Running head: TEST VALIDITY

Test Validation: A Literature Review

Jeffrey M. Miller

University of Florida

Jeffrey M. Miller

Table of Contents

Origins

Campbell

One of the earliest distinctions between types of validity is Campbell's (1957) notion of internal and external validity. Internal validity asks the question – "Does the test measure what is purports to measure?" (American Psychological Association, 1954). This view in which validity is inherent within the test itself has shifted drastically to one demanding external evidence due to the seminal work of Samuel Messick (Suen, 1990) External validity, which can be defined as the degree to which test scores can be generalized to other populations, settings, times or events, can be traced to Campbell's distinction of this form of validity from validity internal to an experiment (Albright & Malloy 2000).

External validity is often used synonymously with the term "generalizability". Campbell advocated random assignment and the use of control groups to ensure internal validity as well as MTMM techniques to ensure external validity (Albright & Malloy, 2000).

Cook & Campbell

In 1979, Cook joined Campbell and expanded the list of validities to four (Cook & Campbell, 1979). Internal validity remained the most important type. This was followed by construct validity, or the generalization from the experimental operations to the hypothesized constructs. Next, they added statistical conclusion validity, which involved the validity of the inferences made using data from the experiment. Finally, external validity remained a crucial part of the validation process. Much has been written about these and other related aspects of validity as discussed by Campbell independently and in collaboration with Cook; however, the scope of this paper is limited to validity in testing. Readers interested in experimental validity should consult Cook and Campbell (1979), Dooley (2001), and Albright & Malloy (2000). Readers interested in a more detailed history of defining validity should consult Angoff (1988),

The Many Faces of Validity

The Holy Trinity

The first attempt to demarcate types of validity in testing standards resulted in four specific types (American Psychological Association, 1954). These were content validity, predictive validity, concurrent validity, and construct validity. Content validity refers to the degree to which the content of the items reflects the content domain of interest. Predictive validity refers to the degree to which scores predict future performance. Concurrent validity refers to the degree to which scores relate to scores obtained from other measures. Construct validity refers to the degree to which scores reflect the unobservable trait of interest.

As far back as the APA Technical Recommendations of 1954 (American Psychological Association), predictive and concurrent validity have been viewed as one criterion-related validity. The resulting three types of validity inspired Guion’s (1980) reference to a holy trinity of validity – content validity, criterion-related validity, and construct validity. Cronbach (1984) considered the three types to be different methods of inquiry reflecting a unitarian paradigm to be expanded upon by Messick in 1989.

Criterion-Related Validity

Both predictive validity and concurrent validity refer to the correlation between test scores and some criterion. For predictive validity, this correlation is between the test score and a score on some future performance. Ultimately, such tests are used to make decisions (Nunnally, 1978). For concurrent validity, the correlation is between the test score and some current performance on either a substitute for the current test (e.g., computer-based versus pencil-and-paper administration) or some related criterion (e.g., SAT scores versus GPA) (Cronbach & Meehl, 1955).

By far, the most common means of validating a test has been the calculation of a correlation coefficient between the test score and some external criterion. Nunnally (1978) explains that the interpretation of this coefficient is the “possible improvement in the average quality of persons that would be obtained by employing the instrument in question (p. 91).” Shepard (1993) explains that the use of such a predictive correlation coefficient was so prevalent that its use became synonymous with the term validity. This was supported by Guilford’s (1946) infamous quote that “a test is valid for anything with which it correlates (p.429)”. However, Shepard (1993) argues that a single correlation is insufficient.

Four points are worth noting. First, the inference from the test to the criterion may be a function of the mode of assessment. Second, there is the possibility that both the test and the criterion are similarly biased. Cronbach & Meehl (1955) explain “any distinction between the merit of the test and criterion variables would be justified only if it had already been shown that… [both] were excellent measures of the attribute (p. 285)”.

Further, the items, although perhaps reliable, may not reflect the content universe of interest (Suen, 1990).

Third, validity seeks some degree of causation; however, it well known that correlation coefficients do not necessary imply a causal connection. “It is ironic that a field so attuned to the fallacy of mistaking correlation for causation in experimental contexts would be willing to accept correlations in the measurement sphere as immediate proof of test validity (Shepard, 1993, p.411).”

Fourth, the variance of the correlation coefficient is a function of sampling characteristics making it sensitive to restrictions in range. Validity has been quantitatively defined in terms of variance as “the proportion of observed variance that is relevant to the purposes of testing (Brown, 1982, pp. 67-68)”. Hence, the correlation between SAT scores and first year college grades would vary as a function of a college’s disclosure of a minimum grade point average required for consideration of admission.

In line with Shepard’s second and third arguments regarding criterion-related validity is the desirability of multiple sources of evidence. Cronbach & Meehl (1955) discuss this in terms of the dual invalidity of the test and criterion; however, their explanation suggests a potential invalidity of linkage between the test and criterion. For example, are creative thinkers likely to be ambitious graduate students? Submitted poetry could be evaluated and correlated with a measure of ambition. Suppose that a student submitted a haiku and received a low score but received a high score on a multiple-choice measure of ambition. Can it be determined that writers low in creativity are more ambitious? Certainly not if the essay evaluator considered haikus to be a rigid writing style, focusing on the form rather than the content. However, if several other measures of creativity were included and the correlation remained low or negative, then the argument that writers low in creativity are less ambitious would be substantiated. Better yet, several measures of ambition would be included in the analysis.

Regardless, the use of correlation coefficients to determine predictive validity is omnipresent. Hence, interpreters are forewarned to be realistic in their interpretation of the coefficient. “In most prediction problems, it is reasonable to expect only modest correlations between a criterion and either an individual predictor test or a combination of predictor tests. People are far too complex to permit a highly accurate estimate of their proficiency in most performance-related situations from any practicable collection of test materials. (Nunnally, 1978, pp.90-91)”.

Content Validity

Haynes, Richard, and Kubany (1995) define content validity as “the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose (p. 238)”. They mention three hindrances to successful content qualification in psychological assessment. Extending these to educational measurement, they are 1.) the zero presence of items representing a defined facet of the construct; 2.) the presence of item measuring irrelevant facets; and, 3.) disproportionate weighting of items in deriving an aggregate score.

Interestingly, content validity addresses features of the test, not the scores. In fact, content validation often occurs before scores are even obtained. “Content validity deals with inferences about test construction; construct validity involves inferences about test *scores*. Since by definition, all validity is the accuracy of inferences about test scores that which has been called “content validity” is not validity at all. (Tenopyr, 1977, p.50).”

Hence, some would argue (e.g., Messick, 1975) that content validity not even be included in test validation while others (e.g., Yalow & Popham, 1983) concur that content validation is “a necessary precursor to drawing reasonable inferences based on test scores (Shepard, 1993, p.415)”. They would argue that the test score is evidence-in-waiting. “If a test is constructed so that it constitutes a representative sample of the domain of interest, then we expect that an examinee’s score on the test reflects how the examinee will perform on the domain of interest (Yalow & Popham, 1983, p.11)”.

One problem with expanding validity beyond score interpretation toward the inclusion of test content is the “foot-in-the-door” dilemma. Once the content becomes a necessary condition of validity, then the argument can be made that whether or not the content was instructed by the teacher is now in the scope of validity (i.e., instructional validity). Further, once the instruction by the teacher becomes part of the scope of validity, then whether or not the content and instruction are included in the curriculum becomes part of the scope of validity (i.e., curricular validity). Hence, two other validity terms enter the literature. First, curricular validity has been defined by McClung (1978) as “a measure of how well test items represent the objectives of the curriculum (p. K-3)”. Second, instructional validity (1978) has been defined by McClung (1978) as “an actual measure of whether the schools are providing students with instruction in the knowledge and skills measured by the test (p. K-4).”

An Aside on Instructional Validity, Curricular Validity, and Debra P.

The necessity for these two new aspects of validity arose due to political circumstances when a class-action suit, initiated by the parents of Debra P., was filed with the U. S. Fifth Circuit Court of Appeals. Debra P. had been denied high school graduation; however, the argument was made that Debra P. did not have the opportunity to learn the curriculum from which scores were obtained to make graduation decisions. The Supreme Court ruled that sufficient instructional preparation is a necessary ingredient in determining validity; hence, curriculum and instructional validity are tied to content validity (Yalow & Popham, 1983).

Given that people can conjure precursors to the validity of curriculum and instruction hence perpetuating the “foot-in-the-door” dilemma, it is not surprising that the most current standards maintain a focus on the interpretation of scores alone. The argument has been made that such adequacy-of-preparation issues are in the realm of education, not psychometrics (Yalow & Popham, 1983).

Certainly, uses and consequences of the interpretations are currently accepted as an important ingredient of validity. Perhaps, the Debra P. case should be seen not as a movement towards an idealistic situation where every student learns every objective, but rather a movement towards individual consideration of the interpretation of test scores and their resulting consequences on high school graduation.

Content Validation Techniques

As mentioned by Linda Crocker in her presidential address to the American Educational Research Association (2003), procedures for content validation may be the most neglected aspect in test documentation. This tendency to neglect documentation of content validation procedures was discussed almost fifteen years earlier in Messick’s (1989) treatise on validity. Indeed, Yalow and Popham’s (1983) prophecy may have been fulfilled: “We fear that efforts to withdraw the legitimacy of content representativeness as a form of validity may, in time, substantially reduce attention to the import of content coverage (p. 11)." The reluctance to address content validity is surprising when considering the impact of such cases as Debra P in political legislation. It is clear, at least politically, that scores are linked to content as well as the curriculum (Crocker, 2003). And from the a testing perspective, she states, "content representation is the only aspect of validation that can be completed prior to administering the test and reporting results. If this process yields disappointing results, there is still time to recoup."

The standard protocol usually involves some team of content experts to rate and judge the match between an item and the content domain of interest (Dunn, Bouffard, & Rogers, 1999; Hambleton, 1980). Crocker and Algina (1986) outline the following steps for content validation: "1.) Defining the performance domain of interest; 2.) Selecting a panel of qualified experts in the content domain; 3.) Providing a structured framework for the process of matching items to the performance domain; and, 4.) Collecting and summarizing the data from the matching process. (p. 218)".

Often, the panel of experts is a group of professionals familiar with the content domain. For example, a group of algebra teachers may individually rate the quality of an item using a Likert scale followed by the calculation of an average rating that determines the fate of the item. Often the process of matching items to the performance domain is driven by a table of specifications; however, Crocker and Algina (1986) emphasize that such matching does ensure validity. Similar to many other aspects of measurement (e.g., reliability), the table of specifications can be viewed as a necessary but not sufficient condition of content validation.

Haynes, Richard, & Kubany (1995) outline 35 steps to conducting a detailed content validation. For the sake of space, the authors summarized seven of the guidelines.

First, similar to Crocker and Algina (1986), the domain needs to be defined; however, they also included the importance of defining the facets of the construct. Second, all elements of the measure should be validated (e.g., instructions, response format, scales). Third, experts generate the initial items. Fourth, “Use multiple judges using formalized scaling procedures (p. 244)”. This becomes especially important in determination of accuracy discussed earlier in this paper. Fifth, the proportional distribution and weighting of items should be examined. Sixth, if the instrument is new then the results of the content validation should be published. Finally, seventh, conduct additional analyses such as item analysis and factor structure.

Hambleton (1980) provides the following twelve steps for test development and validation: “1. Objectives or domain specification must be prepared or selected before the test development process can begin; 2. Test specifications are needed to clarify the test’s purposes, desirable test item formats, number of test items, instructions to item writers, etc.; 3. Test items are written to measure the objectives included in a test (or tests, if parallel forms are required); 4. Initial editing of test items is completed by the individuals writing them; 5. A systematic assessment of items prepared in steps 2 and 3 is conducted to determine their match to the objectives they were written to measure and to determine their “representativeness”; 6. Based on the data from step 5, additional item editing is done. Also, test items are discarded that do not at least adequately measure the objectives they were written to measure; 7. The test(s) is assembled; 8. A method for setting standards to interpret examinee performance is selected and implemented; 9. the test(s) is administered; 10. Data addressing reliability, validity, and norms are collected and analyzed; 11. A user’s manual and technical manual are prepared; 12. A final step is included to reinforce the point that it is necessary, in an on-going way, to be compiling technical data on the test items and tests as they are used in different situations with different examinee population (pp. 81-82)".

A Note on Face Validity

The content validation methods described above by Crocker and Algina (1986), Haynes, et al (1995), and Hambleton (1980) emphasize a predominantly qualitative process. In fact, Brown (1986) argued that a rational and judgmental process is necessary. Such a paradigm may lead many to confuse content validity with face validity (Brown, 1986), which is defined as "when the items look like they measure what they are supposed to measure (Friedenberg, 1995, p.251)". Although this author delineates the two aspects of validity by the fact that content validity involves experts, it appears that even the distinction between an expert and a non-expert is a qualitative and perhaps arbitrary decision.

Face validity is usually treated as unwarranted and similar to using anecdotal evidence in a professional research publication. This is because face validity concerns how the test appears to look (Anastasi, 1988). Test validation reflecting face validity most often occurs when no formal process of content validation occurs (Suen, 1990). Bornstein (1996) went so far as to criticize Messick's (1995) model for not including it as an aspect of validity. Bornstein argues a difference between content and face validity based on how the appearance of the test alters responses (e.g. self-presentation and self-report biases). Bornstein (1996) states, "It is ironic that although the original (1954) version of the American Psychological Association's *Technical Recommendations for Psychological Tests and Diagnostic Techniques* included an extensive discussion of face validity, the 1966 and 1974 versions of this document included only brief discussions of this issue. The most recent (1985) version of the APA's *Standards for Educational and Psychological Testing* included no mention of face validity. Face validity not only influences test-takers' motivations and test-taking strategies before and during testing, but also affects the degree to which construct-irrelevance variance contaminates respondents' test scores in a variety of situations and setting (p.984).

Accountability for Content Validation

Although initial content validation may include the views of experts, there are quantitative techniques to be presented. Regardless, the process of content validation is rarely described beyond the mentioning of the use of the experts (Crocker, 2003). In Crocker's (2003) presidential address to the National Council on Measurement in Education, she pleas for better reporting of procedures and specifies the following areas of content validation that need more research: 1. the design of test specifications and item generation; 2. item review tasks; 3. subject matter expert reliability; 4. data analysis techniques; and 5. ways to improve efficiency and reduce costs.

The next section addresses quantitative indices of validity. However, it should be noted that there is not one number that can be used to determine the validity of a test (Suen, 1990). Rather, support for validity is gained through indices and other forms of judgment from multiple sources that add to or take away from the degree of validity.

Quantitative Techniques in Content Validation

Katz's proportion

The most basic technique for content validation is to sum the number of items that judges match to the objectives and divide by the product of the number of items and the number of judges. This technique proposed by (Katz, 1957) suffers from sensitivity to the number of items, ignored item weighting, ignored degrees of expert certainty, and the phrasing of the content validity question (Crocker, Miller, & Franks, 1989)

Brown's pretest-posttest

Brown (1986), aware of the judgmental nature of content validation, proposed a pretest-posttest design in which scores are validated by increasing significantly from the pretest through training to the posttest. This technique is limited by the necessity of examinees having no prior knowledge at pretest. He also proposed correlating the scores with those on other tests with content covering the same domain. The obvious limitation here is identifying and collecting data from two tests that measure exactly the same domain. Further, there is the potential that both tests will correlate highly but measure the wrong domain (Nunnally, 1978).

Likert Scales

Item rating

Content experts sometimes rate the degree to which the item fits the domain in such dimensions as relevance, representativeness, specificity, and clarity (Haynes, Richard, & Kubany, 1995). This can be quantified using a Likert-scale rating sheet resulting in scores than can be subjected to descriptive and inferential statistical analyses such as interrater agreement (Brown, 1986; Friedenberg, 1995).

Klein and Kosecoff's Objective to Item Correlation

Klein and Kosecoff (1975) proposed an index that not only includes information regarding the degree of fit between the item and the objectives but information regarding the objectives as well. Using this procedure, the importance of the objectives is obtained from experts using a 5-point Likert scale. Then, the experts decide whether the items fit the objectives. The proportion of judges who match an item to an objective are summed for all items for an objective which can then be correlated with the mean or median Likert-scale rating for that objective. Crocker, Miller, & Franks (1989) caution that this index is influenced by the number of items for a particular objective, the number of objectives, sensitivity to the number of judges, and the wording of instructions to judges. Further, they caution that if an item does not match an objective at all then the statistics for that item will not influence the correlation.

Morris and Fitz-Gibbon’s Relevance Index

Morris and Fitz-Gibbon (1978) devised an index similar to Klein and Kosecoff’s Objective To Item correlation; however, an additional 2-point Likert scale is included to measure item format suitability as well as another 2-point Likert scale measuring object appropriateness. Calculations based on these measures result in an index of coverage, an index of relevance, a grand average, and a decision statistic based on the comparison of the first three indices. This approach is limited by calculations using scores measuring different constructs (e.g., importance, suitability) as well as variation due to differences in the Likert scales (e.g., 5-point vs. 2-point).

Limitations

One problem with content validation techniques using Likert scales is the degree of measurement error in the experts’ agreement. Crocker, Llabre, and Miller (1988) applied generalizability theory to content validity ratings and suggested that variability will be affected by the particular research question (e.g., ratings should be more similar for a spelling test than a geography test (Nunnally, 1978)), the population (e.g., local vs.

state), ad content breadth (e.g., narrow vs. broad). Such test-exclusive features will affect the level of rater agreement and, subsequently, the desirable number of raters. This problem is also faced in correlation analysis; however, the researcher can provide measures of absolute accuracy (e.g., standard errors) and relative accuracy (e.g., squared multiple correlation coefficients) (Agresti & Finlay, 1997; Crocker & Algina, 1986).

Crocker, Llabre, and Miller (1988) addressed the nature of variation in content ratings. After mentioning several authors of content validation procedures, “none of these procedures allows the test user to systematically isolate sources of variation that may influence the content experts’ ratings (p.287)”. It is commonly known that a distribution reaches normality as the sample size increases (viz., the central limit theorem) and that measures of dispersions (e.g., standard deviations) should be included when reporting measures of central tendency (e.g., means). Haynes, Richard, & Kubany (1995) state that “confidence in the robustness of the ratings (the standard error of measurement) will increase as the number of judges increases. In addition, increasing the number of raters (e.g., more than five) facilitates the detection and exclusion of rater outliers. (p. 244)”. In fact, despite the arbitrary nature of mandating a specific number of raters (Crocker, Llabre, & Miller, 1988), Lynn (1986) argued that five or more raters should control for rater agreement by chance alone. Unfortunately, the ease of procuring a large enough number of experts to ensure a trustworthy standard error of measurement is not always feasible. Further, such estimates of sample size should be extended to the number of Likert-scale categories which immediately suggests poor confidence intervals when the number of categories is less than or equal to five options.

Penfield’s Score Interval

Penfield (2004) developed a confidence interval that is not constrained by large numbers of raters or Likert-scale categories. The Score interval outperforms the traditional Wald interval (viz., the mean +/- (t * the standard error)) especially when there are five or fewer categories and twenty or fewer raters. Hence, herein may lie the solution of obtaining a more accurate estimate of content validity when the number of raters is small (i.e., less than 20) and/or the number of Likert-scale categories is small (i.e., less than 5) (Penfield & Miller, in press).

Rovinelli and Hambleton's Indice of Item Congruence

The method prescribed by Rovinelli and Hambleton (1977) results in indices of item congruence in which experts rate the match between an item and several constructs assuming that the item taps only one of the constructs which is unbeknownst to the experts. The permitted ratings are 1 (viz., the item clearly taps the construct), 0 (viz., the expert is unsure whether or not the item taps the construct, and –1 (viz., the item clearly does not tap the construct). The result of their equation is an index on a scale of –1.0 to +1.0 in which +1.0 can be interpreted as all experts agreeing that the item clearly taps the construct. Unfortunately, this index does not provide a uniform cut-off dictating when to retain and when to discard or modify an item. Hambleton (1980) suggests the following strategy: "In my work, when I feel it desirable to set a cut-off score, I create the poorest set of content specialists' ratings that I would be willing to accepts as evidence that an item was in the domain of interest. The value of the index for this set of minimally acceptable ratings serves as the cut-off score for judging the item-objective match of each of the test items. (p. 89)". In other words, an index is computed for the minimally accepted rating and is then used as the cut-off score. Hambleton (1980) also mentions the drawback of time that must be invested into computing an index for each item. However, Turner, Mulvenon, Thomas, and Balkin (2003) have designed SAS computer programs to eliminate this problem.

Aiken's V

One index of validity that can be used with Penfield's (2003) Score interval is Aiken's (1985) V coefficient that was initially designed for small samples. Simply put, raters evaluate an item using a Likert-scale. For each rating, s is computed which is the difference between the rating and the lowest rating possible. The s for each rater is summed to produce S which is then divided by n*(c-1) where n is the number of raters and c is the number of integers in the rating scale. For example, an item on a basic algebra exam may read, "Which of the following best illustrates the distributive property of mathematics?" Three raters determine that the item moderately taps algebraic thinking and select "3" on a 5-point Likert-scale, and one rater determines that the item poorly taps algebraic thinking and selects "1" on a 5-point Likert-scale. The s for the first three raters is 2 (i.e., 3-1), and the s for the fourth rater is 0 (i.e., 1-1). Summation results in S = 2 + 2 + 2 + 1 = 7. This is divided by 4*(5-1) resulting in an Aiken's V between 0 and 1.0 which, in this case, is .438. Using Aiken's published table, the statistical significance of this coefficient can be determined. "When the sample of items or raters is large,…, the central limit theorem can be applied to determine the statistical significance of the mean value of V (Aiken, 1985, p.135)." Further, Aiken's V can be interpreted descriptively since the coefficient will be 1.0 only when all raters agree to the highest possible rating for the item.

Lawshe’s Content Validity Ratio

Another index of content validity has been called the content validity ratio or CVR (Lawshe, 1975). This index is useful when the test is to be used for decisions such as job placement or school acceptance. The CVR is equal to the difference between the number of raters indicating the success on the item as essential and the number of raters, which is then divided by the number of raters. The resulting CVR for the item can then be compared to a probability table (provided by the author) displaying the minimum CVR necessary given the number of raters to achieve significance at the one-tailed .05 level. However, Crocker, Miller, & Franks (1989) note that the CVR is merely a linear transformation of Katz’s proportion.

Construct Validity

Construct Ontology

“A construct is some postulated attribute of people, assumed to be reflected in test performance (Cronbach & Meehl, 1955, p. 283).”Often is the case in both psychological and educational research that no universe of construct and no criterion are available (Cronbach & Meehl, 1955). Construct validity addresses the degree to which scores represent the unobservable trait operationalized through test items. However, there are situations where a construct is operationalized after using the test for criterion-related purposes. One example is the use of intelligence tests, having had the original intentions of identifying special needs children but now being used to classify all people as having a certain quantity of this hypothesized ‘intelligence’.

In one sense, the entire notion of constructs is contradictory to a scientific view of psychology. From a positivistic standpoint, the postulation of such a commonly researched and applied construct as intelligence is absurd. What is relevant is that the operationalisation of a construct is, indeed, based on observable behaviors.

MacCorquodale and Meehl (1948) explain this as a hypothetical process in which the construct is defined in terms of expressed experiences. “The ultimate ‘reality’ of the world in general is not an issue here; the point is merely that the reality of hypothetical constructs like the atom, from the standpoint of their logical relation to grounds, is not essentially different from that attributed to stones, chairs, other people, and the like. When we say that hypothetical constructs involve the notion of ‘objective existence’ of actual processes and entities within the organism, we mean the same sort of objective existence, defined by the same ordinary criteria that is meant when we talk about the objective existence of Singapore. (MacCorquodale & Meehl, 1948, p.107).” In test validation, one must be content to set ontological debates aside, especially when a test is necessary not to confirm or disconfirm the existence of “algebraic reasoning” but to use observable scores to place a student into an appropriate algebra course.

The Nomological Net

Cronbach and Meehl (1955) explain that “to make clear what something is means to set forth the laws in which it occurs (p. 213)”. They conceptualized the nomological (viz. law-like) net in which the construct “is defined by a network of associations or propositions in which it occurs [with validation] possible only when some of the statements in the network lead to predicted relations among observables [based on] many types of evidence. High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct (1955, pp. 299-300)”. However, such notions that the relationship between constructs is hard-wired have been abandoned by most theorists. Shepard (1993) explains that “it is less tenable for social scientists to assume either that human behavior is governed by laws, akin to the laws of physics, waiting to be uncovered or that constructs and observables can be meaningfully separated given that observation occurs through an interpretive lens (p. 417)”.

Despite the qualitative reasonableness of this argument, it seems worthwhile to further discuss Cronbach and Meehl’s nomological net on the grounds that 1.) the net provides a framework for describing and testing constructs in an organized manner and 2.) the results of tests can be based on probabilistic decisions (i.e., traditional hypothesis testing). Cronbach and Meehl admit to this latter point when stating that, “a rigorous (though perhaps probabilistic) chain of inference is required to establish a test as a measure of a construct (p. 293)”. Further, since the nomological net is a theoretical structure, no permanence in mandate and structure is necessary. Rather, researchers of test validity should be aware of the possible influence of human interpretation in the development of the net, as they should be with the interpretation of test scores in the absence of the nomological net. The authors support this line of thought; “We do not worry about such advanced formal questions as “whether all molar behavior statements are decidable by appeal to the postulates” because we know that no existing theoretical network suffices to predict even the known descriptive laws. Nevertheless, the sketch of a network is there; if it were not we would not be saying *anything* intelligible about our constructs…the vague, avowedly incomplete network still gives the constructs whatever meaning they do have. (p. 294)”. However, Cronbach and Meehl’s inclusion of “acceptance” in which several investigators must agree or disagree to the tenets of the net may no longer be a practical argument given the current emphasis on test use and consequences.

Prediction of the role of a construct is tested by hypothesizing a certain amount of test variance accounted for by the construct in the network. Testable propositions include the relationship between the test and construct, between the construct and other constructs, and between constructs and observable variables. The near infinite number of experiments that can be held introduces a non-utilitarian aspect into the theory; test validators, in line with Shepard's (1993) suggestions, should base their experiments on the test's uses and consequences.

Cronbach and Meehl (1955) provide explanations for disconfirming results. They suggest three possibilities: "1.) the test does not measure the construct variable, 2.) the theoretical network which generated the hypothesis is incorrect, and 3.) the experimental design failed to test the hypothesis properly. (p. 296). The first possibility is well worth considering in high-stakes decisions. However, for such high-stakes decisions where a considerate amount of time and money has been invested in test construction, it may be worthwhile to consider reexamining the network, gathering new data, and improving the experimental design before throwing out the proverbial baby with the bath water.

Construct-ion

Shepard (1993) delineates four features of construct validation. First, construct validity includes both an internal and external structure. Internally, test validators should identify and describe, "subdomains or subconstructs, the expected interrelationships among dimensions of the construct, and the processes believed to underlie test performance (pp. 417-418)". Techniques include correlation analysis and item response theory. Externally, test validators should identify the relationship between the construct(s) and other constructs.

Second, test validators may use both correlational evidence and experimental manipulations. Third, test validators should not only seek the confirmatory evidence but also identify plausible rival hypotheses. This follows from the earlier discussion of the utility of falsification and null hypothesis testing. Indeed, it is peculiar that test validators who have often had significant training in experimental methods emphasizing falsification fail to apply such techniques to test validation. Possibly, later revisions to the Standards definition should include not only a confirmatory degree of theory and evidence but also a disconfirmatory degree of theory and evidence.

Finally, and similar to the call for confirmatory and disconfirmatory evidence, is the use of both convergent and discriminant validity evidence. One common technique is Campbell and Fiske's (1959) multitrait-multimethod (MTMM) matrix in which several constructs are measured using several methods. The correlation of one construct between methods should be higher than the correlation between separate constructs using the same method. Although MTMM, provides an objective matrix of correlation coefficients, there is no set scheme of determining when a correlation coefficient is high enough or too low. Shepard (1993) explains, "the interpretation of results cannot be reduced to a set of algorithmic decisions (e.g., the principal component of the convergent correlations is *not* the true construct). (p.423)".

Construct Validation Techniques

Group Differences

If groups should differ on a construct, then tests between groups should signify this difference (Cronbach & Meehl, 1955). For example, mentally healthy elderly and adolescent populations should differ on the construct 'impulsivity'. If a measure of compulsivity yields a reliable correlation of .93, then the measure should be investigated for potential invalidity.

MTMM

The multitrait-multimethod (MTMM) technique was discussed previously in the process of establishing constructs. However, it is not limited to a priori analysis; MTMM can be used as a post hoc measure of convergent and discriminant validity to ensure the valid interpretation of test scores. For more discussion of MTMM and its application, consult Campbell (1960) and Campbell and Fiske (1959)

Factor Analysis

The purpose of factor analysis is to reduce the items to a set of independent factors that then provide information as to how many constructs are contained in the test. (Suen, 1990). "A factor analysis provided three types of information: (1) how many factors are needed to account for the intercorrelations among the tests, (2) what factors determine performance on each test, and (3) what proportion of the variance in the test scores is accounted for by these factors (Brown, 1986, p.141)." A factor analysis of a measure may support the intended number of factors and their interdependence. On the other hand, "a factor analysis might establish that the items can be subdivided into several subscales but that the initial pool does not contain enough items to assess each of these content domains reliably (Clark & Watson, 1995)".

Internal Structure

At a purely technical level, measures of internal structure such as KR-20 and coefficient alpha are the subject of reliability. However, if test measures a particular construct, then the items expected to be related to the construct should correlate highly with each other (Cronbach & Meehl, 1955); irrelevant items (i.e., suppressor variables) should concurrently correlate negatively and/or weakly with these items (Cronbach, 1945). The weakness of such analyses is that they do not include any relationship between the scores of the test and the scores of other variables (Brown, 1986).

Change Over Occasions

Cronbach & Meehl (1955) discuss their extension of test-retest reliability to validity. For example, does the GRE Verbal assessment adequately measure reading comprehension? Suppose a student receives a score of 300 on one occasion. Following testing, suppose the student pays for coaching in order to improve the score. Suppose training includes identifying key features in the paragraph most likely to appear in the questions as well as techniques for discarding questions based on knowledge of how the test is constructed. Suppose a second administration of the exam yields a score of 420. Now, does the GRE Verbal assessment include a measure of reading comprehension or a measure of the ability to successfully perform on this mode of assessment? This is certainly a question of construct validity.

Studies of Process

Certainly, a measurement of a construct should be representative of the processes utilized in the performance of the behavior intended to represent the construct. For example, suppose one wishes to measure flexibility in problem solving. Suppose the student is asked to think aloud while solving a problem. If the student verbalizes several options then the test is a more valid measure of flexibility in problem-solving than if the student verbalizes mathematical algorithms.

How Many Validities Are There?

It is worth note that the reduction of types of validity can not only be seen as from four types to three types but from many types to three types. Hubley and Zumbo (1996) clarify that, although that the four types (i.e., predictive, concurrent, content, and construct) are most common, many other types have been proposed including factorial and practical validity (Guilford, 1946), face validity (Mosier, 1947), intrinsic validity (Gullliksen, 1950) and Anastasi's (1954) conglomeration of face, content, factorial, and empirical validity. Although the status quo paradigm supports a unified theory of validity, the current 1999 Standards describe the following sources of validity evidence:
1. evidence based on test content, 2. evidence based on response processes, 3. evidence based on internal structure, 4. evidence based on relations to other variables including convergent and discriminant evidence, test-criterion relationships, and validity generalization, and 5. evidence based on the consequences of testing. Due to their popularity in use, the holy trinity is next described.

Convergence of Multiple Sources of Evidence & Judgment

A pioneer in the field of validity, Lee Cronbach, argues that "validation of an instruments calls for an integration of many types of evidence. The varieties of investigation are not alternatives any one of which would be adequate (Cronbach, 1980, p.99)". Further, such techniques for validation are not purely objective. Because test scores are used to make decisions that affect human lives, validation does not occur in a social, political, and economic vacuum. "Empirical evidence, task analysis, formal assumptions, everyday beliefs, and valuations are intertwined in the argument that supports a test use (Cronbach, 1980, p.101)." Even those tempted to use only objective measures for validation should recognize that the selection and development of such techniques are value-laden (Cronbach, 1980).

The status quo mentality in test validation emphasizes convergence of evidence. However, this does not imply sole use of techniques aimed at confirming validity. The falsification paradigm declares a null hypothesis that is the opposite of that which is in attempt to prove. Only with a sufficiently low probability of being incorrect will the researcher reject this hypothesis and contend probabilistic support for that which is in attempt to prove. Inspired by the philosophy of Popper (1963), Cronbach argues, "the job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it (Cronbach, 1980, p.103)." Another advocate of this technique is Landy (1986).

Validity as a Unitary Concept

Primacy of the Construct

Many concur that validity is a unitary concept. In fact, that unitary concept would be the validity of the construct subsuming all other aspects of validity (Cronbach, 1980). Interestingly, early writers on the topic of validity sought primarily to promote construct validity to equal status among the other ‘types’; “Without in the least *advocating* construct validity as preferable to the other kinds (concurrent, predictive, content), we do believe it imperative that psychologists make a place for it in their methodological thinking, so that its rationale, its scientific legitimacy, and its dangers may become explicit and familiar. (Cronbach & Meehl, 1955, p.301).” However, even with the most reliable and elaborate statistical techniques, the element of human judgment must not be ignored. Many argue that the FCAT is not a valid measure for high school graduation decisions (e.g., statistical theory easily destroys the notion that one prompt on the writing assessment is sufficient). This did not stop the state from using the FCAT from making such decisions. Will 200,00 protestors threatening a boycott of Florida’s oranges, amusement parks, and lottery make a difference? “Hard facts influence judgments, but so do anecdotes and personal experience. Judgments embody tradeoffs, not truths (Cronbach, 1980, p.102)”.

In theory, there has been a shift toward acceptance of a unified concept of validity that seeks multiple sources of evidence. The evolution of testing standards marks 1974 as the year such unification began: “These aspects of validity [e.g., predictive, content] can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1974)”. In practice, many testing and measurement specialists neglect the overarching framework of a unified theory in interpreting validity (Shepard, 1993). However, some argue that the abstract nature of an overarching unified interpretation is more suitable for academic discussion than for formal practice (Meehl, 1977).

One example of the unifying nature of construct validity is its implementation in the validation of content. Certainly, it is intended that the content of the measure be representative of the construct being measured. Hence, content validation lends support to the validity of the construct to the extent that the construct is clearly defined and delimited (Haynes, Richard, & Kubany, 1995).

Samuel Messick

Messick provided a comprehensive definition of validity. He defined validity as "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 measurement (1989, p.13)". The impact of this definition is clear when comparing it to the most recent Standards (1999) definition of validity. Messick included the importance of evidence (theoretical and empirical), a focus on test uses, and an emphasis on the interpretation of scores.

Although Cronbach may have been the first to push for a unified theory of validity, Messick is certainly to be credited with over 100 dense and thorough pages targeting this issue. His framework is a four-cell table. The first cell (upper-left) is where the evidential basis meets the test interpretation calling for construct validity. In the second cell (upper-right), the evidential basis meets test use adding relevance and utility to construct validity. The third cell (lower-left) engages the consequential basis to test interpretation to produce value implications. Finally, the fourth cell (lower-right) unites the consequential basis and test use to bear social consequences.

Shepard's Response

Lorrie A. Shepard (1993) noted a potential contradiction in Messick's theory. On one hand, Messick promoted a unified theory of validity in which all types (e.g., predictive, concurrent) are subsumed by construct validity. On the other hand, Messick presented a four-cell table, which according to Shepard "invites a new segmentation of validity requirements (p.426)" further confused by the location of construct validity in the first part. Is construct validity the unified theory or the starting point of a new theory? Further, she argues that Messick did not explicate pertinent and relevant concerns necessary for the support of a particular test's use.

Messick's intended interpretation of the table was that construct validity would appear in each cell, that each cell was to include all components of previous cells. Hence, Shepard's concerns are that of utility. In academic circles, Messick's conception of validity is a mansion with many rooms for discussion and debate. In the world of testing, Messick's conception may not be practical for practitioners. Shepard argues, "we are in dispute only about how these issues should be communicated, by theorists to measurement practitioners and ultimately by the field of external audiences. (p.428)".

Shepard (1993) argues that Messick's four-cell table promotes an unrealistic validation process that would require unlimited time and resources to test a vast number of hypotheses. She calls instead for a validation test that is in alignment with the current Standards definition grounded on the proposed uses of the test. "If a test is proposed for a particular use, a specific network of interrelations should be drawn focused on the proposed use (Shepard, 1993, p.429)."

Messick's Six Aspects (but not really)

In 1996, Messick delineated six aspects of validity – 1.) consequential, 2.) content, 3.) substantive, 4.) structural, 5.) external, and 6.) generalizability. Of course, he clarified that these six aspects should still be viewed within a unified framework. This shift in ideology is interesting because it suggest the possibility that the unified theory and the multiple perspectives theory are not mutually exclusive. Possibly, the distinction is similar to a Gestalt where types, aspects, or facets of validity are parts and the unified theory is the whole. Where the whole is radically other than the sum of its parts, in this case the unified theory is radically other than the integration of multiple validities. Perhaps the function of the unified vs. multiple validities debate is due to a desire to develop academic theory as well as a desire to develop practical validation methods.

Intentions, Uses, and Consequences

The most current Standards (1999) definition of validity is "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. (p.9)". The most striking feature of this definition is the emphasis on the test score. This feature is not new; Cronbach (1970) referred to the accuracy of test score inferences in establishing validity.

When discussing validity, a popular misconception is that a test, survey, or other means of measurement is capable of even possessing validity. A test is not valid or invalid in of itself (Suen, 1990). "One validates, not a test, but an interpretation of data arising from a specified procedure (Cronbach, 1971, p.447)." The most current definition is explained similarly as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (AERA, APA, & NCME, 1999, p.9)". Such uses may include the purposes of placement, diagnosis, assessment, prediction, and evaluation (Katz, 1957). The operative terms here are interpretation, data, procedure, and uses. Hence, the validity of the test is composed of how quantities are linked from behavioral responses (viz., a concrete manifestation of an objective (Hambleton, 1980) to theoretical constructs, how such quantities are obtained, and how they are used ultimately to make substantive human decisions. Because the theoretical constructs are rarely observable, the emphasis must be data-driven with a skeptical eye given toward establishing and claiming the validity of means used to gather such data. The newer definitions of validity emphasize the necessity of establishing the validity of measures in terms of scores and how those scores will be used to make human decisions. Even a quantitative definition of validity relates observed variance in scores to purposes (Brown, 1982).

One example is the Florida Comprehensive Assessment Test (FCAT). Starting in 2003, satisfactory performance on this test precludes both advancement from the third to fourth grade as well as high school graduation. The results were that approximately 43,000 third graders would be retained and approximately 14,000 seniors would not graduate high school. One argument is that the use of one test (viz., the FCAT) is not a valid measure of performance for such high-stakes decisions. As of this writing, the governor of Florida is considering the use of college-entry exam scores as an alternative assessment of high-school graduation readiness. However, this proposal did not stop a protest by approximately 200,000 parents and students in south Miami Beach who argued that students with high grade points averages are failing the FCAT and that the test scores are disproportionately low for blacks and Hispanics. Perhaps Cronbach foreshadowed such incidences when stating, “the public intends to judge specific test uses for itself, directly or through courts and legislatures” (Cronbach, 1980, p.100).

Note that this places the responsibility of test validation upon the test score interpreter (Cronbach, 1980). An example of the necessity of this assignment of responsibility is the use of the Graduate Record Examination (GRE) scores for graduate school admission decisions. This test is but one assessment used by admissions committees to determine acceptance into their respective programs. Other assessments include undergraduate grade point average, personal statements, and recommendation letters. Hence, the use of GRE scores will vary from institution to institution. Further, the relevance of GRE scores may vary from program to program. Hence, it is incumbent on each individual graduate admission committee to determine the validity of GRE scores in making admission decisions.

It should be noted that an overt emphasis on intentions, uses, and consequences can also hurt validity. Campbell (1996) refers to use in social control where validity drops because of well-known testing purposes. For example, teachers aware of testing for accountability learned to “teach to the test” prompting Crocker’s (2003) call to “teach for the test”. Further, potential graduate students may score high on admission exams because of coaching techniques irrelevant to the constructs targeted on the exam.

When Is Enough Enough?

Validity is currently seen as a never-ending process of accumulating evidence that either supports or fails to support an interpretation and use of a test score (Hambleton, 1980). “Confidence in a theory is increased as more relevant evidence confirms it, but it is always possible that tomorrow’s investigation will render the theory obsolete

(Cronbach & Meehl, 1955, p.300)." Further, claiming that a test is valid ignores the conditional nature of validity (Haynes, Richard, & Kubany, 1995; Nunnally, 1978). Indeed, the word "degree" should not be overlooked when reading the Standards definition of validity (APA, AERA, & NCME, 1999).

Although an advantage to this conditional mandate is flexibility, a disadvantage is a lack of any boundaries in determining completion. On one hand, due to the political, social, and economic interaction with test scores, perhaps there should be no stopping point. On the other hand, at some point, a test administrator must decide whether or not to use a test. Shepard (1993) contends that the never-ending process gives "practitioners permission to stop with incomplete and unevaluated data".

A more pragmatic approach would base the choice of validation techniques on relevant questions clarifying "which validity questions must be answered to defend a test use and which are academic refinements that go beyond the immediate, urgent questions (Shepard, 1993, p.407)". Shepard's argument is that other models (e.g., Messick) fail to reach the evaluation of consequences. The structure of test validation should be grounded upon when the claims of a particular test practice. In other words, if the Law School Admission Test (LSAT) claims to yield scores that can be effectively used for law school admission decisions, then the primary question is, "Does the use of LSAT scores result in the admission of the most qualified law school students?"

Continuing Shepard's advice on starting with the proposed uses of the test, the question should be answered from the onset of validation. What evidence is necessary to defend the use of test and warrant its implementation? She explains that for research purposes the first cell of Messick's table requiring solely construct validity may be adequate. However, if a test is being used to determine the academic future of millions of student, all four cells are required in the validation process. Kane (1992) presents an outline equating test validation to the process of analyzing any argument: 1.) clearly state the argument, 2.) connect argument with reasonable assumptions, and 3.) support assumptions with plausible explanations and evidence.

Conclusion: Limitations in Score Interpretation

The key problem in defining validity and implementing techniques for validation is that such processes are usually considered during or after test construction. In a sense, validity is a form of quality control. We tend to address validity as a necessary process similar to reliability while ignoring the fact that validity issues guide testing from its origin.

It is imperative that we consider the root of validation, which is the initial rationale for constructing the test in the first place. Why has the test been constructed? The test has been constructed in response to a need. These needs have traditionally taken place in the forms of expeditious methods to select qualified job candidates, to determine academic readiness and placement, and to assess psychological functions. Why test? The scores will be used to make a human decision with the intent that more faith should be placed in quantification than in human judgment (although it is, indeed, human judgment that guides all aspects of test construction, implementation, and interpretation).
Hence, validation begins before the test is constructed with the primary questions, “Why is the test deemed necessary?” and “What decisions will be made using the scores?”

Ultimately, these questions should guide all subsequent questions regarding such aspects of validity as the adequacy of the content domain, the sampling of content, the operationalisation of constructs, relationships between constructs in the nomological net, prediction, and generalization. The first question addresses situation and the second question addresses function. To situate validity means to place it within its pertinent social, economic, and political time frame. What forces prompt the development of the test? What guiding ideologies and values are underlying the process? Note that even the evolution of the definition of validity in the Standards is in part a function of politics (Shepard, 1993). Function addresses the pragmatic and utilitarian purposes of the test. What are the intended consequences? Are there multiple motives shifting between social, economic, and political forces? How does each introduce potential unintended consequences?

A key feature of these questions is that their answer is not necessarily stable and enduring. Rationales for testing and interpretations of resulting scores interact with the ideologies grounded in a particular social, political, and economic timeframe. Hence, test validity at its core is value-laden. Certainly, objective techniques have been constructed in attempt to quantify test validity. However, the pervasive vociferations to validate consequences and uses beyond statistical techniques suggest that quantification is not sufficient.

According to Shepard (1993), “validity investigations cannot resolve questions that are purely value choices…However, to the extent that contending constituencies make competing claims about what a test measures, about the nature of its relations to subsequent performance in school or on the job, or about the effects of testing, these value-laden questions are integral to a validity evaluation. (p.428)”.

An already validated test may require revalidation. This is normally attributed to differences in characteristics between the test validation sample and the sample with which the test is to be used. However, Haynes, Richard, & Kubany (1995) state that “content validity often degrades over time as new data are acquired and theories about the targeted construct evolve. (p. 241)”. Further, as tests are constructed due to political, economic, and social forces and because all phases of testing include value judgments, test validity will change because of changes in the forces and values. If an established test is to be implemented, then it is incumbent upon the test user to question the influences of social, political, and economic changes on the interpretation of scores from this test in tandem with any further quantifications to ascertain the degree of validity.

REFERENCES

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of Ratings. *Educational and Psychological Measurement, 45*, 131-142.

Aiken, L. R. (1996). Rating scales and checklists: Evaluating behavior, personality, and attitudes. New York: Wiley.

Albright, L., & Malloy, T. E. (2000). Experimental validity: Brunswik, Campbell, Cronbach, and enduring issues. *Review of General Psychology, 4*, 337-353.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques: Preliminary proposal. *American Psychologist*, 7, 461-476.

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

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

Anastasi, A. (1954). *Psychological testing* (1st ed.). New York: Macmillan.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement, 10*, 67-78.

Anastasi, A. (1988). Psychological testing (6th ed.). New York: Macmillan Publishing.

Angoff, W. Hy. (1988). Validity: An evolving concept. In R. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bornstein, R. F. (1996). Face validity in psychological assessment. *American Psychologist*, *51*, 983-984.

Brown, F. G. (1986). *Principles of educational and psychological testing* (3rd ed.). New York: Holt, Rinehart, & Winston.

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

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Review, 54*, 297-312.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7,* 309-319.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago: Rand McNally.

Crocker, L. (2003). Teaching for the Test: Validity, Fairness, and Moral Action. *Educational Measurement: Issues and Practice*. (In Press).

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Belmont, CA: Wadsworth Group.

Crocker, L., Llabre, M., & Miller, M. D. (1988). The generalizability of content validity ratings. *Journal of Educational Measurement, 25*, 287-299.

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing The fit between test and curriculum. *Applied Measurement in Education, 2,* 179-194.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (3rd ed., pp. 201-219). Washington, DC: American Council on Education and National Council on Measurement in Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade* (pp. 99-108). San Francisco, CA: Jossey-Bass.

Cronbach, L. J. (1984). Essentials of psychological testing (4th ed.). New York: Harper & Row.

Dooley, D. (2001). *Social research methods.* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Dunn, J. G. H., Bouffard, M., & Rogers, W. T. (1999). Assessing content-relevance in Sport psychology scale-construction research: Issues and recommendations. *Measurement in Physical Education and Exercise Science*, *3*, 15-36.

Friedenberg, L. (1995). *Psychological testing: Design, analysis, and use*. Boston: Allyn & Bacon.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurements, 6,* 427-439.

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist, 5,* 511-517.

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

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Berk (ed.), *Criterion-referenced measurement: the state of the art.* Baltimore: John Hopkins University Press.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123*, 207-215.

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

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

Katz, M. (1975). *Selecting an achievement test: Principles and procedures (3rd. ed.)*.. (Evaluating and Advisory Series). Princeton, NJ: Educational Testing Service.

Klein, S. P., & Kosecoff, J. P. (1975). *Determining how well a test measures your objectives*. (CSE Rep. No. 94). Los Angeles: University of California, Center for the Study of Evaluation.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183-1192.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28,* 563-575.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35,* 382-385.

MacCorquodale, K. & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55,* 95-107.

McClung, M. S. (1978). Competency testing programs: Legal and educational issues. *Fordham Law Review, 47*, 651-712.

Meehl, P. E. (1977). Specific etiology and other forms of strong influence: Some quantitative meanings. *Journal of Medicine and Philosophy*, 2, 33-53.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: ACE and Macmillan.

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

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.

Messick, S. (1996). *Validity of performance assessment.* In Phillips, G. (1996). Technical Issues in Large-Scale Performance Assessment. Washington, D.C.: National Center for Educational Statistics.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379-387.

Morris, L. L., & Fitz-Gibbons, C. T. (1978). *How to measure achievement*. Beverly Hills, CA: Sage.

Mosier, C. J. (1947). A critical examination of the concept of face validity. *Educational and Psychological Measurement, 7,* 191-205.

Nathan, B. R., & Tippins, N. (1990). The consequences of halo “error” in performance Ratings: A field study of the moderating effect of halo on test validation results. *Journal of Applied Psychology, 75,* 290-296.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Penfield, R. D. (in press). A score method of constructing asymmetric confidence intervals for the mean of a rating scale item. *Psychological Methods*.

Popper, K. (1963). *Conjectures and refutations*. London: Routledge and Kegan Paul.

Reiter-Palmon, R., & Connelly, M. S. (2000). Item selection counts: A comparison of empirical key and rational scale validities in theory-based and non-theory based item pools. *Journal of Applied Psychology, 85*, 143-151.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research*, 2, 49-60.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education, Vol. 19*. Washington, D.C.: American Educational Research Association.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Erlbaum.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology, 30,* 47-54.

Turner, R. C. Mulvenon, S. W., Thomas, S. P., & Balkin, R. S. (2003). Computing Indices of item congruence for test development validity assessments. *International Journal of Testing*. (In press).

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12*, 10-14.