Mid-Western Educational Researcher

Volume 23 | Issue 1 Article 6

2010

Contemporary Treatment of Reliability and Validity in Educational Assessment

Dimiter M. Dimitrov
George Mason University

MWERA 2009 Presidential Address

Contemporary Treatment of Reliability and Validity in Educational Assessment

Dimiter M. Dimitrov
George Mason University

Abstract

The focus of this presidential address is on the contemporary treatment of reliability and validity in educational assessment. Highlights on reliability are provided under the classical true-score model using tools from latent trait modeling to clarify important assumptions and procedures for reliability estimation. In addition to reliability, indices of measurement precision that provide information about error tolerance are also discussed. Regarding validity, the focus is on moving from the discrete construct-based model of validity (Cronbach & Meehl, 1955), which still seems to dominate education assessment research and practices, to the unified construct-based model of validity (Messick, 1989, 1995).

The topic of this presidential address was motivated primarily by my work as the editor of *Measurement and Evaluation in Counseling and Development* (*MECD*)—the official journal of the Association for Assessment in Counseling and Education (AACE). A particular concern in my editorial experience has been that, despite the availability of contemporary approaches to evaluating scale reliability and validity, a large number of manuscripts still involve outdated methods that yield potential threats to valid interpretations and decision making in education, psychology, and related fields. I will leave to your judgment whether this is also the case in some (if not most) dissertations at graduate schools of education nationwide. Typical problems relate to a lack of testing for assumptions in evaluating reliability, limited perspective on measurement precision, and methodological drawbacks in validation processes. I hope that this presentation will provide highlights that can be useful to researchers in studies that involve evaluation of scale reliability and validity for assessment in education.

Highlights on Contemporary Treatment of Reliability

What is Reliability?

In general, the **reliability** of measurements indicates the degree to which they are *accurate*, c*onsistent*, and *replicable* when (a) different people conduct the measurement, (b) using different instruments that purport to measure the same trait, and (c) there is incidental variation in measurement conditions. That is, the reliability of scores shows the degree to which they are "free" of random error. Before I comment on limitations of traditional approaches (e.g., using Cronbach's *alpha*) and advantages of some contemporary approaches to evaluating scale reliability in the classical (true-score) framework, the introduction of some basic concepts seems appropriate.

True-Score Model

A basic assumption in the classical (true-score) model of measurement is that the observed score, *X*, is a sum of a true score, *T*, and random error, *E*. That is,

$$X = T + E. \tag{1}$$

In general, a person's *true score*, *T*, is the mean of the theoretical distribution of scores that would be observed in repeated independent measurements using the same test. Clearly, *T* is a hypothetical concept because it is not practically possible to test the same person infinity times in independent repeated measurements, given that each testing could influence the subsequent testing (e.g., due to "carry over" effects of practice or memory). From the definition of true scores, it follows that the variance of the observed scores is a sum of the variance of the true scores and the error variance (e.g., Zimmerman, 1975). That is,

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{2}$$

As to the error scores (residuals), *E*, it is assumed that they are random and follow a normal distribution with a mean of zero and a variance σ_E^2, that is $E \sim N\left(0, \sigma_E^2\right)$.

The *reliability* of a measurement scale, denoted here ρ_{XX}, is defined as the correlation between the observed scores on two *parallel tests*—i.e., tests with equal true scores and equal error variances for every population of examinees taking both tests. Equivalently, ρ_{XX} indicates what proportion of the observed score variance is true score variance. That is,

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2}. \tag{3}$$

Perfect reliability ($\rho_{XX} = 1$) can theoretically occur when $\sigma_T^2 = \sigma_X^2$ or, equivalently, when $\sigma_E^2 = 0$. The error standard deviation, σ_E, referred to also as the *standard error of measurement* (*SEM*), is typically estimated as

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX}}. \tag{4}$$

The assumptions underlying scale reliability and its estimation involve the concepts of congeneric measures, parallel measures, tau-equivalent measures, and essentially tau-equivalent measures. To better understand the meaning of these concepts, they are defined here in a latent trait framework. For simplicity, let's consider the case depicted in Figure 1, where three test items, X_1, X_2, and X_3, serve as indicators of a single latent trait, η, being measured by the test (e.g., η can be reading ability, test anxiety, etc.)

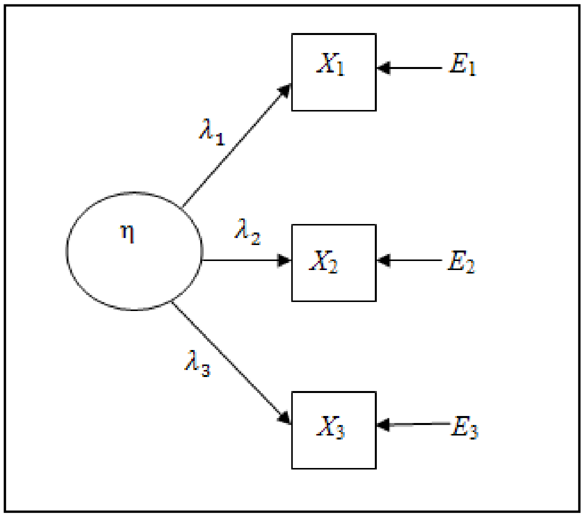

Figure 1. A unidimensional construct η, as measured by three indicators X_1, X_2, and X_3

Analytically, the observed scores X_1, X_2, and X_3 in Figure 1 can be presented as follows:

$$X_1 = (\lambda_1\eta + a_1) + E_1,$$
$$X_2 = (\lambda_2\eta + a_2) + E_2,$$
$$X_3 = (\lambda_3\eta + a_3) + E_3, \qquad (5)$$

where the expression in parentheses (λη + *a*) represents the predicted value of the observed score, *X*, from the latent trait, η, via a simple linear regression, and *E* stands for the error term. As the predicted value of an observed score is, in fact, the true value for this score, *T*, we have:

$$T_1 = \lambda_1\eta + a_1,$$
$$T_2 = \lambda_2\eta + a_2,$$
$$T_3 = \lambda_3\eta + a_3. \qquad (6)$$

Thus, the true scores T_1, T_2, and T_3 on the three items that measure a single latent trait, η, are obtained by regressing the observed scores (X_1, X_2, and X_3) on η. The regression coefficients, referred to also as *factor loadings*, are λ_1, λ_2, and λ_3, and the intercepts are a_1, a_2, and a_3.

Congeneric Measures

Congeneric measures represent the most general case of unidimensional measures in the sense that they may have different scale origins, different units of measurement and may vary in precision. In the context of Figure 1 (see also Equations 6), (a) different scale units means that the regression coefficients (λ_1, λ_2, and λ_3) may differ, (b) different scale origins means that the intercepts (a_1, a_2, and a_3) may differ, and (c) variation in precision means that the variances of the error terms, VAR(E_1), VAR(E_2), and VAR(E_3), may differ.

Parallel Measures

Parallel measures represent the most restricted case of unidimensional measures in the sense that they have the same units of measurement, scale origins, and error variances. In the context of Figure 1, X_1, X_2, and X_3 would be parallel measures under the following restrictions

$$\lambda_1 = \lambda_2 = \lambda_3,$$
$$a_1 = a_2 = a_3, \text{ and}$$
$$\text{VAR}(E_1) = \text{VAR}(E_2) = \text{VAR}(E_3). \qquad (7)$$

As one can also notice, parallel measures have equal true scores and equal error variances.

Tau-equivalent measures

Tau-equivalent measures have the same units of measurement and scale origins, but their error variances may differ. In Figure 1, X_1, X_2, and X_3 would be tau-equivalent measures under the following restrictions

$$\lambda_1 = \lambda_2 = \lambda_3 \text{ and}$$
$$a_1 = a_2 = a_3. \qquad (8)$$

Essentially tau-equivalent measures

Essentially tau-equivalent measures have the same units of measurement, but dissimilar origins and unequal error variances. In Figure 1, X_1, X_2, and X_3 would be essentially tau-equivalent measures under the following restrictions:

$$\lambda_1 = \lambda_2 = \lambda_3. \qquad (9)$$

Limitations of Cronbach's alpha

It would be fair to say that Cronbach's alpha (Cronbach, 1951) is still the most commonly used index of internal consistency reliability. It should be emphasized, however, that Cronbach's alpha is an accurate estimate of the population scale reliability only under the assumptions that (a) the measures are essentially tau-equivalent and (b) there are no correlated error terms. In case that the latter assumption is in place, but the measures are not essentially tau-equivalent (i.e., the measures may differ in units of measurement), Cronbach's alpha underestimates the population scale reliability (e.g., Novick & Lewis, 1967; Raykov, 1997). In case of correlated errors Cronbach's alpha typically overestimates the population scale reliability (e.g., Zimmerman, Zumbo, & Lalonde, 1993). Correlated errors may occur, for example, with adjacent items in a multicomponent instrument, with items related to a common stimulus (e.g., same paragraph or graph), or with tests presented in a speeded fashion (Komaroff, 1997; Raykov, 2001). Thus, Cronbach's alpha cannot be in general considered a dependable estimator of scale reliability. Presented next is a contemporary approach to evaluating reliability in the general case of congeneric measures (i.e., measures that may have different scale origins, different units of measurement, and unequal error variances).

Evaluation of Scale Reliability Using Latent Variable Modeling

For specificity, consider again the unidimensional test model depicted in Figure 1 (see also Equations 5 and 6). In this context, if $X = X_1 + X_2 + X_3$ is the total test score, Equation 3 for the reliability of X, ρ_{XX}, can be translated as follows (e.g., Bollen, 1989):

$$\rho_{XX} = \frac{\left(\lambda_1 + \lambda_2 + \lambda_3\right)^2}{\left(\lambda_1 + \lambda_2 + \lambda_3\right)^2 + VAR\left(E_1\right) + VAR\left(E_2\right) + VAR\left(E_3\right)}. \quad (10)$$

With correlated errors (assuming model identification), the right-hand side of Equation 10 needs to be extended by adding twice the sum of error covariances in the denominator (Bollen, 1989, p. 220). This extension assumes that the model with the added error covariances is identified.

A readable discussion of the latent variable modeling approach to evaluating reliability through the use of Equation 10 is provided by Raykov (2009). He also provides a syntax code in the computer program M*plus* (Muthén & Muthén, 2008) for point and interval estimation of scale reliability of congeneric measures. A different approach to point evaluation of reliability for scales with binary items is proposed by Dimitrov (2003). This approach allows researchers to evaluate the reliability of the composite scale for a test, as well as the reliability of individual test items, based only on estimates of the items parameters obtained with the one-, two-, or three-parameter model in items response theory (IRT). Using formulas developed by Dimitrov (2003), Raykov, Dimitrov, and Asparouhov (in press) applied the latent variable modeling approach to point and interval estimation of reliability for scales with binary items.

Multiple Aspects of Precision in Measurement

In a seminal article on precision of measurements, Kane (1996) argued that the standard error of measurement and reliability coefficients are very useful, but do not capture all aspects of the precision of measurements. He noted that "a more fundamental way to evaluate precision is to compare errors of measurement with the tolerance for error in a particular context. The tolerance for error specifies how large the errors can be before they interfere with the intended use of the measurement procedure and is based on an analysis of the requirements for precision in that context" (Kane, 1996).

Error-Tolerance Ratio (E/T)

To address the evaluation of tolerance for errors, Kane (1996) introduced the *error-tolerance ratio* (*E*/*T*). In the context of the classical true-score model, he defined *E*/*T* as the ratio "error standard deviation to true-score standard deviation," that is

$$E / T = \frac{\sigma_E}{\sigma_T}. \quad (11)$$

The rational behind this definition of *E*/*T* was that "the tolerance for error for each individual can be defined as the individual's true deviation score, and in this context, the root mean square tolerance is simply the standard deviation of the true scores" (Kane, 1996).

Signal-to-Noise Ratio (S/N)

The inverse of the *E*/*T* is referred to as *signal-to-noise ratio* (*S*/*N*), that is

$$E / T = \frac{\sigma_T}{\sigma_E}. \quad (12)$$

The signal-to-noise ratio (*S*/*N*) provides somewhat different perspective on precision in the sense that differences among examinees in the population are taken as the "signals" to be detected, and the true-score standard deviation is taken as an index of the overall strength of this signal. On the other hand, the errors are viewed as noise, and the standard error is taken as an index of the potential impact of this noise in obscuring the signal (Kane, 1996).

It is important to note that the scale reliability can be represented as an explicit function of the error-tolerance ratio (*E*/*N*) or the signal-to-noise ratio (*S*/*N*). Specifically,

$$\rho_{XX} = \frac{1}{1 + (E / T)^2} = \frac{(S / N)^2}{(S / N)^2 + 1}. \quad (14)$$

Relative Errors within a Margin of Tolerance

As noted earlier, Kane (1996) argued that a more fundamental way to evaluate precision is to compare errors of measurement with the tolerance for error in a particular context. He also indicated that "the tolerance for error for each individual can be defined as the individual's true deviation score" (Kane, 1996). In the original metric of measurement, this view on precision translates into the ratio $E/(T - \mu)$ which shows what proportion is the measurement error for an individual from the true-deviation score for that individual. In this ratio, *E* and *T* are the error and true score, respectively, for an individual, whereas μ is the population mean of true scores (which is also the population mean of observed scores, *X*). That is, $E/(T - \mu)$ represents the *relative error of measurement* (*REM*) for an individual true deviation score.

An important question is then what percent of the population scores have *REM* which is smaller in absolute value than a prespecified *margin of tolerance*, δ. In probability parlance, this question translates as follows "What is the probability that a randomly selected score will have *REM* between –δ and δ?" (the margin of tolerance is a positive number, δ > 0). Denoting this probability *PREM*(δ), Dimitrov (2009) showed that

$$PREM(\delta) = P\left(-\delta < \frac{E}{T - \mu} < \delta\right) = \frac{2}{\pi} arctan\left(\delta \frac{\sigma_T}{\sigma_E}\right), \quad (15)$$

where π is the well known mathematical constant (π ≈ 3.1416), *arctan*(.) stands for *arctangent* — the inverse of the trigonometric function *tangent*, tan (.), and δ is a prespecified margin of tolerance for the relative error. By representing the signal-to-noise ratio (σ_T/σ_E), which appears in Equation 15, as a function of the reliability, ρ_{XX}, Equation 15 becomes

$$PREM(\delta) = \frac{2}{\pi} arctan\left(\delta\sqrt{\frac{\rho_{XX}}{1-\rho_{XX}}}\right). \qquad (16)$$

Thus, given the scale reliability, ρ_{XX}, researchers can determine what percent of the population scores have a tolerable relative error, 100**PREM*(δ), which will allow them to better generalize the precision of measurements in making validity judgments. Moreover, *PREM*(δ) can be computed using hand-held calculators that have the *arctan*(.) function; (tan^{-1} is used to denote *arctan* in some calculators).

The margin of tolerance, δ, is selected by the researcher based on his/her judgment about how much relative error is tolerable to allow for valid interpretations of the measures within a specific context. Interestingly, if we select $\delta = \sigma_E/\sigma_T$, i.e., the Kane's (1996) error-tolerance ratio (see Equation 11) and use Equation 15, we obtain *PREM*(δ) = 0.5. Thus, 50% of the individual relative errors, E/(T – μ), are smaller than the Kane's E/T in absolute value. In other words, E/T represents the population median of the distribution of absolute relative errors (Dimitrov, 2009).

From another angle, suppose the scale reliability is ρ_{XX} = .90 and we want to know what percent of the population scores have a relative error, $E/(T - \mu)$, smaller than 0.1 in absolute value. Replacing δ with 0.1 and ρ_{XX} with .90 in Equation 16, we obtain:

$$PREM(\delta) = \frac{2}{\pi} arctan\left(0.1*\sqrt{\frac{0.9}{1-0.9}}\right) = 0.1855.$$

Thus, *PREM*(δ) indicates that 18.55 percent of the relative errors in the population of individual scores are smaller in absolute value than the prespecified margin of tolerance (δ = 0.1).

It is important to note that the relative error of measurement, $E/(T - \mu)$, remains invariant across linear transformations of the scores thus allowing to generalize findings about the percent of relative errors within a margin of tolerance, *PREM*(δ), across such transformations.

Highlights on Contemporary Treatment of Validity

What is Validity?

While reliability of scores deals with their accuracy and consistency, *validity* has to do with whether an instrument measures what it purports to measure. One validates not a test, but an *interpretation of data arising from a specified procedure* (Cronbach, 1971). Historically, there are three major stages in the development of validity models:

1. *Criterion-based model* (e.g., Cronbach & Gleser, 1965) in which validity of measures is viewed as the degree to which these measures are consistent with (or "predict") the measures on a specific "criterion,"
2. *Construct-based model* (Cronbach & Meehl, 1955) which considers three different types of validity— content validity, criterion validity, and construct validity, and
3. *Unified construct-based model of validity* (Messick, 1989, 1995).

Under the **criterion-based model**, the validity of test scores was depicted as the degree to which these scores were accurate representations of the values of a specified *criterion.* A major drawback of the criterion-based conception of validity is that (a) it is too limited and does not capture some basic (e.g., content-related) aspects of validity and (b) it is not possible to identify criterion measures in some domains.

While the **construct-based model** of validity does a better job in this regard, it's major problem is that content validity, criterion validity, and construct validity are depicted as different types of validity. This can mislead test users to believe that these three "types of validity" are comparable or, even worse, that they are equivalent and, thus, collecting evidence for any of them is sufficient to label a test as valid. Messick (1995) argued that the different kinds of inferences from test scores require different kinds of evidence*,* not different kinds of validity.

The **unified construct-based model of validity** is based on a definition of validity provided by Messick (1989): "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13). This conception of validity represents a *unified construct-based model of validity*, by providing a comprehensive view that integrates content-related and criterion-related evidence into a unified framework of construct validity and empirical evaluation of the meaning and consequences of measurement.

A comprehensive **definition of the construct** under validation allows one to identify the behavioral boundaries of the construct, differentiate the construct from other (similar or dissimilar) constructs, and specify relationships between the construct and other constructs. For example, the construct measured by the reading comprehension section on the verbal part of a large-scale standardized test is defined as "one's ability to reason with words in solving problems," and it is expected that "reasoning effectively in a verbal medium depends primarily on ability to discern, comprehend, and analyze relationships among words or groups of words and within larger units of discourse such as sentences and written passages" (ETS, 1998).

Typically, the core definition of a construct is embedded into a more general theory and then refined and operationalized in the context of the theory and practice in which inferences and decisions are to be made based on assessment scores. Based on the adopted construct definition, instrument developers should build a detailed *construct model* that specifies (a) the internal structure of the construct—i.e., its componential structure, (b) the external relationships of the construct to other constructs, (c) potential types of indicators (items) for measuring behaviors that are relevant to assessing individuals on the construct, and (d) construct-related processes—e.g., causal impacts that the construct is expected to have on specific behavior(s).

Messick (1995) specifies six aspects of the unified conception of construct validity—content, substantive, structural, generalizability, external, and consequential aspects. In addition, r*esponsiveness and interpretability* aspects of validity were proposed by the Medical Outcomes Trust (1995) to complete these six criteria under the unified construct-based model of validity.

Content Aspect of Validity

The *content aspect* of validity includes evidence of content relevance, representativeness, and technical quality. In educational assessment, evidence of content validity is gathered primarily through curriculum analysis and inquiry into the nature of knowledge, skills, and other characteristics targeted with the assessment.

Substantive Aspect of Validity

The *substantive aspect* of validity refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks. Evidence about the substantive aspect of validity can be collected through cognitive modeling of the examinees' response processes, observations of behaviors exhibited by the examinees when answering the items, analysis of scale functioning, consistency between expected and empirical item difficulties, and other relevant procedures.

Structural Aspect of Validity

The *structural aspect of validity* appraises the fidelity of the scoring structure to the structure of the construct domain at issue. Typically, evidence of the structural aspect of validity is sought by correlational and measurement consistency between the constructs and their indicators (test items). This is done primarily through the use of factor analysis. An exploratory factor analysis (EFA) is used when there is no enough theoretical or empirical information to hypothesize how many constructs underlie the initial set of items and which items form which factor. EFA is typically used earlier in the process of scale development and construct validation.

A confirmatory factor analysis (CFA) is used in later phases of scale validation after the underlying structure has been established on prior empirical and/or theoretical grounds. Thus, CFA is employed when the goal is to test the validity of a hypothesized model of constructs (factors) and their relationships with a set of observable variables (items, indicators).

Generalizability Aspect of Validity

The *generalizability aspect of validity* examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks, including validity generalization of test criterion relationships. To collect evidence related to the generalizability aspect of validity means to identify the boundaries of the meaning of the scores across tasks and contexts. Typical procedures for collecting such evidence deal with testing for invariance of targeted constructs across groups and/or time points, item bias, consistency of predictions across groups, contextual stability, and reliability.

External Aspect of Validity

The *external aspect of validity* includes convergent and discriminant evidence from multitrait-multimethod comparisons, as well as evidence of criterion relevance and applied utility. The operational definition of a construct is based on a specific theory and, therefore, the validity of the measurable indicators of the construct depends on the correctness of this theory. For example, if we adopt Rosenberg's (1965) theoretical argument that a student's level of "self-esteem" is positively related to participation in school activities, high positive correlation between students' scores on Rosenberg's self-esteem scale and measures of their involvement in school activities will provide convergent evidence of the external aspect of validity for the self-esteem scale.

Consequential Aspect of Validity

The *consequential aspect of validity* appraises the value implications of score interpretations as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice. Both short-term and long-term consequences should be evaluated. It is important to make sure that negative consequences have not resulted from drawbacks of the assessment such as (a) *construct underrepresentation*—the assessment is too narrow and fails to measure important dimensions or facets of the construct, and/or (b) *construct-irrelevant variance*—the assessment allows for variance generated by sources unrelated to the target construct (e.g., item bias).

Responsiveness and Interpretability Aspects of Validity

Responsiveness and interpretability are proposed by the Medical Outcomes Trust (1995) to complement the six criteria described by Messick (1995) under the unified construct-based model of validity (see also Wolfe & Smith, 2007a; 2007b). While responsiveness is considered for support to the external aspect of validity, interpretability is considered as an aspect of validity which reveals the degree to which qualitative meaning can be assigned to quantitative measures. Thus, the *interpretability aspect of validity* indicates how well the meaning of assessment scores is communicated to people who may interpret the scores but are not necessarily familiar with the psychometric terminology and concepts in assessment. For example, the proper communication of *norm-referenced* versus *criterion-referenced* assessment scores is critical for their valid interpretation by a relatively large audience (e.g., practitioners, clients, parents, social workers, policy makers, etc.).

Conclusion

I hope that this presentation provides some important highlights on the contemporary treatment of reliability and validity in educational assessment. In addressing reliability issues, I tried to focus your attention on two major issues. First, researchers should be aware of potential problems and limitations of the (still) commonly used Cronbach's alpha as an index of scale reliability. A more accurate and flexible approach to evaluating scale reliability, which works in the general case of congeneric measures (i.e., different origins, units of measurement, and error variances), is available in contemporary treatments of scale reliability using latent variables modeling (e.g., Raykov, 1997, 2009; Raykov, Dimitrov, & Asparouhov, in press). Second, I tried to emphasize the argument that, along with reliability and standard error of measurement, important aspects of the precision of measurements are addressed via evaluating error-tolerance ratio (E/T), signal-to-noise-ratio (S/N), and proportion of relative errors of measurement that are smaller in absolute value than a prespecified margin of tolerance, *PREM*(δ). Researchers can use information on the precision of measurements provided by E/T, *S*/*M*, and *PREM*(δ) in making validity judgments.

Speaking of validity, my concern is that the contemporary treatment of validity, based on the unified construct-based model of validity (e.g., Messick, 1989, 1995), still does not seem to dominate designs, procedures, and terminology involved in developing, validating, and using instruments for assessment in education. I hope that this presentation will sharpen the focus of educational researchers and practitioners on this issue and will help them in reaching higher standards of quality in education.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

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

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27,* 440-458.

Dimitrov, D. M. (2009). *Estimation of some familiar and new indexes of precision of measurements: A latent variable modeling approach*. A manuscript submitted for publication.

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education, 9*(4), 355-379.

Medical Outcomes Trust Scientific Advisory Committee (1995). Instrument review criteria. *Medical Outcomes Trust Bulletin*, 1-4.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103).

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

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

Raykov, T., Dimitrov, D. M., & Asparouhov, T. (in press). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling*.

Raykov, T. (2009). Evaluation of scale reliability for unidimensional measures using latent variable modeling. *Measurement and Evaluation in Counseling and Development*, *42*(3), 223-232.

Wolfe, E. W., & Smith, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I—Instrument development tools. *Journal of Applied Measurement, 8*(1), 97-123.

Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement, 8*(2), 204-234.

Zimmerman, D. W. (1975). Probability measures, Hilbert spaces, and the axioms of classical test theory. *Psychometrika*, *40*, 221-232.

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