DOCUMENT RESUME

ED 387 508 TM 023 685

AUTHOR Sireci, Stephen G.
TITLE The Central Role of Content Representation in Test Validity.
PUB DATE Apr 95
NOTE 38p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Construct Validity; *Content Validity; *Definitions; Item Analysis; Responses; Scoring; Test Construction; *Test Content; Test Items; Test Theory

ABSTRACT

The purpose of this paper is to clarify the seemingly discrepant views of test theorists and test developers about terminology related to the evaluation of test content. The origin and evolution of the concept of content validity are traced, and the concept is reformulated in a way that emphasizes the notion that content domain definition, relevance, and representation are necessary and fundamental qualities on which all tests should be evaluated. An important area of convergence for test theorists and test developers has been that adequately defining and representing the construct measured is of critical importance. Broad definitions have asserted that content validity is concerned with test and response properties, but narrow definitions limit content validity to investigations of items, tests, and scoring. Future descriptions of test validity must emphasize the important role of content validity in the construction and evaluation of test. Theories and applications of content validity belong within the framework of construct validity. (Contains 1 table, 2 figures, and 69 references.) (SLD)

The Central Role of Content Representation in Test Validity

Stephen G. Sireci[1]

GED Testing Service
American Council on Education

Paper Presented at the Annual Meeting of the National Council on Measurement in Education, as Part of the Symposium, "The Construct of Content Validity: Theories and Applications," San Francisco, CA, April, 1995

---

[1]The quality of this paper was enhanced by recent conversations with Kurt Geisinger, Leon Gross, and Douglas Jackson, and editorial comments provided by Lynn Shelley-Sireci. The opinions expressed in this paper are my own and do not represent an official position of the GED Testing Service or the American Council on Education.

Introduction

An important component in demonstrating test validity is confirming that the test accurately reflects the underlying domain it purports to measure. The term *content validity* was introduced in the early 1950s to describe this desirable test quality. However, for more than twenty years, theorists have argued against use of this term (e.g., Fitzpatrick, 1983; Guion, 1977, 1980; Messick, 1975, 1980, 1989; Tenopyr, 1977). Nevertheless, test developers continue to strive to demonstrate that their tests are content-valid. The purpose of this paper is to clarify the seemingly discrepant views between test theorists and test developers over terminology related to the evaluation of test content. To accomplish this clarification, the origin and evolution of content validity is traced. Finally, a re-formulation of the concept of test validity is proposed that emphasizes the notion that content domain definition, relevance, and representation are necessary and fundamental qualities upon which *all* tests should be evaluated.

Currently, the unitary conceptualization of test validity is prominent (Messick, 1989b). This conceptualization asserts that because the purpose of testing is to make inferences from observed test scores to unobservable constructs, the evaluation of a test requires evaluating the construct validity of these inferences. Given this definition, other established "categories" of validity, such as content- and criterion-related validity, are subsumed under the more general construct validity rubric.

The idea that construct validity is the general, unifying characterization of test validity is not new. In fact, shortly after Cronbach and Meehl (1955) formulated construct validity, Loevinger (1957) argued that "all validity is construct validity." However, the notion of different "types," "aspects," or "categories" of validity persevered. Even today, construct validity is not universally accepted as equivalent to validity in general. In particular, many test specialists have trouble abandoning the notion of content validity (e.g., Yalow & Popham, 1983). Why has the concept of content validity persevered throughout the ages? Why does it continue to frustrate comprehensive and sophisticated formulations of test validity? The answer to these questions is facilitated by a review of the origin and evolution of content validity.

The Origin and Evolution of Content Validity

Early conceptions of validity: criterion correlations

From the earliest days of educational and psychological testing, validation procedures attempted to demonstrate the utility of a test by correlating test scores with an external criterion

(Bingham, 1937; Kelley, 1927; Thurstone, 1932). The external criterion with which scores were correlated was one considered germane to the purpose of the testing (e.g., school grades, supervisor ratings). These correlational studies promoted "validity coefficients", which provided an empirical index of the degree to which a test measured what it purported to measure.

Validity coefficients were often taken as exclusive evidence of a test's validity. However, early test evaluators gradually became critical of their shortcomings. One major problem with validity coefficients was demonstrating the relevance of the chosen criterion to the purpose of the testing (Thorndike, 1931). Another, more serious, problem was demonstrating the validity of the criterion itself (Jenkins, 1946). To redress the limitations of this validation procedure, attempts were made to define validity in theoretical, as well as empirical, fashion.

Kelley (1927), for example, while supporting correlational evidence of validity, expressed concern regarding its limitations, and suggested professional judgment be used to supplement evaluations of test validity. This position was characteristic of the growing concern of the early test specialists that a purely empirical perspective on validation was too restrictive (e.g., Thorndike, 1931; Toops, 1944). In reaction to this realization, new conceptions of validity began to emerge. These new thoughts promulgated different "types" of validity and different "types" of tests.

Guilford (1946), for example, classified validity into two categories: factorial validity and practical validity. Practical validity referred to the traditional correlations among test scores and relevant criteria, while factorial validity referred to the factor loadings of the test on "meaningful, common, reference factors" (p. 428). As can be inferred from his two "types" of validity, Guilford championed an empirical approach to test validation. He maintained that "... only by an objective, empirical procedure such as factor analysis can we know what abilities and traits are represented in either tests or jobs" (p.433). However, he also postulated the plausibility of "validity by inspection":

> When validation data are lacking, the construction or adaptation of a test...must proceed on the basis of considerable guesswork...or professional judgement. A natural and relatively safe approach is to devise a 'job sample' test...Such tests have a fair probability of being valid for that particular job. (p. 436)

Rulon (1946) also recommended an operational approach to test validation. The central elements of his approach were: 1) a test cannot be labeled "valid" or "invalid" without respect to a given purpose; 2) an assessment of a test's validity must include

an assessment of the content of the test and its relation to the purpose of the testing; 3) different forms of validity evidence are required for different types of tests; and 4) some tests are "obviously valid" and need no further study. This approach was innovative in that it required that the purpose of the testing and the appropriateness of test content be evaluated as part of the validation process.

Rulon refrained from creating a new "type" of validity for the "obviously valid" tests; however, some researchers used the term "face validity" to describe this quality. Mosier (1947) expressed concern over use of the term "face validity" and the multiple meanings it acquired. He identified three distinct connotations typically attributed to the term "face validity": 1) validity by assumption, 2) validity by definition, and 3) appearance of validity.

To Mosier, "validity by assumption" referred to the idea that a test could be considered valid if "... the items which compose it 'appear on their face' to bear a common-sense relationship to the objective of the test" (p. 208). He dismissed this "type" of validity as a "pernicious fallacy" (p. 209). His second type of validity, "validity by definition", referred to situations where test questions defined the objective of the testing. In such cases, the validity of the test was represented by the square root of the reliability coefficient. This notion was consistent with Rulon's (1946) description of "obviously valid" tests, and according to Mosier, was the initial intention of the concept of face validity. His last definition, "appearance of validity," referred to the additional requirement that tests appear pertinent and relevant to test consumers and examinees. Mosier noted that this last "type" of validity is not validity at all, but rather an "... additional attribute of the test which is highly desirable in certain situations" (p. 208).

The only connotation associated with face validity that Mosier supported was validity by definition. He stated that this type of validity was important and could be established through the use of subject matter experts. In his formulation,

> ... the test is considered to be valid if the sample of items appears to the subject matter experts to represent adequately the total universe of appropriate test questions. (p. 208)

Mosier argued that validity by definition could be accomplished through subjective, rather than empirical analysis. However, he did not assert that this method of validation was appropriate for all types of tests.

Goodenough (1949) supported Rulon's notion of different types of validity for different tests by classifying tests into two broad

categories: tests as samples, and tests as signs. "Tests as samples" referred to tests considered representative samples of the universe (domain, trait, etc.) tested. "Tests as signs" referred to tests that point to some external universe and provided guidance for a description of the universe. Goodenough's taxonomy related educational achievement tests to "sample" tests, and aptitude and personality tests to "sign" tests. She described sample tests as representative samples of behaviors of the universe tested, and stressed that it was important for tests to attempt to represent this universe; even if complete representation was infeasible. Taken together, Mosier's (1947) description of a "universe of appropriate test questions," and Goodenough's (1949) description of "tests as samples," paved the way for the notion that tests were linked to an underlying content domain, and that evaluating the validity of a test should consider how well the tasks which comprise a test represent that domain.

Thorndike (1949) also stressed consideration of underlying domains when evaluating and constructing tests. He described two approaches to test construction useful for tests designed for personnel selection: the trait approach and the job approach. The trait approach involved identifying individual traits important to a given job and incorporating tests measuring these traits into a test battery. The job approach involved the construction of tests that reproduced key features of the job. It defined the test or test battery as a sample of tasks that represented the total universe of job tasks.

For job sample tests, Thorndike stressed the importance of evaluating the content specifications of the test including: 1) the function(s) that the test is trying to measure, 2) the range of content to be covered in the test, and 3) the editorial and statistical procedures to be used in selecting and refining test items. He considered it important to define the domain to be tested, define the relationship between the test and the domain, and evaluate the items in terms of their appropriateness in measuring this domain. This latter concern he described as "relevance to the ultimate goal" (p. 125). Thorndike asserted that concerns of the appropriateness of test content could not be resolved on strictly empirical grounds, and acknowledged the importance of including the judgments of content experts in the validation process. Thorndike supported subjective evidence of validity, but in keeping with previous conceptualizations, he did not consider such information sufficient for validation, or on a par with empirical evidence of validity.

Like Rulon (1946), Mosier (1947), and Thorndike (1949), Gulliksen (1950a) acknowledged the importance of evaluating test content when validating a measure. However, Gulliksen stressed that evaluations of test content should be empirically based. He proposed three empirical procedures that could be used to

evaluate what he termed "intrinsic validity": 1) evaluate test results before and after training in the subject matter at hand, 2) assess the consensus of expert judgement in evaluations of the test content, and 3) assess the relationship of the test to other tests measuring the same objective. Gulliksen's rationale in recommending the above procedures was that if the content of the test was appropriate, then posttraining scores would exhibit superiority over pretraining scores, there would be a fair degree of consensus among the judges (regarding the appropriateness of the content), and the test would agree with other tests measuring the same objective. The influence of Gulliksen's recommendations are evident in contemporary evaluations of tests involving pretest-posttest comparisons, subject matter expert consensus, and concurrent validity.

The writings of Rulon (1946), Mosier (1947), Goodenough (1949), Thorndike (1949), and Gulliksen (1950a, 1950b) delineated the fundamental precepts that eventually emerged as content validity: domain definition, domain representation, and domain relevance. These researchers, among others, signaled a change in the conception and practice of test validation. This change expanded validity beyond the notion of correlating test scores with criterion measures, and stressed that validation must consider the appropriateness of test content in relation to the purpose of the testing.

The emergence of content validity

Given that several new and varied thoughts concerning test validity were being advanced, there was a need in the early 1950s to summarize and clarify the multiple meanings it acquired. An early synopsis of these varying conceptualizations was Cureton's (1951) "Validity" chapter that appeared in the first edition of *Educational Measurement* (Lindquist, 1951). Cureton presented the newer ideas of content validation along with the older notions that defined validity in terms of correlations of test scores with external criteria. His chapter marked an early introduction of the term "content validity" into the literature of educational and psychological testing.

Cureton described two "aspects" of validity: relevance and reliability (p. 622). "Relevance" referred primarily to the degree of correspondence between test scores and criterion measures, and "reliability" referred to the accuracy and consistency of test scores. These two "aspects" of validity closely paralleled earlier notions of validity and reliability. However, Cureton also acknowledged the existence of "curricular relevance or content validity" (p. 669) that was appropriate in some educational settings.

Cureton's description of content validity was congruent with the formulations of Rulon (1946), Mosier (1947), and Thorndike

(1949), in that it related the importance of the sample of tasks to the domain tested. Like Guilford (1946) and Thorndike (1949), Cureton did not consider evidence of content validity to be as valuable as empirical evidence. However, he noted its importance in educational and industrial testing and provided guidance in gathering such evidence:

> If we validate items statistically, we may accept the judgement that the working criterion is adequate, along with all the specific judgments that lead to the criterion scores. We may, alternatively, ask those who know the job to list the concepts which constitute job knowledge, and to rate the relative importance of these concepts. Then when the preliminary test is finished, we may ask them to examine its items. If they agree fairly well that the test items will evoke acts of job knowledge, and that these acts will constitute a representative sample of all such acts, we may be inclined to accept these judgments. (p. 664)

Thus by 1951, techniques and procedures for evaluating test content were alive and well. In fact, perceptions of test validity were changing so rapidly that the American Psychological Association (APA) commissioned a panel to offer a formal proposal of test standards to be used in the construction, use, and interpretation of psychological tests. This committee, the APA Committee on Test Standards, dramatically changed the conception and terminology of validity.

The first product from the Committee was the *Technical Recommendations for Psychological Tests and Diagnostic Techniques: a Preliminary Proposal* (APA, 1952). This publication promulgated four categories of validity: predictive validity, status validity, content validity, and congruent validity. The Committee did not explicitly define content validity, but rather described it in terms of specific testing purposes:

> Content validity refers to the case in which the specific type of behavior called for in the test is the goal of training or some similar activity. Ordinarily, the test will sample from a universe of possible behaviors. An academic achievement test is most often examined for content validity. (p. 468)

Although the Committee proposed content validity as a category of validity, several caveats concerning its use were raised. For example, the introductory paragraph describing content validity read:

> Claims or recommendations based on content validity should be carefully distinguished from inferences established by statistical studies ... While content

> validity may establish that a test taps a particular area, it does not establish that the test is useful for some practical purpose. (p. 471)

In the introduction to the recommendations concerning validity, the notion of content validity is nearly dismissed for non-achievement type tests: "few standards have been stated for content validity, as this concept applies with greatest force to achievement tests" (p. 468). Ironically, though the Committee limited the relevance of content validity, they proposed strict standards to govern it:

> If content validity is important for a particular test, the manual should indicate clearly what universe of content is represented ... The universe of content should be defined in terms of the sources from which items were drawn, or the content criteria used to include and exclude items... The method of sampling items within the universe should be described (p. 471)

The 1952 *Recommendations* asserted that content validity referred primarily to achievement tests and that the goals of domain definition and domain sampling were attainable goals.

In response to the preliminary *Recommendations*, a joint committee of the APA, American Educational Research Association (AERA), and the National Council on Measurements Used in Education (NCME) was formed, and published the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954). This publication featured several modifications of the 1952 proposed recommendations. For example, the four "categories" of validity noted in the 1952 proposal were referred to as "types" or "attributes" of validity. "Congruent validity" was renamed "construct validity", and "status validity" was renamed "concurrent validity." Another significant difference between the 1952 proposal and the formal publication in 1954 was the increased consideration given to content validity. The description of content validity in the 1954 *Recommendations* read:

> Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn. Content validity is especially important in the case of achievement and proficiency measures. In most classes of situations measured by tests, quantitative evidence of content validity is not feasible. However, the test producer should indicate the basis for claiming adequacy of sampling or representativeness of the test content in relation to the universe of items adopted for reference. (p. 13)

This divergence from the description of content validity

presented in the 1952 document was significant in that content validity was not limited to cases where the function of testing was to assess the impact of training. Rather, content validity was considered relevant to industrial and personality testing. In fact, content validity was elevated to a level of importance equivalent to that of the other aspects of validity:

> It must be kept in mind that these four aspects of validity are not all discrete and that a complete presentation about a test may involve information about all types of validity. A first step in the preparation of a predictive instrument may be to consider what constructs or predictive dimensions are likely to give the best prediction. Examining content validity may also be an early step in producing a test whose predictive validity is ultimately of major concern. (p. 16)

Although the importance of content validity was increased in the 1954 version, caveats referring to its limitations were retained.

<u>All validities are not created equal: the content validity controversy begins</u>

After publication of the 1954 *Technical Recommendations*, the notion of four separate but equal types of validity became accepted terminology in the psychometric literature. However, not all test specialists agreed that the types of validity were of equal import. For example, Anastasi, in her first edition of *Psychological Testing* (1954) articulated several caveats concerning content validity. In keeping with the 1952 *Proposed Recommendations* she described content validity as "especially pertinent to the evaluation of achievement tests" (p. 122), and warned against generalizing inferences made from the test scores to more general content areas, or to groups of people who might be differentially affected by the content. Furthermore, Anastasi did not support the use of content validity for validating aptitude or personality tests.

Cronbach and Meehl (1955) clarified the four types of validity and emphasized construct validity, which they believed applied to all tests. They described validity as being of three major types: criterion-related (which included both predictive and concurrent validity), content, and construct. They asserted that construct validity "... must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured" (p. 282). Though Cronbach and Meehl did not subsume content validity under construct validity, their work facilitated the notion that construct validity is always involved in test validation.

Lennon (1956) recognized the ambiguity in the descriptions of

content validity and proposed a formal definition. He noted that "the term content validity is not defined in the text of APA's Technical Recommendations, but the meaning intended may readily be inferred ..." (pp.294-295). He proceeded to define content validity in terms of the degree of correspondence among responses to test items and responses to the larger universe of concern:

> We propose in this paper to use the term content validity in the sense in which we believe it is intended in the APA Test Standards, namely to denote *the extent to which a subject's responses to the items of a test may be considered to be a representative sample of his responses to a real or hypothetical universe of situations which together constitute the area of concern to the person interpreting the test.* (p. 295)

Lennon's definition differed from previous descriptions of content validity in that it included *responses* to test items rather than only the items themselves. He stated that, "this is to underscore the point that appraisal of content validity must take into account not only the *content* of the questions but also the *process* presumably employed by the subject in arriving at his response" (p. 296). Thus, Lennon viewed content validity as an interaction between test content and examinee responses. This view was consistent with Anastasi's (1954) concerns regarding the differential effects of test content on different groups of examinees. Both Anastasi and Lennon implied that if different groups of examinees employed different processes in responding to test items, content-by-process interactions could threaten score interpretation. Therefore, evaluations of test content must consider the population tested, because inferences drawn from these evaluations may not generalize to other populations.

Lennon's work was an important clarification of the concept of content validity. He provided justification for its use and asserted that, like the other forms of validity, "content validity ... is specific to the purpose for which, and the group with which, a test is used" (p. 303). Lennon also listed three assumptions governing the use of content validity: 1) the area of concern to the tester must be conceived as a meaningful, definable, universe of responses; 2) a sample must be drawn from this universe in some useful, meaningful fashion; and 3) the sample and the sampling process must be defined with sufficient precision to enable the test user to judge how adequately performance on the sample typifies performance on the universe.

Loevinger (1957) viewed content validity differently from Lennon. She pointed out that content domains were essentially hypothetical constructs, and borrowing from Cronbach and Meehl (1955), asserted that "... since predictive, concurrent, and content validities are all essentially ad hoc, construct validity

is the whole of validity from a scientific point of view" (p. 636). Loevinger described other "types" of validity as mutually exclusive "aspects" of construct validity. She did not dismiss content validity as unimportant, but rather described it as an important stage of the test construction process.

Loevinger developed the concept of "substantive validity" to incorporate the concerns of test content within the framework of construct validity. She described substantive validity as "... the extent to which the content of the items included in (and excluded from?) the test can be accounted for in terms of the trait believed to be measured and in the context of measurement" (p. 661). Her implication that test content should be assessed in terms of its relation to a measured trait, rather than to a measured content domain, illustrated her rationale in subsuming the concept of content validity under the rubric of construct validity.

Though Loevinger (1957) attempted to provide a cohesive theory of test validity centered on construct validity, other test specialists were dissatisfied with such a formulation. Ebel (1961), for example, rejected philosophical descriptions of validity and called for "... a more concrete and realistic conception of the complex of qualities which make a test good" (p. 641). He asserted that the "types" of validity were scientifically and philosophically weak and pointed out that the descriptions of validity by Cureton (1951), Loevinger (1957), and others, had not led to practical or scientific improvement: "so long as what a test is supposed to measure is conceived to be an ideal quantity, unmeasurable directly and hence undefinable operationally, it is small wonder that we have trouble validating our tests" (p. 643). Ebel recommended that use of the term "validity" be abandoned altogether, and replaced by "meaningfulness" (p. 645).

Ebel's pragmatism supported the notion that the meaningfulness of a test could be established by a substantive analysis of the test's content. He described three characteristics useful for determining test quality: 1) the importance of inferences that can be made from the test scores, 2) the meaningfulness of test scores, and 3) the convenience of the test in use. An important aspect of the meaningfulness of test scores he described as "an operational definition of the measurement procedure" (p. 646). Thus, although Ebel approved of abolishing the label "content validity," he underscored the requisite that tests must be able to define and represent the content domain.

Although the writings of Loevinger (1957) and Ebel (1961) employed different terminology, there was clear agreement that the appropriateness of test content was dependent upon sound test construction procedures. Therefore, evaluation of test content required evaluation of those procedures.

Further "clarification": the 2nd version of the *Standards*

In response to the issues raised by Cronbach and Meehl (1955), Lennon (1956), Loevinger (1957), Ebel (1961), and others, APA, AERA, and NCME revised the 1954 *Technical Recommendations* and published the *Standards for Educational and Psychological Tests and Manuals* (1966). This collaborative effort resulted in several changes in the description of validity that supported the notion of content validity.

The 1966 *Standards* reduced the four "types" of validity to three "aspects" of validity, subsuming concurrent and predictive validities under the rubric of "criterion-related validity" (as suggested by Cronbach and Meehl, 1955). Another modification incorporated into the 1966 *Standards* was the notion that test users were also responsible for maintaining validity. It was strongly recommended that the test users apply adequate judgement when considering use of a test for a particular purpose. The 1966 standards stated, for example, that "...even the best test can have damaging consequences if used inappropriately. Therefore, primary responsibility for the improvement of testing rests on the shoulders of test users" (p. 6). Consideration of test content was an important part of evaluating the appropriateness of a test for a given purpose.

The 1966 *Standards* elevated the importance of content validity in the evaluation of achievement tests. It explicitly stated that, for achievement tests, content validation was necessary to supplement the evidence gathered through criterion-related studies:

> Too frequently in educational measurement attention is restricted to criterion-related validity. Efforts should also be directed toward both the careful construction of tests around the content and process objectives furnished by a two-way grid and the use of the judgment of curricular specialists concerning what is highly valid in reflecting the desired outcomes of instruction. (p. 6)

Although content validity was described as imperative for educational tests, its use was not limited to achievement testing: "content validity is especially important for achievement and proficiency measures and for measures of adjustment or social behavior based on observation in selected situations" (p. 12). This description was much broader than that provided in the 1954 *Recommendations*, implying that content validity is relevant for psychological and industrial testing.

Another central theme of the 1966 *Standards* was the assertion that particular testing purposes called for specific forms of

validation evidence. Each of the three "aspects" of validity required different forms of validity evidence considered relevant to one of three "aims of testing." With respect to content validity, the 1966 *Standards* asserted that it applied to all measures assessing an individual's current standing with respect to a substantive domain.

The 1966 *Standards* expanded previous descriptions of content validity by defining it in operational terms; i.e., as an evaluation of the operational definition of the content domain tested, and the accuracy of the sampling of tasks from that domain:

> Content validity is demonstrated by showing how well the content of the test samples the class situations or subject matter about which conclusions are to be drawn... The [test] manual should justify the claim that the test content represents the assumed universe of tasks, conditions, or processes. A useful way of looking at this universe of tasks or items is to consider it to comprise a definition of the achievement to be measured by the test. In the case of an educational achievement test, the content of the test may be regarded as a definition of (or a sampling from a population of) one or more educational objectives ... Thus, evaluating the content validity of a particular test for a particular purpose is the same as subjectively recognizing the adequacy of a definition. (pp. 12-13)

Because the operational definition of the content domain tested is formally represented by the content specifications of a test, the 1966 *Standards* suggested that these specifications be evaluated in examining a test's content validity. They also asserted that the test manual should define the content domain tested and indicate how well the test represents the defined domain. The requirements listed for meeting this standard involved ensuring that: the universe was adequately sampled, the "experts" who evaluated the content were competent, and that there was substantial agreement among content experts. In addition, the test manual was required to report: the classification system used for selecting items, the blueprint specifying the content areas measured by the items along with the processes corresponding to the content areas, and the dates when content decisions were made. It was further required that test manuals clearly identify those validation procedures that were the result of "logical analysis of content" from those that were empirically-based.

The 1966 *Standards* presented a comprehensive description of the content validity approach to test validation. Content validity was not described as inferior to the two other "aspects" of

validity, rather it was portrayed as the essential type of validity required for a large category of tests. Unlike the 1954 *Recommendations*, content validity was not criticized for its lack of empirical verifiability; rather, it was portrayed as being superior to empirical forms of evidence in certain situations. The basic principles underlying content validity (domain definition, domain representation, and domain relevance), were stated explicitly, and practical suggestions for evaluating content validity were provided.

The controversy continues: re-emergence of a unitary conceptualization of validity

The second edition of *Educational Measurement* (Thorndike, 1971), featured a chapter on test validity by Cronbach. This chapter summarized and clarified the fundamental precepts of validity presented in the 1966 *Standards*, and set the stage for future revision of the philosophy and practice of test validation.

Cronbach (1971) defined test validation as " ... a comprehensive, integrated evaluation of the test" (p. 445). He described the three aspects of validity as complimentary rather than exclusive, and recommended that all forms of evidence be considered in test validation. In keeping with the 1966 *Standards*, he maintained that certain types of validation evidence were more desirable according to the purpose of the testing.

Cronbach (1971) described content validation as an investigation of the alignment of the test to the universe specifications denoted in the test blueprint. The question asked in content validation was "do the observations truly sample the universe of tasks the developer intended to measure or the universe of situations in which he would like to observe?" (p. 446). In order to address this question the test validator was required:

> To decide whether the tasks (situations) fit the content categories stated in the test specifications ... [and] To evaluate the process for content selection, as described in the manual. (p. 446)

Cronbach stated that content validation involved an assessment of the universe (domain) definition and an assessment of how well the test matched that definition. Because the universe is often defined in terms of a test blueprint, the degree to which the test is congruous with its blueprint was described as a crucial element of content validation. Content validation then, involved ensuring the adequacy of the definition of the content domain tested and the extent to which the test represented that definition:

> Content validation asks whether the test fits the developer's blueprint, and ... whether the test user

would have chosen the same blueprint. (p. 452)

Cronbach's description of content validation was consistent with the writings of Loevinger (1957), Ebel (1961), and Nunnally (1967) in that all asserted that content representation is best achieved by appropriate test construction procedures. Cronbach described a favorable content validation study as one which "... is fully defined by the written statement of the construction rules" (p. 456). He reinforced this notion by later asserting that "To be sure, test construction is no better than the writers and reviewers of the items" (p. 456).

Cronbach clarified the role of judgments used in content validation and distinguished them from judgments employed in construct validation. The former he restricted to the "operational, externally observable side of testing," while the latter required empirical verification (p. 452). Cronbach's treatise provided support and guidance for the practice of content validation. In contrast to Goodenough (1949) and Loevinger (1957), he argued that the assessment of the degree to which a test samples or represents its domain was an attainable goal, accomplished via expert judgement. His contention was that if the items on a test were representative of, and relevant to, the tested domain, then they would hold up under the scrutiny of subject matter experts.

Cronbach affirmed that subjective judgement was the only form of evidence applicable in content validation. He asserted that, "correlations have nothing to do with content validation" and that "... nothing in the logic of content validation requires that the universe or the test be homogeneous in content" (p. 457). Thus evidence of construct and criterion-related validity, as well as intra-test homogeneity (i.e., internal consistency reliability), were not to be adduced as evidence of content validity.

Cronbach's (1971) validity chapter brought many of the divergent theories and practices of validation together within a cohesive framework. Because some of his formulations deviated from the 1966 *Standards*, the time was ripe for a new version. Borrowing largely from Cronbach (1971), AERA, APA, and NCME revised the 1966 *Recommendations*, and published the *Standards for Educational and Psychological Tests* (APA, 1974). This revision retained the notion of three unique "aspects" of validity and made only minor changes in the description of content validity.

The 1974 *Standards* re-emphasized the importance of the domain definition in evaluating test content. Test developers were required to provide relevant operational definitions of the universe tested. The practice of content validation was defined in terms of the degree to which the test corresponded to the operational definition of the domain tested:

> ... a definition of the performance domain of interest must always be provided by a test user so that the content of a test may be checked against an appropriate task universe ... In defining the content universe, a test developer or user is accountable for the adequacy of his definition. An employer cannot justify an employment test on grounds of content validity if he cannot demonstrate that the content universe includes all, or nearly all, important parts of the job. (pp. 28-29)

This excerpt illustrates the contention that, like validity in general, content validity is not a unique feature of a test. Rather, a test's content is valid only with respect to a given purpose. The 1974 revision maintained the general descriptions of content validity promulgated in 1966. However, content validation was described in operational, rather than theoretical terms. The 1974 *Standards* clearly separated concerns of content validity from those of construct and criterion-related validities:

> The definition of the universe of tasks represented by the test scores should include the identification of that part of the content universe represented by each item. The definition should be operational rather than theoretical, containing specifications regarding classes of stimuli, tasks to be performed and observations to be scored. The definition should not involve assumptions regarding the psychological processes employed since these would be matters of construct rather than of content validity. (p. 48)

The 1974 *Standards* endorsed the notion that evaluations of test content should focus on the test's representation of the content domain as defined in the test blueprint. However, because evaluation of the psychological processes measured by test content was now described as purview to only construct validity, the notion that content validity was a separate, but equal, form of validity was severely undermined. Test specialists began to refrain from referring to content validity as a "type" of validity and began to regard construct validity as the most general and complete form of validity.

After publication of the 1974 *Standards*, two schools of thought prevailed regarding validation theory: one school promulgating the idea that validity consisted of three "separate but equal" aspects; the other school advocating a unitary conceptualization centered on construct validity. Proponents of the unitary conceptualization (e.g., Messick, 1975) disqualified content validity as a "type" of validity because validity referred to inferences derived from test scores, rather than from the test itself. The unitary conceptualization of validity also dismissed

criterion-related validity as a separate type of validity. It argued that in criterion-related studies, information is gained about both the test and the criterion. Because no one criterion is sufficient for the validation of a test, and because criteria must also be validated, criterion-related studies were only a part of the larger process of construct validation.

Further challenges to content "validity"

Messick (1975, 1980, 1988, 1989a, 1989b) argued strongly for a unitary conceptualization of validity. He asserted that different forms of *evidence* of validity do not constitute different *kinds* of validity. While he maintained that different types of inferences derived from test scores may require different forms of evidence, he repudiated labeling these forms of evidence "validity." Messick (1980) described construct validity as validity in general and asserted that its specific facets should be differentiated from the general concept:

> ... we are not very well served by labeling different aspects of a general concept with the name of the concept, as in criterion-related validity, content validity, or construct validity, or by proliferating a host of specialized validity modifiers ... The substantive points associated with each of these terms are important ones, but their distinctiveness is blunted by calling them all 'validity'. (p. 1014)

Messick (1975, 1980) recommended use of the terms "content relevance," "content representation," or "content coverage" to encompass the intentions associated with the term content validity. Similarly, he recommended that "criterion relatedness" replace the term "criterion validity." Messick's (1988, 1989a, 1989b) formulation of validity also called for validating the value implications and social consequences that result from testing. He asserted that this "consequential basis" of test interpretation and use also fell under the rubric of construct validation.

Guion (1977) supported Messick's (1975) contention that concerns of test content should not be denoted "validity," and recommended the terms "content representativeness" and "content relevance" for describing a test's congruence to the domain tested. Content representativeness referred to how well the test content sampled the universe of content and how well the response options sampled the universe of response behaviors. Content relevance referred to the congruence between the test content and the purpose of the testing. Guion (1977) did not condone accepting a test as valid based solely on an evaluation of its content; however, he proposed five conditions that would support the content representativeness and relevance of a test:

> First: The content domain must be rooted in behavior with a generally accepted meaning ... Second: The content domain must be defined unambiguously ... Third: The content domain must be relevant to the testing ... Fourth: Qualified judges must agree that the domain has been adequately sampled ... [and] Fifth: The response content must be reliably observed and evaluated. (pp. 6-8)

The assessment of the ability of a test to satisfy these conditions has become a principal goal in appraising test content. However, Guion pointed out that even when all these conditions are met, the test cannot be considered validated. He asserted that content representativeness was a necessary, but not sufficient, condition for validity.

Guion (1980) modified the conditions for evaluating content relevance for employment testing applications. These modifications resulted in a four-step process that "would assure a work sample test of unquestionable job relevance" (p. 391). The four-step process involved: 1) defining a *job content universe*, 2) identifying the *job content domain*, 3) defining a *test content universe*, and 4) defining a *test content domain*. Guion asserted that content relevance involved representative sampling from both a universe of content and a universe of potential tasks used to measure the content. The job content domain and test content domain were operational definitions of these universes, and content validation was the assessment of the adequacy of these operational definitions. Guion also suggested that a universe of scoring procedures be identified when evaluating test content.

Like Guion, Tenopyr (1977) advocated a process-oriented conception of content assessment. She argued that because all tests intended to measure constructs, content validity was not "validity," but rather an assessment of the test construction process. She stated that

> The obvious relationship between content and construct validity cannot be ignored; however, content and construct validity cannot be equated ... Content validity deals with inferences about test construction; construct validity involves inferences about test *scores*. (p. 50)

Tenopyr asserted that if the test construction process was to be adduced as evidence of "validity," then the process must focus on "well-defined constructs with easily observable manifestations" (p. 54).

The writings of Messick, Guion, and Tenopyr, indicated that although content representation was not to be considered a form

of validity, it was still a necessary goal of the test construction process. In keeping with this notion, Fitzpatrick (1983), admonished use of the term content validity, but described four "prevailing notions" of content representativeness desirable in test construction: domain sampling, domain relevance, domain clarity, and technical quality in test items. Fitzpatrick asserted that evaluation of these desirable characteristics need not be labeled "validation." Rather, she labeled the adequacy of domain sampling as "content representativeness," and the relevance of test content as "content relevance." Her evaluation of the usefulness of the concept of content validity led her to conclude that "... content validity is not a useful term for test specialists to retain in their vocabulary". (p. 11)

Although most test specialists were critical of the term "content validity, they continued to support the fundamental principles comprising this concept. Loevinger (1957), for example, stated that "... considerations of content alone are not sufficient to establish validity even when the test content resembles the trait, [but] considerations of content cannot be excluded when the test content least resembles the trait" (p. 657). Similarly, Fitzpatrick (1983) pointed out that "fit between a test and its definition appears important to establish, but it is not a quality that should be referred to using the term 'content validity'" (p. 6). Finally, as asserted by Messick (1989a),

> ... so-called content validity does not qualify as validity at all, although such considerations of content relevance and representativeness clearly do and should influence the nature of score inferences supported by other evidence (p. 7).

These views are evident in contemporary conceptualizations of validity (e.g., Angoff, 1988; Cronbach, 1988; Geisinger, 1992; Shepard, 1993), which demonstrate that the fundamental principles underlying content validity have persevered. In fact, the most recent version of the *Standards for Educational and Psychological Testing* (1985), while emphasizing a unitary conceptualization of validity, retained the importance of content domain representation. In this version, the "aspects" of validity denoted in the 1971 *Standards* were described as "categories" of validation. This modification of terminology changed the phrasing from "content validity" to "content-related evidence of validity," which "demonstrates the degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content" (p. 10).

A Re-formulation of Test Validity

The preceding literature review illustrated the historical roots of convergent and divergent theories of test validity. A conspicuous area of convergence is the claim that adequately defining and representing the construct measured is of critical importance. However, there is considerable divergence regarding the terminology used to describe this process. Given the fact that use of the term content validity (e.g. Cureton, 1951) preceded its formal definition (Lennon, 1956), it is not surprising that this term has been controversial since its inception.

In this section, the arguments surrounding validity nomenclature are temporarily forestalled to more fully examine the concept of content validity. After describing the components that comprise what until 1975 was termed content validity, its relationship within the unitary conceptualization of validity is discussed.

Defining content validity

As demonstrated in the literature review, content validity can be described either broadly or narrowly. Broad definitions assert that content validity is concerned with test and response properties, whereas narrow definitions limit content validity to investigations of items, tests, and perhaps, scoring procedures. The broad and narrow conceptualizations can be contrasted by excerpting from and Lennon (1956) and Messick (1989b):

> ... content validity ... denote[s] *the extent to which a subject's responses to the items of a test may be considered to be a representative sample of his responses to a real or hypothetical universe of situations which together constitute the area of concern to the person interpreting the test.* (Lennon, 1956; p. 295)

> Content validity is based on professional judgements about the relevance of the test content to the content of a particular behavioral domain of interest and about the representativeness with which item or task content covers that domain. Content validity as such is not concerned with response processes, internal and external test structures, performance differences and responsiveness to treatment, or with social consequences. (Messick, 1989b; p. 17)

Messick's narrow description of content validity supports his unitary conceptualization of validity centered on construct validity. In so doing, he attributes some of the qualities ascribed to content validity in Lennon's formulation to the purview of construct validity. These two exemplar definitions differ with respect to the "domain" of content validity. However, all descriptions of the desirable content qualities of a test

include four critical elements: domain representation, domain relevance, domain definition, and appropriate test construction procedures. As demonstrated in the literature review, these four elements define the concept of content validity. Table 1 presents these fundamental elements along with some of the test specialists who acknowledged their importance to test validity.

Hypothetical constructs versus hypothetical content domains

Divergent definitions of content validity stem from long-standing ambiguity regarding what is a "construct" and what is a "content domain". Distinctions between constructs and content domains have typically been made on the basis of tangibility; constructs are described as unobservable and undefinable, and content domains are characterized as observable and definable. In fact, some descriptions of content domains equate the content domain with the test specifications governing the test construction process. However, test specifications represent an operational definition of the content domain, not the domain itself. Hence, test specifications are tangible and observable, but the content domains they describe are not. Such construct/content confusion was noted in the 1985 *Standards*:

> ... methods classed in the content-related category thus should often be concerned with the psychological construct underlying the test as well as the character of the test content. There is often no sharp distinction between test content and test construct." (p. 11)

Messick (1989b) expounded on this excerpt noting that "the word *often* should be deleted from both of the quoted sentences--that as a general rule, content-related inferences and construct-related inferences are inseparable" (p. 36). Messick asserted that a conceptualization of the domain tested must be made "in terms of some conceptual or construct theory of relevant behavior" (p 37). Thus perhaps in the Messickian view, "content domain" is synonymous with "construct." Although the distinction between the two concepts is not explicitly discussed by Messick, a close reading suggests that he relates content domains to test-specific behaviors, and constructs to both test and non-test behaviors. Given this view, constructs and content domains are similar in that they are both latent and unobservable; however, they differ in level of abstraction.

Content validity within the construct validity framework

To illuminate the distinction between constructs and content domains, their theoretical relationship is depicted in Figure 1. The two dimensions framing the relationship are "abstract versus tangible" (vertical) and "unobservable versus observable" (horizontal). As indicated in the figure, the construct is the most abstract concept followed by the content domain. These two

Table 1

Selected Publications Defining Content Validity

Domain Representation	Domain Relevance	Domain Definition	Test Construction Procedures
Mosier (1947)	Rulon (1946)	Thorndike (1949)	Loevinger (1957)
Goodenough (1949)	Thorndike (1949)	APA (1952)	Ebel (1961)
Cureton (1951)	Gulliksen (1950a)	Lennon (1956)	AERA/APA/NCME (1966)
APA (1952)	Cureton (1951)	Ebel (1961)	Nunnally (1967)
AERA/APA/NCME (1954)	AERA/APA/NCME (1954)	AERA/APA/NCME (1966)	Cronbach (1971)
Lennon (1956)	AERA/APA/NCME (1966)	Cronbach (1971)	Guion (1977,80)
Loevinger (1957)	Cronbach (1971)	AERA/APA/NCME (1974)	Tenopyr (1977)
AERA/APA/NCME (1966)	Messick (75,80,88,89a,89b)	Guion (1977,80)	Fitzpatrick (1983)
Nunnally (1967)	Guion (1977,80)	Tenopyr (1977)	AERA/APA/NCME (1985)
Cronbach (1971)	Fitzpatrick (1983)	Fitzpatrick (1983)	
AERA/APA/NCME (1974)	AERA/APA/NCME (1985)		
Messick (1975,80,88,89a,89b)			
Guion (1977,80)			
Fitzpatrick (1983)			
AERA/APA/NCME (1985)			

unobservable entities are mediated by test specifications, which operationally define the content domain. The arrow emanating from the construct to the test specifications demonstrates that conception of the construct influences the development of test specifications, as do political and logistic factors, such as the testing purpose. The content/construct relationship presented here is congruent with Shepard (1993) who claimed "content domains for constructs are specified logically by referring to theoretical understandings, by deciding on curricular goals in subject matter fields, or by means of a job analysis" (p. 413).

The test construction tasks depicted under the "observable" column in Figure 1 summarize the elements comprising content validity. The arrows connecting the elements of the figure are labeled according to their role within a "content validity" framework. Similar to the distinction between constructs and content domains, the "observable" elements of Figure 1, also differ in level of abstraction. Test specifications, item pools, tests, test scores, and item responses are all observable, but each of these test construction products is one step further from a concrete representation of the construct.

Although Figure 1 attempts to provide a theoretically correct depiction of the relationship between these concepts, other variations are possible. For example, the association between "item responses" and "test scores" could be transposed (however, the derivation of scores from item responses requires carefully-developed rules and algorithms). Furthermore, not all steps in the test construction process are included in Figure 1. An obvious omission is work conducted to help create test specifications, such as job analyses or reviews of curricula. Thus the relationships depicted in Figure 1 are not claimed to be absolute. Rather, they are presented to emphasize that content validity is fundamental to construct validity, and constructs and content domains are both hypothetical entities mediated by attempts to operationally define the construct measured. These relationships are illustrated more concretely in Figure 2, where they are applied to a hypothetical mathematics achievement test.

It should be noted that the theoretical relationship between construct and content validity presented here is consistent with the Messickian view of validity. Messick has continually pointed out that elements of content validity are an integral part of the construct validity framework. What is different in the "re-formulation" here is the emphasis that content validity is a *necessary component of construct validity*. Similar to test score reliability, content validity sets a lower bound for construct validity. Without defensible test content (i.e., adequate definition of the content domain, relevant and technically correct items and tasks, and sound test construction procedures), construct validity cannot be achieved. This view is congruent with Shepard (1993) who commended Messick's unitary

Figure 1: Theoretical Relation Between Construct And Content Validity

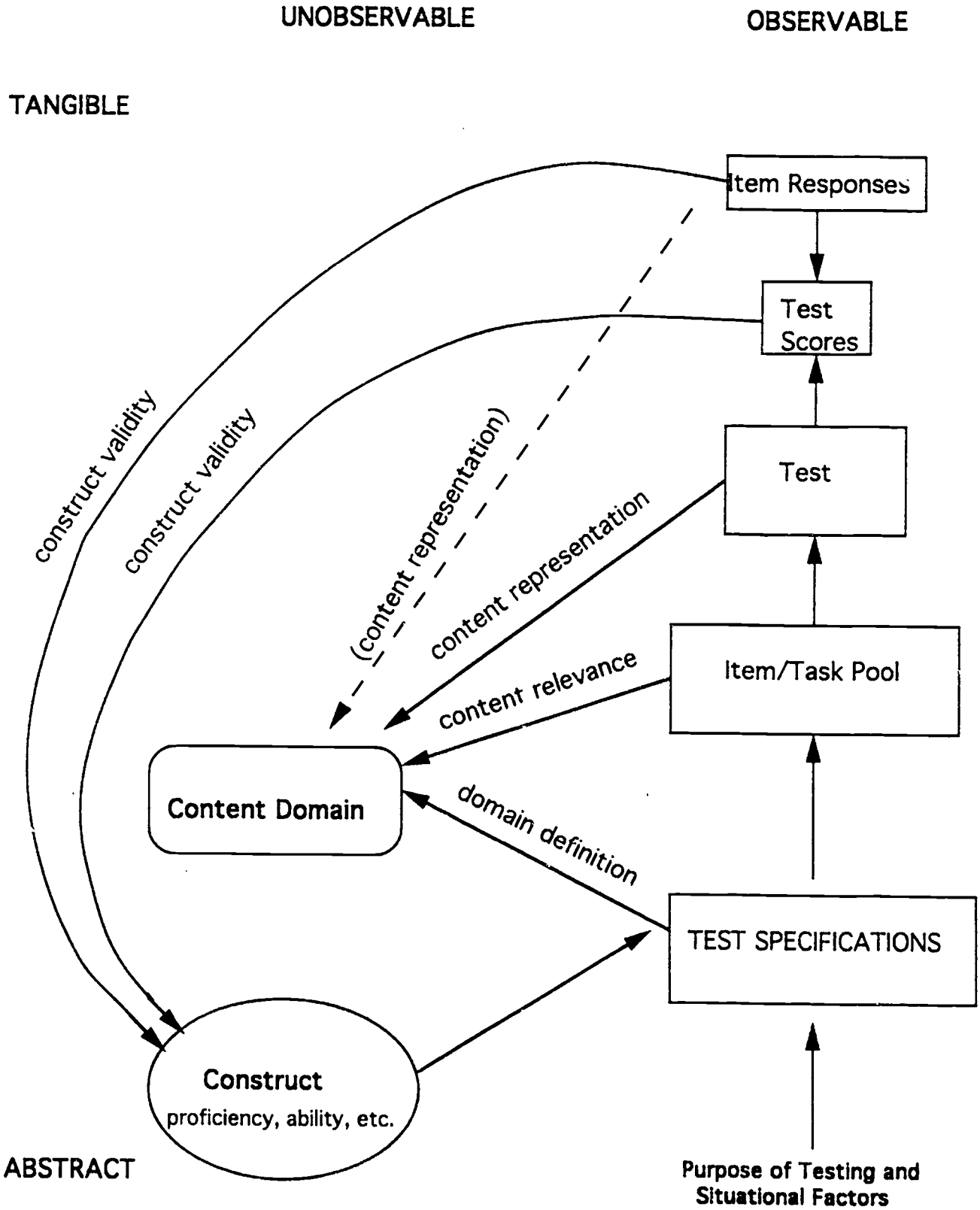
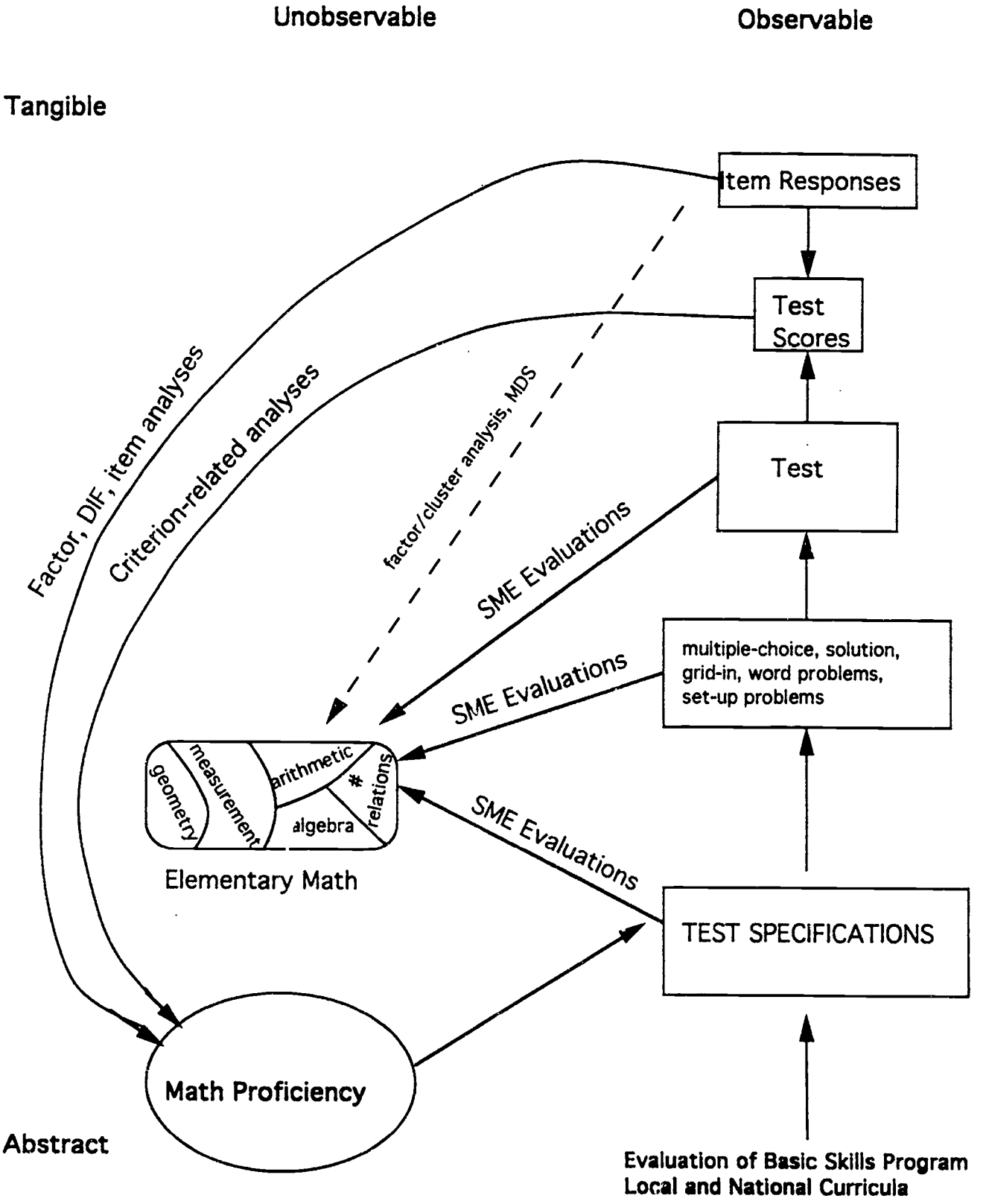


Figure 2: Construct/Content Relation For Hypothetical Math Test



Note: SME=Subject Matter Expert

conceptualization of validity for stressing the importance of testing consequences, but stated that "consequences [are] equal contenders alongside domain representativeness as candidates for what must be assessed in order to defend test use" (pp. 429-430).

The claim that the elements comprising content validity are necessary qualities for construct validity is not revolutionary. In fact, Messick (1989b) argued the same point in his unitary conceptualization of validity:

> It is not enough that construct theory should serve as a rational basis for specifying the boundaries and facets of the behavioral domain of reference. Such specifications must also entail sufficient precision that items or tasks can be constructed or selected that are judged with high consensus to be relevant to the domain. (p. 38)

Although Messick did not state explicitly that content domain representation (and the other aspects of content validity) was necessary to achieve construct validity, such a notion is consistent with his unitary conceptualization of validity.

The point here is that the central role of content validity within the construct validity framework cannot be ignored. For if the sample of tasks comprising a test is not representative of the content domain tested, the test scores and item response data used in studies of construct validity are meaningless. As Ebel (1977) succinctly put it "data never substitute for good judgment" (p. 59).

The assertion that content validity is a necessary component of construct validity should not be misconstrued as implying that a test can be validated based on content analyses alone. Evidence of content validity does not provide sufficient evidence for validating inferences derived from test scores. As accurately asserted by Messick, Fitzpatrick, Guion, Shepard, and others, the elements of content validity do not signify validity. Thus, content validity is a necessary, not sufficient, condition for construct validity (Guion, 1977).

An issue remaining to be resolved is whether analyses focusing on item responses are germane to content validity. Responses to test items are clearly instrumental to construct validity. However, task responses are also used to evaluate content domain representation (e.g., Jackson, 1974). Although this issue is equivocal, it is reasonable to maintain that response properties fall under the purview of both construct and content validity, as illustrated by the dotted lines in Figures 1 and 2. However, this unresolved issue is one of nomenclature, which is likely to be of little importance to practitioners who will carry out these procedures, regardless of the labels theorists use to describe them (Ebel, 1977). For example, Embretson (1983) introduced the

term "construct representation" to describe the process of using item response data to describe what a test measures.

Retaining the term content validity

The term content validity can be used to describe the family of test construction and validation procedures pertaining to measurement of the underlying content domain. It describes essential processes for defending score interpretations with respect to the content domains (and constructs) presumably measured. The terms "content representation," "content relevance," and "domain definition" are discrete. Therefore, use of the term content validity to describe these qualities, as well as to refer to appropriate test construction procedures, provides parsimony. Use of the term content validity does not undermine the unitary conceptualization of validity. Thus contrary to Fitzpatrick (1983), it is affirmed here that content validity is a useful term, and should be retained in the vocabulary of test practitioners and theorists.

Is content validity applicable to all tests?

It was asserted above that content validity is an essential component of test validity. However, content validity has historically been attributed to the domain of educational, industrial, and licensure testing. Therefore, the question remains whether content validity is applicable to psychological tests (e.g., personality tests). Like educational tests, psychological tests purport to measure one or more underlying constructs. These constructs must be defined operationally, prior to the development of tests designed to measure them. Such operational definitions evoke content domains. Therefore, content validity is applicable to all tests purporting to measure an underlying construct. As Anastasi (1986) noted:

> So-called content validation and criterion-related validation can be more appropriately regarded as stages in the construct validation of all tests... validation extends across the entire test construction process; it encompasses multiple procedures employed sequentially at appropriate stages. Validity is built into a test at the time of initial construct definition and the formulation of item-writing specifications... (pp. 12-13)

The Practice of Content Validation

The preceding sections defined content validity and argued that it is a fundamental part of construct validity. This section briefly describes traditional and contemporary procedures used to advance and evaluate content validity.

Procedures used to evaluate test content can be classified

generally as subjective or empirical. Subjective methods refer to studies where subject matter experts (SMEs) are used to evaluate test items and rate them according to their relevance and representativeness to the content domain tested. Empirical methods refer to those procedures that analyze the data obtained from administering the test (test and item score data).

Subjective methods for evaluating test content

Crocker, Miller, and Franks (1989) and Osterlind (1989) reviewed subjective methods for evaluating test content. All methods reviewed provided an index reflecting the degree to which the content of the test held up under the scrutiny of SMEs. Two commonalities existed among the different content indices reviewed. First, each procedure provided at least one quantitative summary of subjective data gathered from SMEs. Second, the SMEs used in each procedure rated each test item in terms of its relevance and/or match to specified test objectives. The major differences between the methods reviewed were in the specific instructions given to the SMEs, and whether or not an item was allowed to correspond to more than one objective.

Two methods for quantifying the judgments made by SMEs are provided by Hambleton (1980, 1984) and Aiken (1980). Hambleton (1980) proposed an "item-objective congruence index" designed for criterion-referenced tests where each item is linked to a single objective. This index reflected SMEs' ratings, along a three-point scale, of the extent to which an item measured its specified objective, versus the extent to which it was linked to the other test objectives. Later, Hambleton (1984) provided a variation of this procedure designed to reduce the demand on the SMEs. He also suggested a more straightforward procedure where SME ratings of item-objective congruence could be measured along longer Likert-type scales. Using this procedure, the mean congruence ratings for each item, averaged over the SMEs, provided a straightforward, descriptive index of the SMEs' perceptions of the item's fit to its designated content area.

Aiken's (1980) index also evaluates an item's relevance to a particular content domain, using SMEs relevance judgments. His index takes into account the number of categories on the scale used to rate the items and the number of SMEs conducting the ratings. The statistical significance of the Aiken index is evaluated by computing a normal deviate (z-score) and its associated probability.

Like other subjective methods used to evaluate test content (c.f. Lawshe, 1975; Morris & Fitz-Gibbon, 1978), Hambleton's and Aiken's methods provide SME-based indices of the overall content quality of test items. The individual item indices can also be averaged to provide a global index of the overall content quality of a test. Popham (1992) reviewed applications of SME-based

indices of content quality for teacher licensure tests. His review illustrated that variation in the rating task presented to SMEs affected their judgments. He noted that criteria for determining whether content representation was obtained were not available, and so he called for further research to establish standards of content quality based on SME evaluations.

Factor and MDS analyses of item ratings

Two additional SME-based methods used to investigate content validity were proposed by Tucker (1961), and Sireci and Geisinger (1992; in press). Tucker factor-analyzed SME ratings regarding the relevance of test items to the content domain tested. Two interpretable factors were related to test content. The first factor was interpreted as "a measure of general approval of the sample items" (p. 584), and the second factor was interpreted as revealing two schools of thought among the SMEs as to what kinds of items were most relevant ("recognition" items or "reasoning" items). Tucker concluded that factor analysis of SME relevance ratings was appropriate for identifying a test with high content validity and for identifying differences in opinion among SMEs.

Sireci and Geisinger (1992; in press) used multidimensional scaling (MDS) and cluster analysis to evaluate SMEs' ratings of item similarity. This procedure was used to avoid informing the SMEs' of the content specifications from which the tests were derived. The rationale underlying the procedure was that items comprising the content areas specified in the test blueprints would be perceived as similar to one another by the SMEs (with respect to the content measured) and would cluster together in the MDS space. Items that comprised different content areas would be perceived as less similar and would not group together. The results illustrated that MDS and cluster analysis of SME item similarity ratings provided both convergent and discriminant evidence of the underlying content structure of a test. For example, in their analysis of a social studies achievement test, Sireci and Geisinger (in press) discovered a distinction between items measuring U.S. history and those measuring world history, which was not specified in the test blueprint.

Empirical evaluations of test content

Most empirical methods for evaluating test content do not employ subjective opinion in the analyses, and so the problem of bias or error in SMEs' ratings is avoided. Empirical investigations of test content include applications of MDS and cluster analysis (Napior, 1972; Oltman, Stricker, and Barrows, 1990), and applications of factor analysis (Dorans & Lawrence, 1987; Jackson, 1974). These investigations uncover dimensions, factors, and clusters presumed to be relevant to the content domains measured. However, interpretation of the results can be problematic, especially when response properties of the data

confound content interpretations (Green, 1983; Davison, 1985).

Both empirical and subjective analysis of test content provide important information regarding content and construct validity. However, both approaches have limitations. Sireci and Geisinger (1992; in press) recommend using both types of analyses to fully evaluate content domain definition and representation.

Conclusions: The Future of Content Validity

The discussion of validity presented in this paper focused on validity issues related to test content. These issues are much narrower in scope than those discussed by Messick, Shepard, and others, in describing construct validity and test validation. Formulations of the unitary conceptualization of validity center on fairness issues related to inferences derived from test scores. By framing "test" validity within the context of values and consequences, the unitary conceptualization broadened the agenda surrounding test equity. Validation centered on the unitary conceptualization requires test developers and users to go beyond demonstrating the validity of a test for a particular purpose. It also requires that the unintended consequences of testing, and associated societal values, be considered.

An unfortunate consequence of the unitary conceptualization of validity is the lack of attention paid to test content. As forewarned by Yalow and Popham (1983) "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). With a few notable exceptions (e.g., Popham, 1992; Sireci & Geisinger, 1992; Smith, Hambleton, & Rosen, 1988) a perusal of recent measurement journals and test publisher's technical manuals reveals a paucity of research and practice in the area of content validation.

Current developments in educational testing invoke a renewed emphasis on evaluating the quality of test content. For example, computerized testing and item selection algorithms threaten representation of the content domain if item selection decisions are based solely on statistical indices of item quality (e.g., item difficulty, item discrimination). A related example is increasing use of the Rasch model for test development. Due to its assumption of equal discrimination among the test items, the Rasch model is not likely to be appropriate for tests measuring heterogeneous content domains. Thus increasing emphasis on statistical criteria for item selection may result in limited representation of the content domain (cf. Wainer & Lewis, 1990).

The resurgence of "authentic" assessments, which strive to represent the constructs measured more accurately (Linn, 1994), also invokes a renewed emphasis on content validity. This resurgence, coupled with advances in assessment technology (e.g.,

interactive video), will yield new types of tests that must be justified with respect to construct representation. Furthermore, recent proposals for more flexible item writing guidelines (Popham, 1994, 1995) necessitate more thorough evaluation of the content quality of tests (and items) vis-a-vis the constructs measured. Therefore, future descriptions of test validity must emphasize the necessary and important role of content validity in the construction and evaluation of tests.

In particular, the forthcoming revision of the *Standards For Educational and Psychological Testing* should emphasize the central role of content validity in test construction and evaluation. Specifically, the revised *Standards* should acknowledge that:

1. Content validity refers to a family of issues and procedures fundamental for evaluating the validity of tests and of inferences derived from test scores.

2. Content validity is a necessary, but not sufficient, requirement for construct validity.

3. Content validity is a fundamental requirement for all tests. It is relevant in psychological and personality testing, as it is in educational and industrial testing.

4. Comprehensive procedures exist for evaluating test content and for assisting test developers in their pursuit of content-valid tests.

These recommendations emphasize the importance of content validity within the unitary, construct validity framework. This emphasis is similar to that advanced by Shepard (1993) who noted:

> The consensus, already emergent before the 1985 standards, has been solidified. Construct validation is the one unifying and overarching framework for conceptualizing validity evaluations. Logical analysis of test content and empirical confirmation of hypothesized relationships are both essential to defending the validity of test interpretations; however, neither is sufficient alone. (p. 443)

Thus it is concluded that the theories and applications of content validity have a home within the unifying framework of construct validity, as well as within other conceptualizations of validation such as Kane's (1992) argument-based approach. Future descriptions of construct validity should emphasize its critical elements that comprise the concept of content validity. As both historical and current practice of test validation demonstrates, if the content of a test cannot be defended with respect to the use of the test, construct validity cannot be obtained.

References

Aiken, L.R. (1980). Content validity and reliability of single items or questionnaires. Educational and Psychological Measurement, 40, 955-959.

American Psychological Association, Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. American Psychologist, 7, 461-465.

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

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

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

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.

Anastasi, A. (1954). Psychological testing. New York: MacMillan.

Anastasi, A. (1986). Evolving concepts of test validation. Annual Review of Psychology, 37, 1-15.

Angoff, W.H. (1988). Validity: An evolving concept. In H. Wainer & H.I. Braun (Eds.), Test validity (pp. 19-32). Hillsdale, New Jersey: Lawrence Erlbaum.

Bingham, W.V. (1937). Aptitudes and aptitude testing. New York: Harper.

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

Cronbach, L.J. (1971). Test Validation. In R.L. Thorndike

(Ed.) Educational measurement (2nd ed., pp. 443-507). Washington, D.C.: American Council on Education.

Cronbach, L.J. (1988). Five perspectives on the validity argument. In H. Wainer & H.I. Braun (Eds.), Test validity (pp. 3-17). Hillsdale, New Jersey: Lawrence Erlbaum.

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

Cureton, E.E. (1951). Validity. In E.F. Lindquist (Ed.), Educational measurement (1st ed., pp. 621-694).

Davison, M.L., (1985). Multidimensional scaling versus components analysis of test intercorrelations. Psychological Bulletin, 97, 94-105.

Dorans, N.J. & Lawrence, I.M. (1987). The internal construct validity of the SAT. (Research Report). Princeton, NJ: Educational Testing Service.

Ebel, R.L. (1961). Must all tests be valid? American Psychologist, 16, 640-647.

Ebel, R.L. (1977). Comments on some problems of employment testing. Personnel Psychology, 30, 55-63.

Embretson (Whitley), S. (1983). Construct validity: construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

Fitzpatrick, A.R. (1983). The meaning of content validity. Applied Psychological Measurement, 7, 3-13.

Geisinger, K.F. (1992). The metamorphosis in test validity. Educational Psychologist, 27, 197-222

Goodenough, F.L. (1949). Mental testing. New York: Rinehart.

Green, S.B. (1983). Identifiability of spurious factors with linear factor analysis with binary items. Applied Psychological Measurement, 7, 3-13.

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

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

Guion, R.M. (1978). Scoring of content domain samples: the problem of fairness. Journal of Applied Psychology, 63, 499-506.

Guion, R.M. (1980). On trinitarian doctrines of validity. Professional Psychology, 11, 385-398.

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

Gulliksen, H. (1950b). Theory of mental tests. New York: Wiley.

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: Johns Hopkins University Press.

Hambleton, R.K., (1984). Validating the test score In R.A. Berk (Ed.), A guide to criterion-referenced test construction. Baltimore: Johns Hopkins University Press, pp. 199-230.

Jackson, D.N. (1974). Personality Research Form : Manual. Port Huron, MI: Research Psychologists Press.

Jenkins J.G., (1946). Validity for what? Journal of Consulting Psychology, 10, 93-98.

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

Kelley, T.L. (1927). Interpretation of educational measurement. Yonkers-on-Hudson, NY: World Book Co.

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

Lennon, R.T. (1956). Assumptions underlying the use of content validity. Educational and Psychological Measurement, 16, 294-304.

Lindquist, E.F. (Ed.). (1951). Educational measurement. Washington, D.C.: American Council on Education.

Linn, R.L. (1994). Criterion-referenced measurement: a valuable perspective clouded by surplus meaning. Educational Measurement: Issues and Practice, 13, 12-15

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694 (Monograph Supplement 9).

Messick, S. (1958). The perception of social attitudes. Journal of Abnormal and Social Psychology, 52, 57-66.

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. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braun (Eds.), Test validity (pp. 33-45). Hillsdale, New Jersey: Lawrence Erlbaum.

Messick, S. (1989a). Meaning and values in test validation: the science and ethics of assessment. Educational Researcher, 18, 5-11.

Messick, S. (1989b). Validity. In R. Linn (Ed.), Educational measurement, (3rd ed.). Washington, D.C. American Council on Education.

Morris, L.L., and Fitz-Gibbon, C.T. (1978). How to measure achievement. Beverly Hills: Sage.

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

Napior, D. (1972) Nonmetric multidimensional techniques for summated ratings. In Shepard, R.N.; Romney, A.K.; and Nerlove S.B. (eds.), Multidimensional scaling: Volume 1: Theory. New York: Seminar Press.

Nunnally, J.C. (1967). Psychometric theory. New York: McGraw-Hill.

Oltman, P.K., Stricker, L.J., and Barrows, T.S. (1990). Analyzing test structure by multidimensional scaling. Journal of Applied Psychology, 75, 21-27.

Osterlind, S.J. (1989). Constructing test items. Hingham, MA: Kluwer.

Popham, W.J. (1994). The instructional consequences of criterion-referenced clarity. Educational Measurement: Issues and Practice, 13, 15-20;39.

Popham. W.J. (1995). Postcursive review of criterion-referenced test items based on "soft" item specifications. A symposium presented at the annual meeting of the National Council on

Measurement in Education, San Francisco, April 20.

Rulon, P.J. (1946). On the validity of educational tests. Harvard Educational Review, 16, 290-296.

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

Sireci, S.G., and Geisinger, K.F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. Applied Psychological Measurement, 16, 17-31.

Sireci, S.G., & Geisinger, K.F. (in press). Using subject matter experts to assess content representation: a MDS analysis. Applied Psychological Measurement.

Smith, I.L., Hambleton, R.K., & Rosen, G.A. (1988). Content validity studies of the examination for professional practice in psychology. Paper presented at the annual convention of the American Psychological Association, Atlanta, GA.

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

Thorndike, E.L. (1931). Measurement of intelligence. New York: Bureau of Publishers, Columbia University.

Thorndike, R.L. (1949). Personnel selection: Test and measurement techniques. New York: Wiley.

Thorndike, R.L. (Ed.). (1971). Educational measurement (2nd ed.). Washington. D.C.: American Council on Education.

Thurstone, L.L. (1932). The reliability and validity of tests. Ann Arbor, Michigan: Edwards Brothers.

Toops, H.A. (1944). The criterion. Educational and Psychological Measurement, 4, 271-297.

Tucker, L.R. (1961). Factor analysis of relevance judgments: an approach to content validity. In A. Anastasi (Ed.) Testing problems in perspective: Twenty-fifth anniversary volume of topical readings from the invitational conference on testing problems, (pp. 577-586) Washington, D.C.: American Council on Education (1966).

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27, 1-14.

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