The Meaning of Content Validity

Anne R. Fitzpatrick
University of Massachusetts, Amherst

The ways in which test specialists have defined content validity are reviewed and evaluated in order to determine the manner in which this validity might best be viewed. These specialists have differed in their definitions, variously associating content validity with (1) the sampling adequacy of test content, (2) the sampling adequacy of test responses, (3) the relevance of test content to a content universe, (4) the relevance of test responses to a behavioral universe, (5) the clarity of content domain definitions, and (6) the technical quality of test items. After the theoretical and practical soundness of defining content validity in terms of each of these notions is evaluated, it is concluded that these notions are best regarded as definitions of concepts other than content validity. Since no appropriate means of defining this type of validity is therefore found, it is concluded that content validity is not a useful term for test specialists to retain in their vocabulary.

Guidelines on test development commonly state that test developers must show that their measures are content valid. The *Standards for Educational and Psychological Tests* (*Standards;* APA, AERA, & NCME, 1974) and most measurement texts indicate that norm-referenced achievement measures must be content valid (e.g., Anastasi, 1976; Brown, 1976). Evidence of content validity has also been described as essential for criterion-referenced measures of good quality (Hambleton & Novick, 1973; Millman, 1978). Finally, federal regulations have stated that content validity is an important property of measures used for professional certification and for employee selection and classification (EEOC, 1978; FEA, 1976).

Support for these guidelines can be found in literature on test validation, where test specialists have consistently noted the pertinence of content validity to the tests of performance and subject matter learning that are commonly administered in academic and employment settings (e.g., Anastasi, 1976; Brown, 1976; Dunnette & Borman, 1979; Glaser & Klaus, 1962; Shimberg, 1981). Responses to such tests are viewed as samples of some behavior (Goodenough, 1949), and content validity is said to play a major role in establishing that these samples are representative and interpretable.

Unfortunately, also found in literature on test validation is evidence that test specialists do not agree on the meaning of content validity. Specifically, authorities appear to differ in their views on (1) what features of a measure are evaluated under the rubric of content validity, (2) how content validity is established, and (3) what information is gained from study of this type of validity. For example, a test developer seeking advice from various sources about validating a reading test might be told either that content validity is based solely upon a logical study of test content (Aiken, 1979; Payne, 1974) or that it may entail empirical studies involving the scores of a measure (Anastasi, 1976).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 1, Winter 1983, pp. 3–13

0146-6216/83/010003-11$1.55

Perusing further, the test developer might be informed that content validity alone provides a sufficient basis for claiming the validity of a reading measure (Millman, 1978; Thorndike & Hagen, 1977), that content validity is necessary but not sufficient for establishing the validity of this measure (Linn, 1980), or that content validity is not any kind of validity at all (Gleser, 1969; Messick, 1975; Tenopyr, 1977).

The purpose of this article is to examine alternative perspectives on content validity in the interest of determining the manner in which this validity might best be viewed. Prevailing notions about content validity and the process of content validation are described as they pertain to the following four concepts: (1) domain sampling, (2) domain relevance, (3) domain clarity, and (4) technical quality in test items. Each notion is discussed in terms of its psychometric import, and the theoretical and practical soundness of using it as a definition of content validity is appraised.

Several terms are utilized below and merit some definition here. The terms "behavior" and "behavior domain" are used interchangeably to refer to any type of knowledge, skill, or performance that a test is claimed to assess. This behavior might be described either in somewhat broad, abstract terms (e.g., "understanding common fractions") or in somewhat narrow, concrete terms (e.g., "finding the sum of two one-digit integers"). In contrast, the term "content domain" is used to refer to any definition that is given of the procedures used to measure a behavior of interest. In its most detailed form, this definition might describe the content and structure of a test procedure and the rules for scoring the responses that are obtained. Because a test typically will be devised to assess several behaviors, several content domains often will be used to describe the items that should constitute a test. For example, if a test of first graders' skills in counting, addition, and subtraction were of interest, three domains that describe certain counting, addition, and subtraction tasks would be specified and used to define the items that should constitute the test. Finally, the terms "behavioral universe" and "content universe" are used to refer to realms of diverse behaviors and tasks, respectively, from which a test developer might draw the behaviors to be measured and the content to be covered by a test that is devised. For example, from the universe of skills taught in the nation's schools, a developer usually will draw the behaviors to be assessed by a standardized achievement test.

The Concept of Domain Sampling

Central to most test specialists' views of content validity is the general concept that this validity refers to the adequacy with which a test samples the domains that the test is claimed to cover. However, when the specific terms used to explain this concept are examined, it becomes clear that these specialists have interpreted and operationalized the concept in different ways.

The Test as a Content Sample

Some authorities have viewed content validity as relevant to test content issues and have indicated that this validity is dependent upon the adequacy with which the items of a measure constitute an adequate sample of the content domains that a test is claimed to cover (Cronbach, 1971; Linn, 1974, 1980; Loevinger, 1957; Messick, 1975). Using the term "universe" when referring to a content domain, Cronbach (1971) described this view when he stated,

> An achievement test is said to represent a body of content outlined in the test manual . . . To ask, "Are the tasks used in collecting data truly representative of the specified universe?" is to examine *content validity*. (p. 451)

To assess the adequacy of a test as a content sample, it has typically been suggested that content experts be asked to judge (1) how well each item of a test corresponds to the defined content domain that the item was written to reflect and (2) how well sets of items represent the content domains to which they are judged to correspond (e.g., Brown, 1976; Thorndike & Hagen, 1977). Measurement specialists have also indicated that these judgments should be gathered systematically by a test developer, and they have described several means for doing so (see Ebel, 1956; Polin & Baker, 1979; Rovinelli & Hambleton, 1977).

The Test as a Behavioral Sample

An alternative view taken by other test specialists relates content validity to the adequacy of test response rather than to test content samples. According to this view, content validity is dependent upon the extent to which responses to a test constitute an adequate sample of the behaviors that the test is designed to assess (Aiken, 1979; APA et al., 1974; Anastasi, 1976; Lennon, 1956; Mehrens & Lehmann, 1978). Given that responses to a test are usually symbolized by test scores, the most recent *Standards* (APA et al., 1974) stated this view well:

> To demonstrate the content validity of a set of test scores, one must show that the behaviors demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired performance domain. (p. 28)

Both logical and empirical procedures have been suggested as means of investigating this version of content validity. In the *Standards* it is said that experts' judgments of how well the items of a test correspond to and represent the test's content domains can be used to determine whether the test adequately samples the desired behaviors (APA et al., 1974). Other specialists have suggested that the items of a test be examined to determine whether they appear to call for a representative sample of the intended behaviors (Anastasi, 1976; Mehrens & Lehmann, 1978; Rozeboom, 1966). According to these specialists, studies of examinees' responses to the test should also be conducted, since analyses of test content alone may not reveal certain uncontrolled or unobservable factors that make test responses reflect behaviors other than the ones intended. Anastasi (1976), for example, recommended the use of certain correlational and experimental techniques to detect the irrelevant influences of reading ability and test speededness on math test performance.

Discussion

To evaluate these divergent views, it is best to first consider them in operational terms, since some similarities and differences between them then become clearer. It was noted above that some specialists have argued that content validity is concerned with the adequacy of a test as a content sample, whereas others have argued that this validity is concerned with the adequacy of test responses as a behavioral sample. Nevertheless, when the specialists who hold these opposing views say that this validity is established by judging the relation of test items to the domains of content the test is said to cover, these specialists are maintaining two perspectives of content validity that are operationally equivalent: According to both perspectives, content validity is operationally defined as the outcome of judging the sampling adequacy of test content. It is most parsimonious to regard the two perspectives as referring to the same concept, and they will be treated as such in the discussion that follows.

In contrast, consider the view taken by other specialists who have argued that content validity is concerned with the adequacy of test responses as a behavioral sample and that it is established by studying both test content and test responses. This view is operationally different from the first-mentioned set of perspectives and therefore can be said to represent a different concept of content validity. Accordingly, it will be examined separately from these perspectives.

Content validity based on judgments about the sampling adequacy of test content. Judgments of the sampling adequacy of test content can be thought of as one means of establishing the scientific soundness of a measure. These judgments indicate the degree to which the content domains of a test are represented by the items of the test; they thereby establish the fit between the definition of a measurement operation and the actual operation that is devised (Cronbach, 1971). It is a basic tenet of the empirical sciences that an operation must fit its definition if the operation is to be considered admissible as a scientific means of collecting data (Brodbeck, 1957; Peak, 1953).

In educational and psychological testing contexts, judgments of the sampling adequacy of a test's content may also serve to support particular test score interpretations that are of interest. For example, these judgments may provide grounds for claiming that examinees' scores on a test are reliable indicants of the true scores examinees would obtain on the domains covered by the measure (Cronbach, Gleser, Nanda, Rajaratnam, 1972; Nunnally, 1967). As Cronbach (1971) noted, when the items of a test are judged to adequately represent well-defined domains of content, it is permissible to view responses to these items as generalizable samples of the responses examinees would exhibit if they were tested on all of the items constituting these domains. Of course, there is some risk of drawing an erroneous conclusion when the generalizability of test responses is logically inferred on the basis of a content analysis, and so empirical confirmation of this generalizability is well advised (Cronbach et al., 1972). Nevertheless, the judgment that a set of items appears representative does provide some basis for inferring that examinees' performance on these items can be considered as a reliable estimate of their true domain scores. This is an inference that users of criterion-referenced tests, in particular, almost always wish to make (Hambleton, Swaminathan, Algina, & Coulson, 1978; Tomko, 1981).

The finding that a test's content is representative can also provide logical support for the interpretation of what is measured by a set of test scores. For example, when the words to be dictated in the spelling section of a third grade achievement test are judged to represent adequately the domains of words that have been taught to third graders, support is gained for the inference that third graders' scores on this test will reflect their levels of spelling skills. Alternatively, when the items of an algebra test are found to represent adequately elementary textbook problems involving linear algebra and logarithms but not those involving trigonometry, there is less basis for inferring that the total scores on this measure will reflect examinees' overall skill in solving elementary algebra problems.

Thus, 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." As Messick (1975) has suggested, test validity refers to the soundness of a test score interpretation; it is established by evidence that results from studies of test scores and shows the degree to which this interpretation is sound. Judgments of the sampling adequacy of test content result from studies of test content rather than test scores, so these judgments do not provide the kind of evidence that establishes the validity of an inference about the meaning of these scores. Hence, the association of these judgments with the term "content validity" is misleading, as it suggests that these judgments constitute a form of test validity. In accord with Messick (1975) and Gleser (1969), it is therefore recommended that the term "content representativeness" is better to use than the term "content validity" when referring to that which is established by judging the sampling adequacy of test content.

Content validity based on studies of test content and test scores. When content validity is said to refer to the adequacy of test responses as a behavioral sample and is operationally defined as the outcome of studies of test content and test scores, then it would seem that this validity is being viewed as pertinent to matters and methods that are central to investigations of construct validity (see also Guion, 1977). Of primary concern here is the matter of whether responses to a test reflect the behaviors that the test has been designed to assess; this is an inquiry about the descriptive meaning of test responses. To establish this meaning, studies of test scores are carried out to show that a proposed interpretation of these responses is supported by other data that are gathered. Construct validity is similarly concerned with the meaning of test responses and is similarly dependent on studies of these responses to establish this meaning (Messick, 1975, 1980).

Because this second view of content validity, as operationalized, is concerned with the meaning of test responses, it is recommended that it be referred to as a perspective on construct rather than content validity. The distinction of this view from the concept of construct validity is difficult to discern and, therefore, not worthwhile to maintain.

The appropriateness of regarding the issue of whether a test measures simple behaviors as a question of construct validity may well be questioned. Traditionally, construct validity has been thought pertinent only when test scores are to be viewed as indicants of some psychological quality such as "logical reasoning" or "anxiety" (Cronbach & Meehl, 1955; Ebel, 1977). However, as Cronbach (1971) has indicated, whenever situations, objects, or people are classified, a construct is being employed to form the class to which these elements belong. To be sure, the constructs used when test responses are classified as indicators of, say, "addition skills" or "performance on word meaning tasks" are much simpler and easier to verify than are those abstract terms that are used when it is suggested that test responses reflect psychological qualities such as "logical reasoning skills" or "anxiety" (cf. Ebel, 1977). Nevertheless, any interpretation of test responses can be said to entail the use of constructs. Thus, constructs are invoked by the claim that responses to a test reflect the particular behaviors that the test is intended to measure. To show the validity of this claim can be said to demonstrate its construct validity.

The Concept of Domain Relevance

A second quality seen by some test specialists as pertinent to the content validity of a measure is that of domain relevance. Perspectives on this quality diverge as they did on the issue of domain sampling, however, with some authorities associating the notion of domain relevance with matters related to test content and others associating this notion with matters related to test responses.

The Relevance of Test Content

Measurement textbooks and test development manuals have commonly indicated that a content valid test must cover important aspects of the content universe that a test user wishes to assess (e.g., APA et al., 1974; Thorndike & Hagen, 1977). They imply, therefore, that a measure's content validity depends, in part, upon whether the content domains that define a measure are relevant to the important parts of some universe of, say, academic or job content that interests a test user.

The recommended methods for assessing the relevance of a test's content domains have been judgmental in nature. For example, it is often suggested that the relevance of a standardized achievement measure be established prior to test construction by having content experts, experienced teachers, and curriculum experts agree upon the aspects of curricula that are important and should be covered by the test (APA et al., 1974; Anastasi, 1976; Brown, 1976). For an employment test, authorities have recommended that the content of a test be selected in light of what job analyses or the judgments of persons knowledgeable about a job indicate to be important tasks entailed in the job of interest (FEA, 1976; Glaser & Klaus, 1962; Miller, 1962). Authorities have also usually noted that the relevance of test content should be judged by an individual test user who, accordingly, should first identify the content universe he or she wishes to measure and then review the content covered by a test to appraise its content validity as a measure of this universe (APA et al., 1974; Brown, 1976).

The Relevance of Test Responses

Test specialists who have been concerned with the relevance of test responses rather than test content have suggested that a test should be considered content valid only to the extent that the behaviors assessed by the test reflect the behavioral universe that a test user wishes to assess (e.g., Cureton, 1951; Ebel, 1956; Green, 1981; Guion, 1978a, 1978b). Cureton (1951) had this meaning in mind when he discussed test "relevance," which he viewed as

> the degree to which the test operations as performed upon the test materials in the test situation agree with the actual operations as performed upon the actual materials in the situation normal to the task. (p. 622)

Guion (1978a, 1978b) and Lawshe (1975) implied concern for this issue when discussing the validity of employment tests. For example, Guion (1978b) expressed the view that an employment test can only be considered content valid when both test performance and scoring procedures are like the tasks and methods of evaluation experienced in the job to which the test is intended to pertain.

Both logical and empirical approaches to assessing this quality of response relevance have been suggested by these authorities. Ebel (1956) indicated that the items of a knowledge or skills test could be examined to determine the relevance of the behaviors assessed by the test to the behaviors that interest the test user. Guion (1978a) and Cureton (1951) stated that relevance could be logically inferred when the content and procedures entailed in a test were judged equivalent to the circumstances an examinee would face in the natural job or academic setting (see also Lawshe, 1975). Guion (1978a) and Cureton (1951) did caution, however, that irrelevant variables might influence test performance and that this possibility should be investigated through empirical studies of this performance. Cureton, for example, noted the use of criterion groups to examine the relation between a tested performance and performance on the job.

Discussion

The relevance of test content. When a person wishes to use a test to appraise performance on a particular content universe of interest, the relevance of the test's content to this universe will be important to determine. Many test users may wish to do this; teachers commonly wish to use tests to appraise students' learning of the material presented in a course, and personnel psychologists often may ask job applicants to perform some of the tasks entailed in a job. By establishing that the content of a test represents important aspects of the content universe that is of interest, a user can gain logical support for the claim that examinees' performance on the test will be indicative of their performance on this universe (Brown, 1976). For example, by finding that the items of a math test present all of the important kinds of math problems covered in a math course, a teacher will gain support for the claim that the test can be used to appraise students' learning of the course materials.

Although this matter of relevance has significance, in the interest of conceptual clarity it is suggested that it not be included under the content validity label. Judgments of the relevance of test content focus on test content rather than test scores, and they do not directly provide evidence that establishes the meaning of these scores. Because these judgments do not, then, have the aforementioned meaning that the term ''validity'' denotes, it would be better not to say that they reflect ''content validity.'' Instead, it can be said that these judgments reflect ''content relevance''; this term more clearly conveys the particular character and meaning of these judgments (see also Messick, 1980).

Under a content relevance label, information on how relevant a test is to some content universe might certainly be important for a developer to provide. Particularly when achievement and employment tests are constructed, a description of what universe was surveyed, and data indicating the consistency of experts' judgments about what were important aspects of this universe to include in a test, might be of interest and concern to the test user.

The relevance of test responses. The relevance of a set of test responses to a particular behavioral universe may also be important to show in some testing contexts. Most notably, this matter might have import when an employment test is to be used and it is desirable to infer, say, that individuals' performance on the test is a sample of the performance that they would show in the job to which the test is intended to pertain (Guion, 1978a, 1978b).

Although studies of the correspondence between the tasks entailed in a test and those entailed in a universe of interest may be informative with regard to response relevance, these studies cannot establish this relevance. As Cureton (1951) and Guion (1978a) cautioned, the finding that the tasks entailed in, say, an employment test are relevant to a particular universe of job content does not ensure that the responses to these tasks will be uninfluenced by irrelevant variables and indicative of the job performance that is desired. Studies of these responses must also be carried out in order to establish that these responses have only the meaning that is intended.

It is probably best to include the notion of response relevance under the rubric of construct rather than content validity, since this notion is concerned with establishing that a set of test responses has a particular meaning. As noted previously, this is a matter with which construct validity is concerned; no special term such as "content validity" is needed by test specialists to refer to it.

The Concept of Domain Clarity

Also included in most test specialists' views of content validity is the idea that this validity is determined, in part, by the clarity with which the content domains of a measure are defined (e.g., Anastasi, 1976; APA et al., 1974; Lennon, 1956; Linn, 1980; Rozeboom, 1966). Test specialists generally have thought that experts' judgments should be used to establish definitional clarity, and they have discussed several systematic methods of collecting these judgments (see Cronbach, 1971; Cronbach et al., 1972; Hambleton & Eignor, 1978; Rovinelli & Hambleton, 1977).

Where authorities appear to have differed most is on the matter of what characteristics a clearly defined domain should possess (Benson, 1981). Traditionally, measurement texts have indicated that achievement measures are well specified when the subject matter and cognitive processes to be measured are indicated and the number and format of the items that are used to measure each behavior of interest are noted (e.g., Brown, 1976; Rozeboom, 1966; Thorndike & Hagen, 1977). However, more rigorous and operational specifications also have been advocated by some test specialists who have taken the view that the definition of a test should specify all aspects of the testing procedure that are likely to significantly affect examinees' performance on a test (Cronbach, 1971; Millman, 1974). Specifically, it has been suggested that this definition comprise a detailed description of the content, structure, and scoring of the items that are used to measure each behavior a test is intended to assess (APA et al., 1974; Cronbach, 1971; Popham, 1978).

Discussion

It is a frequently noted principle of empirical science that an observation can be accepted as factual only if the observation can be reproduced by independent observers (Kaplan, 1964). This principle of reproducibility also has concerned psychologists and social scientists, as is reflected in the significance they ascribe to the quality of reliability in their measures (Hempel, 1965).

One purpose of requiring that a measurement operation be clearly defined is to improve the reproducibility of the results obtained when this operation is used (Dodd, 1942; Peak, 1953). This aim is scientifically sound, so it seems only reasonable that test specialists should consider it important to clearly specify the content domains of tests used in educational, psychological, or employment testing contexts. In fact, it might be said that ideally a test should be defined in such a way that if a second test were devised using the same definition and administered to the same examinees, these examinees should obtain comparable scores on the two measures, with any lack of equivalency in these scores being simply a product of random and/or sampling error (see also Cronbach et al., 1972).

It does not seem reasonable, however, to refer to judgments of the clarity of a domain definition as evidence of "content validity." By doing so, it is erroneously implied that these judgments establish a form of test validity. As previously noted, test validity is established by findings from studies of test responses and not by conclusions drawn from analyses that concern test content. It would therefore be most useful to simply say that these judgments reflect "domain clarity." Then, the import of these judgments, when referred to, will remain clear.

Since test specialists have not agreed upon an operational definition of domain clarity, it is useful to consider here what elements could be included in a domain definition that would help to make it clear. It will be recalled that Cronbach (1971) and Millman (1974) recommended that the definition of a content domain should detail all aspects of a test procedure that are likely to significantly affect examinees' performance on the test. This recommendation seems sound; such a definition, if used, clearly would contribute to the reproducibility of test results.

Popham (1978) has formulated an approach to defining a content domain that requires a test developer to specify what seemingly are the characteristics likely to most influence test performance. According to Popham, the definition of this domain should include

1. Rules for determining what content and structure must be evident in items used to measure a desired behavior.
2. Rules for scoring performance on these items.
3. Test directions relevant to these items.
4. One or more examples of the types of items that are admissible as measures of the desired behavior.

Of course, following these guidelines does not ensure that a domain definition will be clear. It does appear, however, that Popham has identified those features of a domain definition that, if left unspecified, might vary each time a test was devised on the basis of this definition and might significantly affect examinees' test performance.

The Concept of Technical Quality in Test Items

Study of the technical quality of test items, using appropriate empirical and logical procedures, is the final kind of investigation that has been mentioned by test specialists, albeit infrequently, as requisite for establishing the content validity of a measure. Hambleton and Eignor (1979) and Benson (1981) viewed content validity as resting upon how adequately the items of a test represent the test's content domains. These researchers suggested that test items that are flawed cannot be considered adequate representatives of any content domain associated with a test, so that such items, when present, will diminish the content validity of the test. Ebel (1956), who considered content validity pertinent to test responses rather than to test content, suggested that the quality of test items would also influence the degree to which these responses could be regarded as content valid indicators of the behaviors the test is intended to assess; were the items of a reading measure ambiguously stated, for example, the responses to this measure might inappropriately indicate examinees' deciphering powers as well as their skills in reading.

Discussion

The presence of numerous discussions in measurement textbooks of the principles and methods of devising effective items suggests that test specialists have ascribed considerable import to the notion that test items should have good technical properties. Commonly, it is said that such properties are necessary in order for a measure to have the level of difficulty, reliability, and validity that is desired (e.g., Brown, 1976; Mehrens & Lehmann, 1978).

The few empirical studies that have been done show that certain flaws in items can adversely affect a test's empirical properties (Board & Whitney, 1972; Dunn & Goldstein, 1959), but logical considerations alone would also suggest that good technical quality should be featured in any test that is developed. As Hambleton and Eignor (1979) implied, it can be assumed that good technical quality is a requirement implicit in the specifications of a test's content domains. Consequently, poor items that appear in a test will jeopardize the fit between the test and its definition and, therefore, the reproducibility of any test results that are obtained.

Thus, it appears important to establish that the items of a test are technically sound. To do this, analyses of both item content and item responses typically are carried out (Henrysson, 1971). The results of such internal analyses do not establish that the responses to the test reflect the behavior that is intended, so these results cannot be regarded as evidence of validity (APA et al., 1974). However, these results can be used to establish that the test procedure has no apparent defects that would obstruct reliable and valid measurement of this behavior.

Summary and Conclusions

This article has described and discussed the various ways that content validity has been defined by test specialists. It was noted that these specialists have variously associated this validity with (1) the sampling adequacy of test content, (2) the sampling adequacy of test responses, (3) the relevance of test content, (4) the relevance of test responses, (5) the clarity of domain definitions, and (6) technical quality in test items.

An evaluation of the theoretical and practical soundness of using each of these notions to define content validity suggested that these notions are best regarded as definitions of concepts other than content validity. It was argued that the sampling adequacy of test content, the relevance of test content, and the clarity of domain definitions should be associated with the terms ''content representativeness,'' ''content relevance,'' and ''domain clarity,'' respectively, rather than with the term ''content validity'' because these notions do not refer to any kind of test validity. Also, it was indicated that the sampling adequacy of test responses and the relevance of these responses are matters already encompassed by the concept of construct validity. Finally, it was suggested that technical quality in test items is an important feature of a test but should not be said to reflect any form of validity.

Thus, no adequate means of defining content validity was found. In light of this, it seems reasonable to conclude that content validity is not a useful term for test specialists to retain in their vocabulary.

References

Aiken, L. R. *Psychological testing and assessment*. Boston MA: Allyn & Bacon, 1979.

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

Anastasi, A. *Psychological tests*. New York: Macmillan, 1976.

Benson, J. A redefinition of content validity. *Educational and Psychological Measurement*, 1981, *41*, 793–802.

Board, C., & Whitney, D. R. The effect of poor item-writing practices on test difficulty, reliability, and validity. *Journal of Educational Measurement*, 1972, *9*, 225–233.

Brodbeck, M. The philosophy of science and educational research. *Review of Educational Research*, 1957, *27*, 427–440.

Brown, F. G. *Principles of educational and psychological testing* (2nd ed.). New York: Holt, Rinehart & Winston, 1976.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measures*. New York: Wiley, 1972.

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

Cureton, E. E. Validity. In E. F. Lindquist (Ed.), *Educational measurement*. Washington DC: American Council on Education, 1951.

Dodd, S. C. Operational definitions operationally defined. *American Journal of Sociology*, 1942, *48*, 482–489.

Dunn, T. F., & Goldstein, L. G. Test difficulty, validity and reliability as functions of selected multiple-choice item construction principles. *Educational and Psychological Measurement*, 1959, *19*, 171–179.

Dunnette, M. D., & Borman, W. C. Personnel selection and classification systems. *Annual Review of Psychology*, 1979, *30*, 477–525.

Ebel, R. L. Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement*, 1956, *16*, 269–282.

Ebel, R. L. Comments of some problems of employment testing. *Personnel Psychology*, 1977, *30*, 55–63.

Equal Employment Opportunity Commission. *Uniform guidelines on employee selection*. Washington DC: U.S. Government Printing Office, 1978.

Federal Executive Agencies. *FEA guidelines on employee selection procedures*. Washington DC: U.S. Government Printing Office, 1976.

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), *Psychological principles in system development*. New York: Holt, Rinehart, & Winston, 1962.

Gleser, G. C. Discussion. *Proceedings of the 1969 invitational conference on testing problems*. Princeton NJ: Educational Testing Service, 1969.

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

Green, B. F. A primer of testing. *American Psychologist*, 1981, *36*, 1001–1011.

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

Guion, R. M. "Content validity" in moderation. *Personnel Psychology*, 1978, *31*, 205–214. (a)

Guion, R. M. Scoring of content domain samples: The problem of fairness. *Journal of Applied Psychology*, 1978, *63*, 499–506. (b)

Hambleton, R. K., & Eignor, D. R. Guidelines for evaluating criterion-referenced tests and test manuals. *Journal of Educational Measurement*, 1978, *15*, 321–327.

Hambleton, R. K., & Eignor, D. R. *A practitioner's guide to criterion-referenced test development, validation, and test score usage* (Report No. 70). Amherst MA: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluative Research, 1979.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159–170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *48*, 1–47.

Hempel, C. G. *Aspects of scientific explanation and other essays*. New York: Free Press, 1965.

Henrysson, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington DC: American Council on Education, 1971.

Kaplan, A. *The conduct of inquiry*. San Francisco: Chandler, 1964.

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

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

Linn, R. L. Issues of validity in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), *Practice and problems in competency-based measurement*. Washington DC: National Council on Measurement, 1974.

Linn, R. L. Issues of validity for criterion-referenced measures. *Applied Psychological Measurement*, 1980, *4*, 547–561.

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, *3*, 635–695 (Monograph Supplement No. 9).

Mehrens, W. A., & Lehmann, I. J. *Measurement and evaluation in education and psychology* (2nd ed.). New York: Holt, Rinehart, & Winston, 1978.

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

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

Miller, R. B. Task description and analysis. In R. M. Gagné (Ed.), *Psychological principles in system development*. New York: Holt, Rinehart, & Winston, 1962.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley CA: McCutchan, 1974.

Millman, J. *Strategies for constructing criterion-referenced assessment instruments*. Paper presented at the Conference of Large Scale Assessment, Denver, 1978.

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

Payne, D. A. *The assessment of learning*. Lexington MA: Heath, 1974.

Peak, H. Problems of objective observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences*. New York: Dryden, 1953.

Polin, L., & Baker, E. L. *Qualitative analysis of test item attributes for domain-referenced content validity judgments*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Popham, W. J. *Criterion-referenced measurement*. Englewood Cliffs NJ: Prentice-Hall, 1978.

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

Rozeboom, W. W. *Foundations of the theory of prediction*. Homewood IL: Dorsey, 1966.

Shimberg, B. Testing for licensure and certification. *American Psychologist*, 1981, *36*, 1138–1146.

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

Thorndike, R. L., & Hagen, E. *Measurement and evaluation in education and psychology* (4th ed.). New York: Wiley, 1977.

Tomko, T. N. *The logic of criterion-referenced testing*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1981.

Acknowledgments

A version of this paper was presented at the annual meeting of the American Educational Research Association, Boston, 1980. The author thanks Kathleen Burk, Ronald K. Hambleton, and two anonymous reviewers for helpful comments on a draft of this paper.

Author's Address

Send requests for reprints or further information to Anne R. Fitzpatrick, P.O. Box 626, North Amherst MA 01059 U.S.A.