses to conceptualize precision in terms of reliability or in terms of test information, distinctive forms of each usually can be derived for particular quantitative characterizations.

For example, a test whose formal structure can legitimately be paraphrased by a unidimensional linear common factor model (i.e., {Co,Co,1,LI,EIV}), such as any of the classical methods for estimating reliability of (or lower bounds to) an unweighted sum of the item responses, could be employed to produce an estimate of score precision (e.g., Spearman–Brown, Cronbach's alpha); alternatively, the test information function could be used to indicate the relative degree of score precision for individuals located within a specified range on the latent factor dimension. Regardless of which particular strategy is adopted, the specific nature of the composite will dictate to a large extent which strategies for producing information about score precision are legitimate and which are not.

6. Entering the Composite Into External Test Validation Studies

Up to this point in a data-based test analysis, support has been amassed that the items of the test measure in common just one thing (i.e., empirical support for internal test validity has been given) and a composite has been produced to estimate individuals' relative standing with regard to that one thing. However, it has not been settled that this one thing is, in fact, the attribute the test items were designed to measure (i.e., it is possible to have a very precise measure of a single attribute but one that does not measure the particular attribute that it was designed to measure). In fact, at least insofar as the measurement of traits is concerned, no definitive case can ever be made about the identity of the attribute that the composite measures (Cronbach & Meehl, 1955). Certainly a great deal more evidence can be accumulated that has direct bearing on the provisional claim that the scores on the composite are error-laden measurements of the attribute under study. This evidence is accumulated in an ongoing program of (external) construct validation (cf. Cronbach, 1971; Cronbach & Meehl, 1955; Loevinger, 1957; Peak, 1953).

In general terms, the logic of such a program of investigation can be outlined as follows:

1. A given test, T, comprises a set of items, each of which is designed to be an observable indicator of some (typically unobservable) attribute, A.

2. Variation in the responding of individuals in some focal population to the items of T is caused by a complex web of relations involving A, other attributes, and additional sources (e.g., situational and method factors).

3. The extant theory pertaining to A postulates particular relationships (a) between A and its indicators (or composites of its indicators), (b) between A and other attributes, and (c) between indicators (or composites) of both the same and different attributes (cf. Cronbach & Meehl, 1955). If we use the current notation, the theory pertaining to (a) describes the formal structure of T, whereas the theory pertaining to (b) and (c) describes the nomothetic span of T (cf. Embretson, 1983). It is the latter possibility that bears on what we herein refer to as external test validity.

4. If Steps 1–5 of the test analytic framework have been satisfied, some optimal composite of item responses, denoted here by ϕ, is an estimator of the latent variate, which is taken to be a representation of A. Thus, ϕ stands in for the items of T in all external test validation analyses. In particular, evidence of external test validity accrues from ϕ behaving in a manner that is in keeping with testable consequences of the nomothetic span of T.

5. Because there is, in principle, an infinity of testable consequences of the nomothetic span of T and, at any stage in an ongoing program of construct validation, only a small subset of these can be derived and tested, a given test is, in principle, always deemed to be provisionally construct valid.

Finally, it should be noted that external test validity may be investigated in potentially many different ways, including (but certainly not limited to) producing correlations between composite scores and variables external to the test. Guidelines have been provided for investigating distinct aspects of external test validity in terms of everything from classical test-criterion relationships to discriminant and convergent evidence (Cronbach, 1971) to correlational evidence based on the multitrait–multimethod matrix (Campbell & Fiske, 1959) and many others. The current framework does not itself provide an explicit approach for gathering evidence in support of external test validity; rather, our aim here is to bring the test analyst to the point at which embarking on such investigations may be justified on the grounds that the test scores on which such investigations are based have been produced according to sound logic and have subsequently been shown to have adequate precision. Then, these composites may be meaningfully entered into external test validity investigations, whatever their nature.

Summary and Discussion

It has been claimed here that despite a relative preponderance of test theoretic results generated over the past several decades, very little work has been dedicated to the development of explicit test analytic frameworks that stipulate how such results should be employed in passing judgment on the performance of a test in a given application. It is important to keep in mind that, in its current form, the framework is meant to provide a rationale for conducting data-based test analysis. In such analysis, the primary aim is neither to make discoveries about what a test measures or about the attribute it has been designed to measure but, rather, to determine whether the test items may be coherently composited into suitably precise scores that may then be taken as representations of examinees' standings with respect to the attribute thought to underlie the test. The logic on which the framework is based does not, however, preclude the possibility of making discoveries about the test during the course of an evaluative analysis (e.g., that it does or does not perform as it was designed to perform in a given population or particular context of employment). Nor does it rule out the possibility that there will be equivocation over what constitutes an appropriately specified formal structure or that, over time, changes in either the theoretical structure or certain formal properties of the test (e.g., alterations in item stems or in modes of response) may necessarily lead to changes in the specification of a test's formal structure and, hence, in the appropriateness of a given measurement model as a quantitative characterization. So, although the requirements of the individual data-based test analysis call for a sequential approach, such as the one presented here, the larger game of validation, in which the framework has a limited role, need not proceed in such a systematized fashion.

Some Possible Extensions of the Basic Framework

The framework presented herein assumes that the responding to a test's items is caused by an underlying attribute that the test is designed to measure (albeit an attribute for which a latent variable becomes a representation). This is why such a strong emphasis has been placed on translating the formal structure into a quantitative characterization that is unidimensional in some sense, such that compositing the items of the test into a single metric may be justified. In addition, with the current setup, the test analytic scenarios to which the framework may be reasonably applied are restricted to a consideration of inter- but not intraindividual differences. Admittedly, the omission of these two (and arguably other) features of tests limits the applicability of the framework in its present form to a subset of test analytic scenarios that might reasonably be encountered by applied test analysts.

Here, we consider two possible additions to the formal structure as it is originally presented.[10] First, Edwards and Bagozzi (2000) made the distinction between reflective and formative measures.[11] As they noted, reflective measures are those for which item responding is thought to be "caused" by some attribute (again, for which the latent variate is a represention), and formative measures are those relevant to situations in which the item responses induce a latent variable (e.g., socioeconomic status as induced by compositing in some way a set of measures, such as annual income or highest level of education achieved). Whereas the present work deals only with the reflective case, many testing applications may call for a formative conceptualization (Borsboom, Mellenbergh, & van Heerden, 2003). In order to broaden the class of potential quantitative characterizations (i.e., models), the current framework could be expanded to include a causal status of the items component, for which the possible values would be reflective (R) and formative (F).

A second potential expansion of the formal structure concerns the temporal structure of scores generated from applications of a given test. As is, the framework does not include time as a component of formal structures and, hence, does not allow for straightforward consideration of intraindividual change with respect to the attribute being measured. In order to accommodate testing applications in which time series analyses are the aim, one should add a temporal structure of composite component of the formal structure, with static (S) and dynamic (D) as potential values. Such an addition would lead to the inclusion of a variety of time series models as candidate quantitative characterizations (cf. Hamaker, Dolan, & Molenaar, 2005; Molenaar, 1985, 1994; Molenaar, Huizenga, & Nesselroade, 2003; Rovine & Molenaar, 2005; van Rijn & Molenaar, 2005).

Further Development of Test Theory Models

Although the current framework does not bear directly on advances and developments in psychometric theory, its pragmatic utility is in large part dependent on such advances and developments. In particular, due to its reliance on statistical models, test analysis will be generally more fruitful and informative as improvements in statistical modeling procedures are made. Here we identify two specific areas in

[10] We thank a reviewer of an earlier draft of this article for suggesting these important extensions of the formal structure.

[11] Although Fornell and Bookstein (1982) were the first to use this terminology, Blalock (1964) is credited with first making the distinction between measures as effects of constructs versus causes of them; the distinction has also been identified by Bollen and Lennox (1991) in the context of what they refer to as cause and effect indicators.

which further work in psychometrics and quantitative psychology would benefit test analytic practice.

First, in the context of the present work, the proposed framework accommodates only the specification of necessary, but not sufficient, criteria for particular latent structures. That is, in order to answer to the needs of a truly evaluative test analysis, there must be clear and unambiguous means of adjudging the conformity of test behavior to the specified formal structure. This requirement means that one must be able to articulate the empirical requirements for the joint distribution of item responses; if these requirements are shown to hold, this implies that a given latent structure (and, hence, a given formal structure) underlies the data. However, as it stands right now, the framework presented herein furnishes only the specification of necessary conditions, which do not necessarily imply the latent structure of interest. And so, although one can rightly adjudge that a given latent structure does not underlie responding to the test if the particular empirical requirements do not hold for the joint distribution, one cannot, strictly speaking, claim that the latent structure does underlie the data if the empirical requirements for the joint distribution do, in fact, hold.[12] Rather, one simply acts as if they do and thereby concludes that the test does conform to its formal structure. However, this lack of sufficiency conditions is by no means specific to measurement models but, rather, is a feature of latent variable modeling generally. Development of better and stronger criteria for latent structures will benefit test analysis, with its heavy reliance on latent variable modeling, and will strengthen any area of research in which the use of such modeling procedures has become an accepted methodology.

Second, different formal structures can be accommodated by the framework proposed herein only to the extent that there exist mathematical models into which the components of those complex formal structures may be mapped. The development of more and more complex models will accommodate more and more complex formal structures and will lead to better quantitative translations and, hence, sounder decisions by researchers about the quality of the tests they employ. Moreover, the veracity of the framework is also reliant on the general soundness of the particular statistical modeling procedures employed therein. Because the framework relies on inferential techniques, if a particular procedure is performing poorly, the claims born out of the test analysis may clearly be compromised, even if the test analyst adheres strictly to the framework. To the extent that improvements are made in the modeling techniques that are available to the test analyst, the pragmatic value of the framework will also be increased.

[12] See Maraun, Slaney, and Goddyn (2003) for a description of the logic underlying these two distinct senses of criteria of latent structures, and see McDonald (1967) for an example in which a unidimensional, quadratic factor structure and a two-dimensional, linear factor structure imply the same covariance structure.

References

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

American Psychological Association [APA]. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(Pt. 2), 1–38.

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences, 25,* 815–824.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for Beck Depression Inventory–II.* San Antonio, TX: Psychological Corporation.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Blalock, H. M. (1964). *Causal inferences in nonexperimental research.* Chapel Hill: University of North Carolina Press.

Blinkhorn, S. F. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50,* 175–186.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–360). New York: Plenum.

Costa, P. T., Jr., & McRae, R. R. (1985). *The NEO Personality Inventory manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement theory and public*

policy. Proceedings of a symposium in honor of Lloyd G. Humphreys (pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity and psychological tests. *Psychological Bulletin, 52,* 281–302.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 155–174.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research, 19,* 440–452.

Grayson, D. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53,* 383–392.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of unidimensionality. *Educational and Psychological Measurement, 37,* 827–838.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology, 8,* 17–24.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2005). Statistical modeling of the individual: Rationale and application of multivariate time series analyses. *Multivariate Behavioral Research, 40,* 207–233.

Hattie, J. A. (1984). An empirical study of various indices for determining undimensionality. *Multivariate Behavioral Research, 19,* 49–78.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Thousand Oaks, CA: Sage.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*(4), 802–812.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.

Holland, P. W., & Rosenbaum, P. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14*(4), 1523–1543.

Jöreskog, K. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika, 3,* 165–178.

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.). *Educational measurement* (pp. 17–64, 4th ed.). Washington, DC: American Council on Education/Praeger.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. (Eds.), *Measurement and prediction* (pp. 362–412). Princeton: Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* New York: Houghton Mifflin.

Lee, S., & Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25,* 313–334.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs No. 7.* Iowa City, IA: Psychometric Society.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–549.

Maraun, M. D., Jackson, J. S. H., Luccock, C. R., Belfer, S. E., & Chrisjohn, R. D. (1998). CA and SPOD for the analysis of test comprised of binary items. *Educational and Psychological Measurement, 58,* 916–928.

Maraun, M. D., Slaney, K., & Goddyn, L. (2003). An analysis of Meehl's MAXCOV-HITMAX procedure for the case of dichotomous indicators. *Multivariate Behavioral Research, 38,* 81–112.

McDonald, R. P. (1967). Factor interaction in non-linear factor analysis. *British Journal of Mathematical and Statistical Psychology, 20,* 205–215.

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

McDonald, R. P., & Burr, E. J. (1967). A comparison of four methods of constructing factor scores. *Psychometrika, 34,* 381–401.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37,* 113–115.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293–299.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1021–1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–46). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 13–103). New York: MacMillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific enquiry into score meaning. *American Psychologist, 50,* 741–749.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35–44.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–62.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate tiem series. *Psychometrika, 50,* 181–202.

Molenaar, P. C. M. (1994). Dynamic latent variable models in developmental psychology. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 155–180). Thousand Oaks, CA: Sage.

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of interindividual and intraindividual variability: A theoretical and empirical vindication of developmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development: Dialogues with lifespan psychology* (pp. 339–360). Dordrecht, the Netherlands: Kluwer Academic.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71.

Peak, H. (1953). Problems of objective observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 243–299). New York: Holt, Rinehart &Winston.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Raykov, T., & Penev, S. (1999). On structural equation model equivalence. *Multivariate Behavioral Research, 34,* 199–244.

Roskam, E. E., & Ellis, J. (1992). "The irrelevance of factor analysis for the study of group differences": Commentary. *Multivariate Behavioral Research, 27,* 205–218.

Rovine, M. J., & Molenaar, P. C. M. (2005). Relating factor models for longitudinal data to quasi-simplex and NARMA models. *Multivariate Behavioral Research, 40,* 83–114.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*(Suppl. 4).

Samejima, F. (1996). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Schonemann, P. H., & Wang, M.-M. (1972). Some new results on factor indeterminacy. *Psychometrika, 37*(Pt. 1), 61–91.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Slaney, K. L. (2006). *The logic of test analysis: An evaluation of test theory and a proposed logic for test analysis.* Unpublished doctoral dissertation, Simon Fraser University, Burnaby, British Columbia, Canada.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Steer, R. A., Ranieri, W. F., Kumar, G., & Beck, A. T. (2003). Beck Depression Inventory–II items associated with self-reported symptoms of ADHD in adult psychiatric outpatients. *Journal of Personality Assessment, 80,* 58–63.

Steiger, J. H., & Schonemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis* (pp. 136–178). San Franciso: Jossey-Bass.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research, 21,* 309–331.

Tellegen, A. (1982). *Brief manual for the Multidimensional Personality Questionnaire.* Unpublished manuscript, University of Minnesota.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577.

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement, 7,* 211–226.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education, 67,* 335–341.

van Rijn, P. W., & Molenaar, P. C. M. (2005). Logistic models for single subject time series analysis. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 125–145). Mahwah, NJ: Erlbaum.

van Schuur, W. H., & Kiers, H. A. (1994). Why factor analysis often is the incorrect model of analyzing bipolar concepts, and what model to use instead. *Applied Psychological Measurement, 18,* 97–110.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213–217.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58,* 21–37.

Received December 17, 2007
Revision received September 23, 2008
Accepted September 29, 2008 ■