Developing a Strong Program of Construct Validation: A Test Anxiety Example

Jeri Benson
University of Georgia

What would a strong program of construct validation look like for the concept of test anxiety? What are the components of strong validation programs? In particular, how does structural equation modeling fit into such a program?

Validation is the most critical step in test development and use because it is the process by which test scores take on meaning. That is, one does not validate a test per se, but what is validated is the interpretation of the scores derived from the test (Cronbach, 1971; Messick, 1989). Furthermore, validation must be viewed as a matter of degree, not an all-or-nothing property. Thus, one study does not validate or fail to validate the scores from a test. Numerous studies may be required, utilizing different approaches, different samples, and different populations to build a body of evidence that supports or fails to support the validity of the scores derived from a test. As such, validation is a continual process which is not captured in one numerical index. Even when a large body of evidence exists that supports the validity of the use of a score from a particular test (e.g., Wechsler Intelligence Scales), on-going validation studies are needed as our interpretation of the trait changes due to shifting social or cultural conditions. Thus, for scores from a test to remain valid over time, their validity must be reestablished periodically. In this article, the term *test* will be used throughout for consistency. However, the term will encompass such terms as instrument, scale, measure, and inventory and refer to both cognitive and affective measurements.

The Theory of Construct Validation

A construct represents an abstract variable derived from observation or theory. Cronbach and Meehl (1955) defined a construct as an "attribute of people, assumed to be reflected in test performance" (p. 283). It is the attribute about which we want to make an interpretation based on the test score. For example, the attribute we may be interested in drawing an inference about might be the degree of test anxiety, self-efficacy, or motivation as measured by a particular test.

The process by which test scores take on meaning through construct validation is very similar to the way in which scientific theories are developed and evaluated and is illustrated in Figure 1. Starting, usually, with observations and information from previous research, a theory of the construct is formulated. In the formulation of the theory, the relationships among the focal construct and other constructs are described. This set of relationships is referred to as a *nomological network* (Cronbach & Meehl, 1955). Hypotheses involving the constructs within the nomological network are generated as well as rival hypotheses which attempt to explain the observed behavior. The hypotheses are tested one at a time, and conclusions are drawn. The conclusions might: (a) require further observations of behavior, (b) provide an alternative interpretation of previous research, (c) indicate a revision of the theory, (d) suggest additional hypotheses, or (e) offer support for the theory. Each of these alternatives aids in our understanding of the construct. Thus, the establishment of construct validity of a test score is an iterative process whereby the theory and the test are constantly being evaluated and refined as depicted in Figure 1.

Strong Programs of Construct Validation

Cronbach (1989) characterized two general approaches to construct validation research: strong and weak programs. The weak program is exemplified by a heavy dose of exploratory empirical research, such as collections of semi-related correlations among measures of the focal construct and measures of other constructs that appear to be "raked together" (Cronbach, 1989, p. 155). The weak program can partially be attributed to statements from the 1954 and 1966 versions of the *Standards for Educational and Psychological Tests (Standards)* suggesting "the more information the test developer provides, the better" (Cronbach, 1988, p. 13).

Jeri Benson is a Professor in the Department of Educational Psychology, College of Education, 325 Aderhold, University of Georgia, Athens, GA 30602. Her specializations are construct validation and covariance modeling.

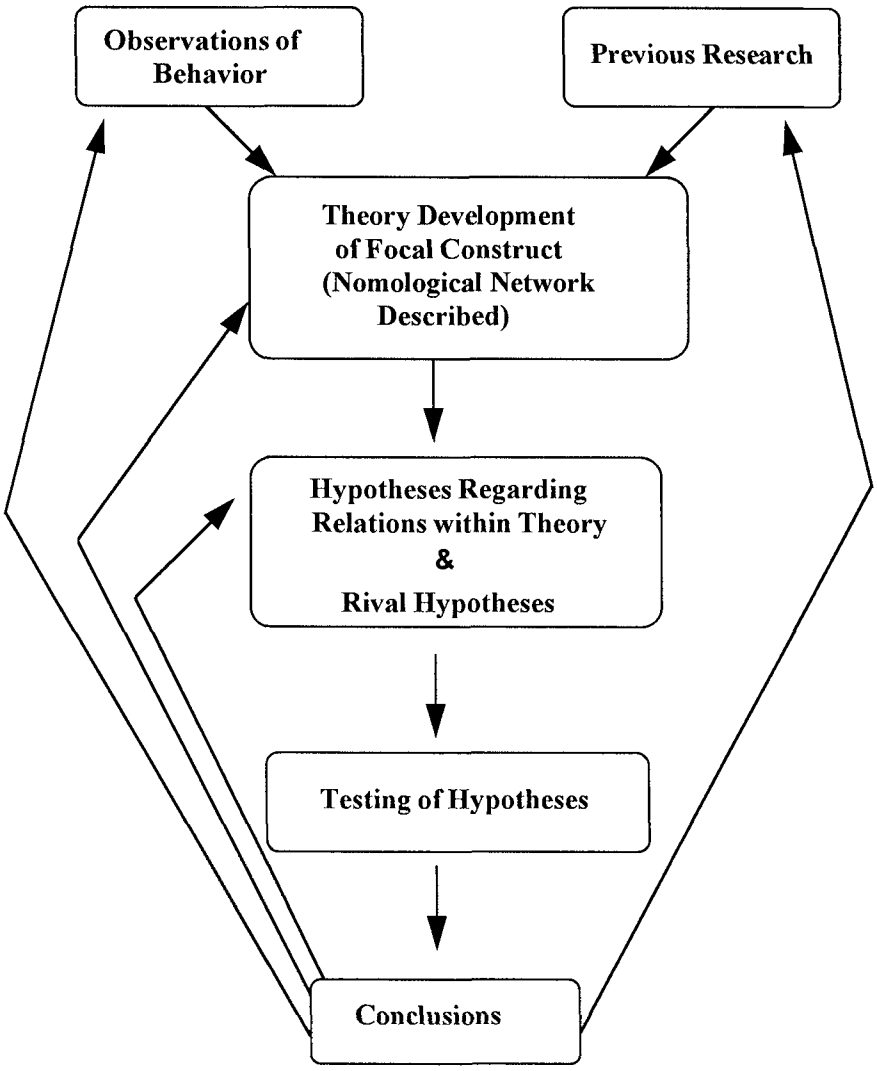

FIGURE 1. *Process of construct validation (theory testing)*

The strong program is reflected in statements from the *Standards* published in 1974 and 1985. A strong program of construct validation is typified by the prominent role theory plays in validation. Actually, Cronbach and Meehl (1955) discussed the importance of theory preceding and guiding test development and validation, followed by the testing of rival hypotheses to evaluate the "validity" of the theory. Over time, if numerous falsification attempts fail, the theory gains strength and acceptability.

The belief in a strong program of validation also was reflected in the early writing of Loevinger (1957) and later by Nunnally (1978) who described the process of construct validation as consisting of three aspects: a substantive component, a structural component, and an external component. The substantive component is where the theoretical domain of the construct is specified and then operationally defined in terms of the observed variables (e.g., the behaviors that reflect the construct). The structural component involves relating the items to the structure of the construct by determining to what extent the observed variables relate to one another and to the construct. It is the external component that begins to give meaning to test scores by determining whether or not the measures of a given construct relate in expected ways with measures of other constructs. The three aspects of construct validity are recast here as stages to convey that construct validation is indeed a process.

Messick (1989) has suggested the varieties of validity evidence are not alternatives but supplements to one another. He defined validity as "an integrated judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores and other modes of assessment" (1989, p. 13). Messick has taken a firm stand on construct validity's being necessary not only from a scientific point of view (which was Loevinger's perspective in 1957) but even in applied settings, where evidence of content- or criterion-related validity might have been sufficient in the past.

Drawing from the suggestions made by Loevinger (1957), Messick (1975, 1980) wrote that content validity was not an indicator of validity at all and that criterion-related validity was too specific to reflect the "accepted" definition of validity. Hence, content- and criterion-related validity provide evidence in the building of the network of relations involving the construct but should not be considered separate and distinct forms of validity. Messick (1980) argued that "the different kinds of *inferences* from test scores require different kinds of *evidence*, not different kinds of validity" (p. 1014, italics added). One drawback to having three kinds of validity has been that users may think they have the option of focusing on one of the forms, as though they were equivalent or comparable. Even worse, naive test users may think that any form of validity evidence is sufficient to label a test as valid.

To incorporate his new thinking regarding validity, Messick (1995) described six aspects of construct validity: content relevance and representativeness; the substantive aspect; the structural aspect; generalizability; the external aspect; and the consequential aspect. The substantive, structural, and external aspects clearly parallel the three components of construct validity identified by Loevinger (1957). It is these three components or aspects of construct validation which are essential to fulfill the requirements of a "strong validation" program, as envisioned by Cronbach (1989).

The purpose of this article is to demonstrate how such a program of

strong construct validation could be applied to the assessment of the construct of test anxiety. Special attention is given to the substantive, structural, and external aspects of construct validation in this article because it is these three aspects which necessitate a strong *psychological* theory to guide the researcher's decisions in designing and evaluating the results from validation studies.

Substantive Stage: Definition of the Theoretical and Empirical Domains of the Construct

In the substantive stage of validation, we are concerned with how the trait and, hence, construct is defined, theoretically and empirically. All constructs are thought to be represented by two domains: a theoretical and an empirical domain. The theoretical domain evolves from the scientific theory surrounding the trait, previous research, as well as one's own observations as shown earlier in Figure 1. The definition of the theoretical domain represents our best understanding of the construct. While no one researcher can hope to define completely the theoretical domain of a construct, over a series of studies and by drawing on the work of others, researchers should be able to bring the boundaries of the theoretical domain into focus. The results from the types of studies suggested under the structural stage (next section) will assist in the refinement of the definition of the theoretical domain.

Constructs also have a corresponding empirical domain which operationalizes the construct. The empirical domain comprises the specific set of observed variables that are used to measure the construct. That is, the empirical domain contains *all* the potential observables (observed variables) and ways those observables can be measured (e.g., performance tasks, self-ratings, clinical observations). As such, the empirical domain is a reflection of the theoretical domain. Initially, it is best to define the theoretical domain as broadly as possible to include all the dimensions and subtleties of the construct. When the theoretical domain is well articulated, the empirical domain will be easier to operationalize and thereby aid in developing measures of the construct.

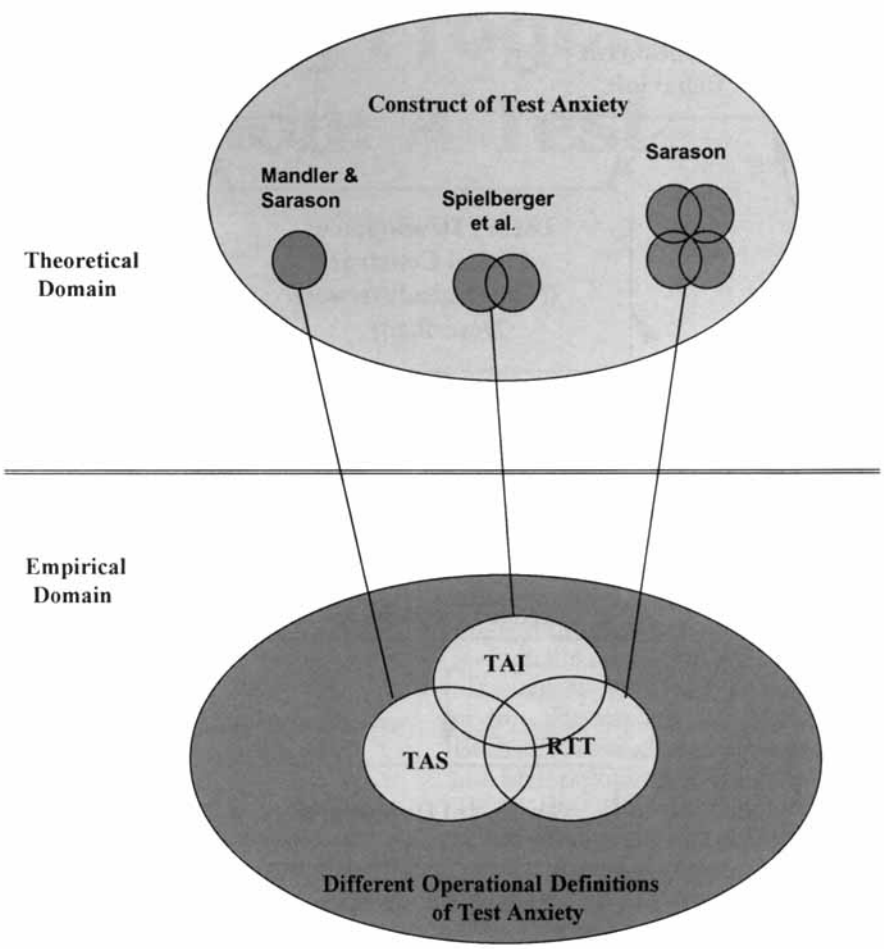

FIGURE 2. *Theoretical and empirical domains of test anxiety*

To better understand the relationship between the theoretical and empirical domains, consider the construct of test anxiety. In Figure 2, the relationship between the theoretical and empirical domains is presented. Within the empirical domain, three different operational definitions of *test anxiety* are provided. These three operationalizations of test anxiety represent three "constructions" based on different theoretical perspectives.

Initially, Mandler and Sarason (1952) envisioned a single latent dimension of test anxiety and developed the Test Anxiety Scale (TAS), which was later revised by Sarason (1978). Next, Spielberger, Gonzalez, Taylor, Algaze, and Anton (1978) questioned the unidimensionality of test anxiety and proposed a two dimensional theory as measured by the worry and emotionality components of their Test Anxiety Inventory (TAI). Later, Sarason (1984) conceptualized test anxiety as being composed of four latent dimensions. In developing his Reactions to Tests (RTT) scale, he retained the worry dimension but reconceptualized emotionality as two distinct dimensions, "persons' bodily arousal and tension" (p. 931), and added a fourth dimension of test irrelevant thinking. These three conceptualizations by no means exhaust the potential theoretical and empirical domains of test anxiety. However, they do represent the dominant themes in this literature over the last few decades and illustrate how the theoretical domain is reflected in the empirical domain. At the empirical level, each measure of test anxiety overlaps the other measures indicating they share items and/or response formats. Furthermore, this empirical overlap is a reflection of the theoretical domain where the construct of test anxiety is conceptualized currently as being multidimensional. While several of the latent dimensions of test anxiety in the theoreti-

cal domain are held in common across the three conceptualizations, they are shown in Figure 2 as distinct for clarity of presentation.

In Messick's (1995) most recent formulation of construct validation, his aspects of content representativeness and relevance could be included as part of the substantive stage. For example, content-related evidence in the form of relevance and representativeness of expert judgment ratings should be gathered and reported during this stage of validation. This evidence can help to ensure the operational definition (specific set of items and response format) adequately reflects all the aspects of the theoretical domain of the construct. An overview of the procedures for gathering content-related evidence is provided by Haynes, Richard, and Kubany (1995).

Messick (1989) has warned that at least two problems can occur during the conceptualization and definition of the theoretical and empirical domains: "construct underrepresentation," when the empirical domain is defined too narrowly, or "construct irrelevancy," when the empirical domain contains reliable variance unrelated to the focal construct (p. 34). Construct underrepresentation can lead to tests being too narrow in content coverage, which then fail to adequately represent the theoretical domain of the construct. For example, the worry dimension of test anxiety may be itself multidimensional (Hagtvet & Sharma, 1995). Construct irrelevancy indicates that the test may contain systematic variability that *does not* relate to the theoretical domain of the construct. For example, when social desirability or motivation contaminates responses to tests, the validity of the test score interpretation is affected. Sources of construct irrelevancy have the potential to impact test interpretation and use. Proposing and testing rival hypotheses is one way of studying the potential sources of construct irrelevancy using methods described in the next two sections.

An illustration of testing for construct irrelevant variance is presented in Hodapp and Benson (1997) and Hagtvet and Benson (1997). In each study, the authors used confirmatory factor analysis procedures to test a series of models attempting to define the theoretical boundary of test anxiety. Measures of constructs such as fear of failure, self-efficacy, and distraction are sometimes included within the domain of test anxiety. The two studies empirically evaluated whether these additional constructs might indeed belong within the domain of test anxiety or should be viewed as antecedents or correlates of test anxiety. Results from studies such as these help to sharpen the theoretical definition of the construct and illustrate the generalizability and boundary of a test score's meaning as described by Messick (1995).

Structural Stage: Internal Relations Among Observed Variables

The structural stage focuses on what Nunnally (1978) referred to as the *internal consistency* of the set of observed variables. Studies conducted under the structural stage are referred to as internal domain studies because they involve only the observed variables for a given test. The objective of an internal domain study is to determine the extent to which the observed variables covary among themselves, and how they covary with the intended structure of the theoretical domain. Many of the statistical methods for conducting construct validity studies are subsumed under this stage. Procedures such as intercorrelations among the items/subscales, exploratory and confirmatory factor analysis, generalizability theory, and item response theory are internal domain procedures.

Probably the most widely used method for this aspect of construct validation has been factor analysis (both exploratory and confirmatory methods). However, a limitation of both exploratory and confirmatory factor analysis is that the methods are totally internally driven. At best, both forms of factor analysis can provide information as to the dimensionality of the construct as guided by the theoretical definition. However, both forms of factor analysis provide no information about what exactly is being measured; factor analysis reveals only that some number of factors can sufficiently explain the covariation among the observed variables. Furthermore, an additional problem with both forms of factor analysis is that what holds the factors together sometimes is not the theory on which the items are based but the process by which subjects responded to the items. French (1965) has referred to this phenomenon as the Achilles' heel of factor analysis. Thus, the results from both forms of factor analysis must be carefully scrutinized to know just what they are reflecting.

Another highly useful method is generalizability theory. Originally conceived of as a method to differentiate types of errors in measurement, generalizability theory can also provide evidence of how well the empirical domain represents the theoretical domain. A very informative set of studies would be to use confirmatory factor analysis to determine how well the specific set of items fit the structure of the theoretical domain. Then use generalizability theory to determine how representative the items are of the empirical domain as well as how adequately the number of items capture the distinctive features of the theoretical domain. Benson and Hagtvet (1996) illustrated how generalizability theory can be utilized to answer questions like those just posed. An advantage of both confirmatory factor analysis and generalizability theory is they require knowledge of the theory on which the construct is based in order to appropriately implement the procedures.

A method that bridges the structural and external stages of validation is the multitrait-multimethod matrix procedure. This method is an internally focused design, in that the convergence of similar constructs can be studied, as well as an externally focused design, in that it studies whether different constructs diverge from one another. Item response theory methods are also useful in determining how well a set of items fits the theoretical structure thought to underlie a construct as well as to evaluate the consequences of test use through differential item functioning. These last four procedures are integral to a strong program of test validation.

Finally, Nunnally (1978) and Guion (1977) have suggested that, when positive results are obtained from studies conducted in the structural stage, evidence of the *necessary* condition for establishing construct validity has been made but the evidence does not meet the *sufficient* condition. When the hypothesized relations are empirically supported, we only know that "something" is being measured, but exactly what it is is still unknown. Test developers and users need to be careful not to fall prey to the nominalistic fallacy, whereby the simple naming of a construct has somehow given rise to it (Cliff, 1983). Thus, the need exists in construct validation to go further and establish not only the convergence of the observed variables with theory but also the divergence of observed variables in accordance with theory. This latter step will fulfill the sufficiency condition and is an essential part of a strong program of construct validation.

External Stage: Relations Among Constructs

Finding positive evidence for the internal structure of a test moves the focus of construct validation to its most crucial stage. Here, we are concerned that the focal construct covaries in theorized ways with different constructs and/or characteristics of the subjects. Crocker and Algina (1986) have stated "an operational definition of a construct is not enough; the meaningfulness or importance of the construct must also be made explicit through a description of how it is related to other variables" (p. 230).

The types of procedures that provide evidence under the external stage have historically been group differentiation and correlation. Group differentiation can take the form of studying the construct in the presence of existing known group differences or of creating group differences by experimental manipulation of the construct. In the first situation, the existing or known groups might be formed from self-referred students. For example, to validate the scores on a measure of test anxiety, one might administer the test to students who had referred themselves to a counseling center for help dealing with test anxiety and a group of nonreferred students. In this example, it might be hypothesized that the self-referred students would have higher means on the measure of test anxiety than the nonreferred group. In the second situation, experimental manipulation of a construct might involve providing test anxiety relaxation techniques to one group of test-anxious subjects while providing no treatment to a similar group of test-anxious subjects. After the treatment, the mean levels on the construct are expected to differ for the two groups according to a given set of theorized expectations.

The most typical procedure used in externally related construct validation studies has been to correlate a measure of the focal construct with measures of other constructs. Since we know that zero-order correlations can be influenced by measurement error as well as other variables, a statistical procedure which can account for these methodological problems is needed. While disattenuating partial correlations may correct for unreliability and the influence of other variables, we are still left with bivariate relationships of what is often a multivariate phenomenon. Therefore, multivariate methods are called for when studying a focal construct conceptualized in a nomological network of constructs.

Benson and Hagtvet (1996) have suggested structural equation modeling (SEM) as an optimal method in which to study the external stage of construct validation. SEM, developed primarily by Jöreskog (1973), is a multivariate statistical technique which combines the fields of factor analysis, path analysis, and econometric modeling. Many of the aspects of the nomological network described by Cronbach and Meehl (1955)—such as, how the observed variables relate to the theorized structure of the construct or how the different constructs involved in the theory surrounding the focal construct relate to one another—relate directly to concepts within SEM. For example, the measurement model in SEM links the observed variables to the hypothesized structure of the construct, and the structural model links the constructs within the nomological network. Thus, what Cronbach and Meehl described philosophically can now be tested empirically within the measurement and structural models of SEM.

An overview of the framework for conducting a strong program of construct validation is provided in Table 1. The statistical and concep-

Table 1
Framework for Conducting a Strong Program of Construct Validation

Substantive stage
Theory-Based (including previous research and observation)
Generate theoretical and empirical definitions
Gather content-related evidence
Consider construct underrepresentation and construct irrelevancy
Structural stage
Item/Subscale intercorrelations
Exploratory factor analysis
Confirmatory factor analysis
Generalizability theory
Multitrait-Multimethod matrix
Item response theory (including differential item functioning)
External stage
Multitrait-Multimethod matrix
Group differentiation
Existing or known groups
Experimental manipulation
Correlations of tests with other tests (including criterion-related evidence)
Structural equation modeling

tual methods used to obtain validity evidence are subsumed under each stage of validation. The framework should be viewed as a continuum versus three discrete stages, where each stage either leads to the next in building evidence for the construct validity interpretation of test scores or suggests the previous stage be reevaluated. Furthermore, a strong program of construct validation would include all three stages and more than one method within each stage.

An Illustrative Application of SEM

Cronbach (1989) portrayed the bulk of construct validity studies as *confirmationist*, rarely testing rival hypotheses. However, falsification of rival hypotheses adds as much to the understanding of the construct as confirming evidence (Popper, 1959). Rival hypotheses can be tested by specifying a priori alternative models using SEM. In fact, Jöreskog (1993) has strongly suggested the specification of alternative theoretically based models prior to testing is how SEM *should* be used.

To illustrate how SEM could be used to test rival hypotheses regarding a construct under investigation, a nomological network for the focal construct test anxiety is presented in Figure 3. Here test anxiety is thought to be composed of two situation specific response behaviors: worry and emotionality as measured by the TAI. The antecedent constructs are academic self-concept and motive to avoid failure. The outcome construct is achievement. These constructs were chosen because they have been included in recent studies of test anxiety (Benson, Bandalos, & Hutchinson, 1994; Covington, 1992; Hagtvet & Benson, 1997; Hodapp, 1989).

Based on these constructs, a theory of test anxiety is postulated where test anxiety is a two-dimensional response concept which has the potential to interfere with one's achievement. However, two personality dispositions, academic self-concept and motive to avoid failure, are thought to trigger the test anxiety responses which in turn affect achievement. Given this theoretical

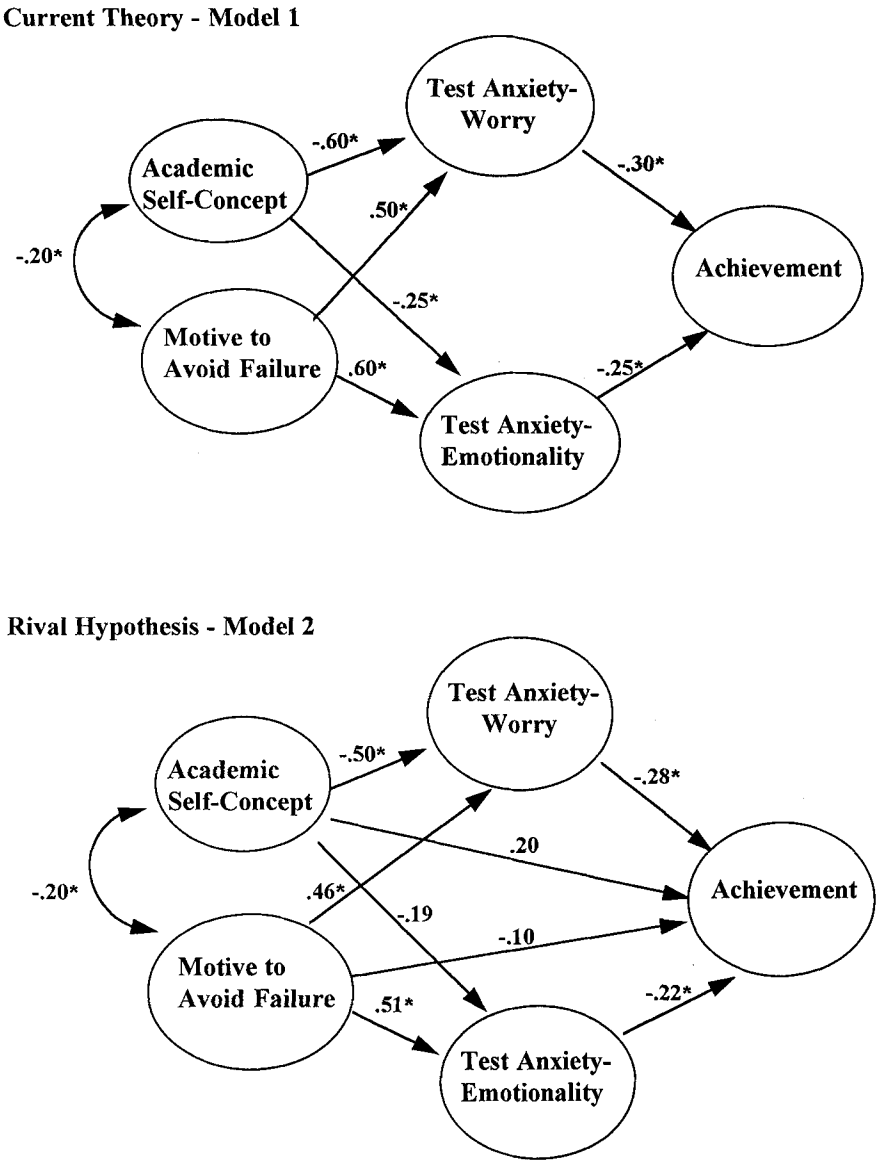

FIGURE 3. *Rival hypotheses of how test anxiety operates in a nomological network*

position, two models were generated to explain how this nomological network of relations involving the focal construct of test anxiety might be evaluated. Model 1 (in the top of Figure 3) represents our current understanding of how test anxiety operates in this nomological network of constructs. The theoretical position advanced in Model 1 is that the two dimensions of test anxiety (worry and emotionality) mediate the effect of the antecedents (academic self-concept and motive to avoid failure) on the outcome (achievement). One possible rival hypothesis is shown in Model 2 (in the bottom of Figure 3), where the antecedents *directly* and indirectly influence achievement. Thus, Model 2 challenges the confirmationist perspective often seen in construct validity studies by posing an alternative explanation of how the focal construct (test anxiety) operates in the nomological network.

The research question guiding the present illustration is: Does test anxiety mediate the influence of academic self-concept and motive to avoid failure on one's achievement, or do academic self-concept and motive to avoid failure have both direct and indirect effects on achievement? Typically, such a construct validation research question would have been answered by a series of zero-order correlations between test

anxiety and achievement and test anxiety and the antecedent variables. The research question might also have been studied via partial correlations between test anxiety and achievement, partialing out the influence of the antecedents, or between the antecedents and achievement, partialing out the influence of test anxiety. However, as stated earlier, observed correlations do not take into account the measurement error in each construct, nor can the network of relations be studied simultaneously. Thus, a multivariate approach is needed which also allows measurement error to be modeled.

For purposes of this illustration, assume that antecedent variables were measured in the beginning of a course or instructional unit and that the test anxiety measure was administered one week prior to the mid-term examination, which served as the outcome measure. Thus, the data were obtained sequentially rather than cross-sectionally. Further, we shall assume 200 examinees responded to the observed measures and the items on the observed measures were combined to form three indicators per latent variable. The measurement model has been omitted from the figure for clarity of presentation.

The analysis of the hypothetical data resulted in the following overall fit information for Model 1: $\chi^2 = 108$, $df = 85$, $p = .045$, GFI = .92, and root mean square error of approximation (RMSEA) = .08; the fit of Model 2 was $\chi^2 = 104$, $df = 83$, $p = .058$, GFI = .93, and RMSEA = .08. The RMSEA (Steiger, 1990) is becoming a popular index of model-data fit, because it assumes the model being evaluated is a good approximation to reality, not a test of perfect fit, as in the case of the chi-square statistic. Browne and Cudeck (1993) have suggested that RMSEA values of .08 or less reflect reasonable model-data fit, per *df*. While the fit of Model 1, in terms of the chi-square statistic and *p* value, reflects the fact that the model-reproduced covariances were not exactly equal to the observed covariances, the GFI and RMSEA would be considered to reflect acceptable overall fit. On the other hand, the *p* value for the chi-square statistic for Model 2 indicated the differences between the model-reproduced covariances and the observed covariances were not significantly different. Thus, Model 2 shows a slightly better fit than Model 1, based on the chi-square and *p* value, whereas the GFI and RMSEA indicated very little difference in the fit of the two models.

However, as pointed out by Jöreskog and Sörbom (1989, p. 41), the assessment of model-data fit includes evaluating the overall model fit as well as examining the parameter estimates (path coefficients, residuals, and their standard errors). Thus, when one is testing competing models based on substantive theory, the parameter estimates also must be examined to see if they are predicted by theory before a model is retained or rejected. Browne and Cudeck (1993) also have urged that models not be mechanically rejected or retained; subjective judgment also is necessary to select a model. Therefore, the selection of which model best represents a set of data is made somewhat easier when models are based on substantive theory.

The hypothetical, standardized path coefficients are presented in Figure 3 for both models and the statistically significant ($p < .05$) coefficients are noted by an asterisk. In Model 2, the two direct paths from academic self-concept and motive to avoid failure to achievement have been added. However, these paths were not found to be statistically significant. Furthermore, the fit of Model 2 did not show a substantial improvement over Model 1, either statistically in the form of a chi-square difference ($\chi^2 = 4$, $df = 2$) or practically. Model 2, while a reasonably plausible model, does not change our understanding of how test anxiety functions in the nomological network portrayed at the top of Figure 3. Therefore, Model 1 still remains tenable, given the overall fit information and parameter estimates which closely parallel what was predicted by our theory.

It appears that academic self-concept and motive to avoid failure do not directly influence achievement, but they operate through test anxiety to influence achievement as predicted by the theoretical position taken in Model 1. Thus, if Model 1 survives numerous testings of various rival hypotheses, this would constitute a body of evidence regarding the meaning of what is being measured by the test anxiety scale and, hence, the theory on which it is based. Additional rival hypotheses could be generated using a multiple-group SEM where, for example, Model 1 could be tested to determine the similarity of the path coefficients among the structural relations for males and females or over different ethnic groups. If the path coefficients and model fit were found to differ over different groups in a series of studies, the theory of test anxiety that generated the scores would need to be reevaluated. However, if the path coefficients and overall model fit do not differ significantly for the groups considered, another piece of evidence is added which indicates test anxiety operates similarly across the groups. Findings such as these relate to the generalizability across populations aspect of construct validity described by Messick (1995) and are at the heart of what Cronbach and Meehl (1955) envisioned when they suggested the testing of rival hypotheses in test validation.

Several caveats need to be made regarding the test anxiety illustration just presented. A necessary precondition to establish the validity of a construct assumes the measures of the other constructs in the network have adequate construct validity themselves. For example, if the negative path coefficient between test anxiety-emotionality and achievement turned out to be a positive association, this finding might be due to the lack of validity for the measures of either or both constructs or to the inaccuracy in the theory that links these constructs. In such a situation, there are many research options to consider. One might be that the theoretical structure of the focal construct requires further differentiation (e.g., test anxiety may be composed of more than two factors). Thus, one would undertake a revision of the TAI as a measure of test anxiety. In this option, one would need to start back with the domain definition to justify theoretically adding additional dimensions and

their content (substantive stage). Then one would need to assess the internal structure of the revised measure (structural stage) and finally to submit the revised measure to external verification using the SEM approach just described. A second option would be to select an already developed measure which is based on more theoretical dimensions (e.g., the RTT). Assuming sufficient evidence existed for the substantive and structural stages for the RTT, one could evaluate the nomological network for the construct of test anxiety using the RTT in a set of structural models similar to those shown in Figure 3. Other options include selecting different or additional measures of the antecedent or outcome constructs.

A second caveat is that there may exist alternative models that fit the data as well as the model being tested. Thus, a strong argument exists for posing and testing alternative sets of models as part of a SEM construct validation study. For example, instead of the models tested in Figure 3, a similar set of rival models could be tested where test anxiety's influence on achievement is theorized to be mediated by academic self-concept and motive to avoid failure. In this situation, the focal construct of test anxiety becomes the antecedent variable and academic self-concept and motive to avoid failure become the mediating variables. To test this set of models, the sequential collection of data would have to be altered from what was described previously. Another rival model might be tested where test anxiety is measured at the same time as the antecedents, but prior to the examination. In this situation, the interpretation of the results would differ because all the antecedents would be taken as trait measures that might be correlated, but for unknown reasons. The results across these three different sets of studies (overall fit indices and parameter estimates) would provide a collection of construct validity evidence to determine how test anxiety functions in a nomological network with other constructs. These examples clearly illustrate the central role substantive theory plays in designing a SEM study.

A third caveat relates to the substantive theory itself. In many disciplines, the substantive theory underlying a construct may not be well understood or agreed on by those in the field. An example of this was just alluded to in the last set of studies described where the role of test anxiety moved from being a moderator variable to being an antecedent variable. Jöreskog (1993) has described the three main uses of SEM as being strictly confirmatory, testing alternative models, and model generating. While the first two approaches are not as frequently found in the literature as the third, one rationale for this may be due to the state of the art in theory building. Certainly SEM and construct validation should be guided by theory. However, where theory is unclear or insufficient, a model generating approach can be used to test relationships in an attempt to clarify controversial aspects of the theory or to build new aspects of the theory. In the model generating mode, the researcher is guided by what is currently known or understood to respecify parts of the model and then can use features of the SEM software (residuals, modification indices, expected parameter change) to suggest where the model may be modified. The resulting model(s) generated from such an approach would clearly need to be cross-validated using a strictly confirmatory approach or by testing alternative models as described previously.

In sum, it is hoped that the proposed framework will be useful in thinking about the stages involved in developing a strong program of construct validation research. The proposed framework pulls together the various statistical methods used in construct validation research into an organized whole. Such a framework is intended to provide guidance to researchers who wish to begin a program of construct validation or to evaluate the validation research literature regarding a construct of interest. Furthermore, the use of SEM in the context of construct validation research is strongly suggested as a method to evaluate nomological networks that are an integral part of construct validation.

Notes

An earlier version of this article was delivered as the Presidential Address to the Society for Stress and Anxiety Research, Prague, Czech Republic, July 1995. The author would like to thank the former editor, Linda Crocker, and several anonymous reviewers for their comments and insights which have helped to shape and strengthen this article.

References

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, (2, Pt.2).

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Benson, J., Bandalos, D., & Hutchinson, S. (1994). Modeling test anxiety among men and women. *Anxiety, Stress and Coping, 7*, 131–148.

Benson, J., & Hagtvet, K. (1996). The interplay between design, data analysis and theory in the measurement of coping. In M. Zeidner & N. Endler (Eds.), *Handbook of coping: theory, research, applications* (pp. 83–106). New York: Wiley.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research, 13*, 115–126.

Covington, M. (1992). *Making the grade: A self-worth perspective on motivation and school reform*. New York: Cambridge University Press.

(Continued on page 22)

Haertel, E. H., & Linn, R. L. (1996). Comparability. In Gary W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59–78). Washington, DC: U. S. Government Printing Office, National Center for Education Statistics.

Khattri, N., Reeve, A. L., Kane, M. B., & Adamson, R. J. (1996). *Studies of education reform: Assessment of student performance: Final report, Volume 1: Findings and conclusions*. Washington, DC: Pelavin Research Institute.

Lambdin, D. V. (1995). An open-and-shut case? Openness in the assessment process. *The Mathematics teacher, 88*(8), 680–684.

Linn, R. L., Baker, E. I., & Dunbar, S. B. (1992). Complex, performance-based assessment: Expectations and validation criteria. *Evaluation Comment*, 2–9.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11(1)*, 3–9, 20.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8*(1), 14–22.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1996). Validity of performance assessments. In Gary W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: U. S. Government Printing Office, National Center for Education Statistics.

National Council of Teachers of Mathematics. (1995). *Assessment Standards for School Mathematics*. Reston, MA: The Council.

Nolen, S. B., Haladyna, T. M., & Hass, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice, 11*(2), 9–15.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice, 10*(4), 12–15.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215–232.

Shepard, L., (1991). Effects of high-stakes tests. *Phi Delta Kappan, 73*(3), 243–247.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46*, 41–47. ■

A Test Anxiety Example

(Continued from page 17)

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart & Winston.

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

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

Cronbach, L. (1989). Construct validation after thirty years. In R. Linn (Ed.), Intelligence: Measurement, theory, and public policy. *Proceedings of a Symposium in Honor of Lloyd Humphreys* (pp.147–167). Urbana, IL: University of Chicago Press.

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

French, J. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement, 25*, 9–28.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1*, 1–10.

Hagtvet, K., & Benson, J. (1997). The motive to avoid failure and test anxiety responses: Empirical support for integration of two research traditions. *Anxiety, Stress, and Coping, 10*, 35–57.

Hagtvet, K., & Sharma, S. (1995). The distinction between self- and other-related failure outcome expectancies: An internal domain study of Indian and Norwegian Students. In A. Oosterwegel & R. Wicklund (Eds.), *The self in European and North American culture: development and process* (pp. 239–255). Boston: Kluwer.

Haynes, S., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*, 238–247.

Hodapp, V. (1989). Anxiety, fear of failure, and achievement: Two path-analytic models. *Anxiety Research, 2*, 301–312.

Hodapp, V., & Benson, J. (1997). The multidimensionality of test anxiety: A test of different models. *Anxiety, Stress, and Coping, 10*, 219–244.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. Goldberger & D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Academic.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: Statistical Package for the Social Sciences.

Jöreskog, K. G. (1993). Testing structural equation models. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

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

Mandler, G., & Sarason, S. (1952). A study of anxiety and learning. *Journal of Abnormal and Social Psychology, 47*, 166–173.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). Washington, DC: American Council on Education.

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.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Sarason, I. (1978). The Test Anxiety Scale: Concept and research. In C. Spielberger & I. Sarason (Eds.), *Stress and anxiety* (Vol. 5, pp. 193–216). New York: Hemisphere/ Wiley.

Sarason, I. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46*, 929–938.

Spielberger, C., Gonzalez, H., Taylor, C., Algaze, B., & Anton, W. (1978). Examination, stress and test anxiety. In C. Spielberger & I. Sarason (Eds.), *Stress and anxiety* (Vol. 5, pp. 167–191). New York: Hemisphere/ Wiley.

Steiger, J. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180.