Content Validity—The Source of My Discontent

Robert M. Guion
Bowling Green State University

The concept of content validity takes on special importance where invoked to justify use of a test. The term 1) refers to psychological measurement, 2) using samples of behavior, sampling both stimulus and response components, and 3) implies representativeness in sampling. Examples are given to show that content sampling may be considered a form of operationalism in defining constructs. Five conditions are proposed as necessary if one is to accept the use of a measuring instrument as a valid operational definition on the basis of content sampling alone.

I must offer a preface of personal history. I first encountered the term *content validity* in the 1954 Technical Recommendations (APA, 1954). My textbook on personnel testing (Guion, 1965) gave it one full page (of more than 500), dismissing it easily as simply content sampling, more appropriate to the classroom than to the personnel office.

A few years later, with a committee of my peers, I had a guilty hand in formulating the document that became, after mutation, the EEOC Guidelines on personnel selection procedures (EEOC, 1970; OFCC, 1968, 1971). Quite casually, the committee inserted two sentences on content validity into that document. One said that content validity might be a permissible means of evaluating employment tests where criterion-related validation was not feasible. The other said that content validity might be a permissible means of evaluating employment tests even where criterion-related validation *might* be feasible. In retrospect, that seems not very enlightening! The source of the confusion is that, although each of us considered content validity a simple concept, we were either unsure of its nature or in disagreement about it. Since, as we often reminded ourselves, we were not writing a textbook, we were spared the painful necessity of explicating clearly either the nature of content validity or of the appropriate defense for its use.

Still later, with a different committee of peers, I had another guilty hand in the formulation of the 1974 *Standards* (APA, 1974). Although it, too, was not to be a text, the space it devoted to explicating content validity was twice what I had allocated to it in my book! And much to my embarrassment, and despite many revisions and hearings, the text material is confusing and even contradictory. No distinction is made between "domain" and "universe." Referring to employment tests, the left hand column on page 29 says the defined domain may be restricted to certain critical, frequent, or prerequisite work behaviors; but the right hand column says the defined

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 1 Winter 1977 pp. 1–10

universe should include all nontrivial parts of the job.

A phone call from C. H. Lawshe last summer set new ideas in motion, and I found other ambiguities in my thinking. Later, Messick (1974) convinced us in his Division 5 presidential address that there is no such thing as content validity; but, because of Lawshe's call, I had already scheduled a conference on it for October—which ended without closure. Then Mary Tenopyr and I, with yet another committee, prepared a document for Division 14 (Division of Industrial-Organizational Psychology, 1975) in which we tried to avoid the term. The committee objected, so we compromised. The section is headed "Content Validity," but our preferred heading is below that in parentheses: "Content-Oriented Test Development." This summer, while that document was in press, another larger, more structured conference emerged from the October one; we called it Content Validity II. As you shall see, it has influenced this paper.

This personal account may help explain my discontent. After all this background, I don't know precisely what content validity is, or if it is, or what to do about it. My discontent is well-expressed in a nursery rhyme I used to read to my children:

> Last night I saw upon the stair,
> A little man who wasn't there.
> He wasn't there again today.
> Oh, how I wish he'd go away!

Unfortunately, content validity, or at least the problems and concerns encompassed in our use of and debate over the term, won't go away. In his role as discussant in Content Validity II, Ebel (1975) said it well:

> Perhaps instead of content validity we should call it content reliability, or job sample reliability.
>
> Perhaps we should, but I doubt that we will. Verbal habits are not that easy to change. We are no doubt fated to live henceforth with somewhat imprecise terminology, and with the confused thinking about test quality it is likely to spawn.

That pessimistic note was the most recent word I've heard. Do you wonder at my uneasiness over content validity? How I wish it would go away!

The Concept of Content Validity

It won't go away because it is sometimes used to justify the use of a test, and any attempt to justify test use is important enough to require serious consideration. We need, therefore, to try to understand what thoughtful people are talking about when they use this term. Three points are basic in their discussions.

First of all, they are talking about psychological measurement, which always has at its base the observation of a sample of behavior. Cronbach (1971) and Messick (1974) have insisted, and correctly, that validity of any sort is an attribute of scores rather than of tests themselves. This is, of course, all very well; but scores are based on responses to carefully standardized stimuli, observed under carefully standardized conditions. In arriving at a score (and I use the term in its broadest sense), we present some stimulus material or situation, we watch the ensuing responses and then we count them, or we classify some of them as right and count those, or we rate them on some scale appropriate to our purposes. In short, in all psychological measurement, we have first of all a specified set of operations for observing and evaluating relevant kinds of behavior. The "content" in discussions of content validity is, therefore, behavior in response to stimuli.

Second, people who talk about content validity are talking about *samples* of behavior. Measurement does not consist of observing every bit of behavior that occurs. A psychologist may observe his subject scratch his left ear, but the observation is rarely recorded in measurement. There must be boundaries to the nature of the stimuli and subsequent behavior to be recorded. Some classes of behavior are within the stated

boundaries; others are not. The boundaries are often quite amorphous. Yet such as they are, they define the "content domain" in discussions of content validity.

Third, people who talk about content validity are talking about how well a *small* sample of behavior observed in the measurement procedure represents the whole class of behavior that falls within the boundaries defining the content domain. This content representativeness is explicitly what people are talking about when they speak of content validity.

Examples of Content Representativeness

Consider now three quite different examples of measurement problems in psychology to which this notion of representative sampling may be applied in developing a measuring instrument.

The first example is a reading comprehension test. Its purpose is to screen out people who are *not* qualified to begin training for a certain job. That training requires the absorption of a great deal of information from a variety of knowledge areas: practical engineering, law, and physiology, as well as information about the organization and its resources. Training lasts for more than three months. Trainees must read and comprehend well enough to remember and to use written material from a variety of sources.

Arbitrarily, we decided that the behavioral content domain would consist of all written assignments in the first two weeks and the answering of all questions that might be asked about the meaning of that material, except questions of implications or of relationships to other material. Notice that this definition implies two sampling problems. One is sampling passages to be read. The other is sampling questions. The first sampling problem was easy. Assignments could be classified by topic, and passages conforming to certain rules could be selected essentially at random. Representativeness of the sample by topic was assured, and it was easily demonstrated that the difficulty level of the selected passages was also representative of the reading difficulties encountered in the total domain.

The second sampling problem was more difficult. There is no existing universe of all possible items, and the variety of possible responses to an item is virtually unlimited. An administrative decision limited the possible item domain to a four-choice format.

Note that we are talking about a "behavioral content domain" and sample. The test consists of *reading* passages and *answering* questions. The response content domain can hardly be conceptualized independently of the stimulus content domain. This is obvious, but many discussions of content validity seem to overlook the fact that a defined content domain necessarily includes both stimulus content and response content.

My second example comes from a thesis by Schimmel (1975) concerned with the measurement of a personality construct called assertiveness. Following Lazarus (1973), he identified four components of an appropriate behavioral content domain: (a) saying "no," (b) asking favors or making requests, (c) expressing both positive and negative emotion, and (d) taking part in conversations, including initiating or terminating them. Schimmel considered existing instruments biased in that they sampled only tendencies to express negative emotion.

He developed a pool of items for a self-description questionnaire. There were content-oriented rules for item development, such as "There must be at least two questions in which the stimulus person is a family member." Each item in the pool was allocated to one of the categories by a panel of judges.

In this example, a theoretical construct defined the boundaries of a behavioral domain. Measurement within that domain *could* have been done using observers and standardized social situations, but it *was* done in a self-description inventory format—an executive decision which further defined the boundaries of a content domain.

My next example is a recommendation for the study of aggressive behavior in mice.[1] One might measure aggression by counting the number of fights, but aggression is thought to be signaled in other ways: sniffing at close range or nosing, climbing, or licking another animal, or a boxing posture where the mouse stands on its hind legs and extends forepaws in the direction of the other animal.

A domain of such behavior could be identified by combing the literature on agonistic behavior and extended by specifying situational variables to define a test situation such as familiarity with the cage, level of illumination, and the like (Scott, 1966). Behaviorally anchored rating scales could be developed by scaling specific behavior statements such as "placed paws on other animal" or "licked fur or extremities of other animal" (Cairns & Nakelski, 1971). The content of an actual rating scale would then be representative of the domain if the behavioral anchors are chosen to represent the entire range of scale values and if the observations are made in a representative set of situations.

By defining a content domain of behaviors known to have some probability of leading to attack within specified stimulus conditions, one could develop a system for rating aggression in mice and that system would be characterized by what mental testers have called content validity.

Content Validity and Operationism

Of these three, the first example is the most conventional in discussions of content validity. The notions of content validity are less likely to be applied to the other two because they involve the measurement of hypothetical constructs. Yet the notion of defining a content domain of stimulus situations and recordable responses, and then developing a standardized sample from that domain, is fundamental to the measurement in all three problems and, I submit, to all psychological measurement.

[1] I am indebted to Thomas F. Sawyer for a summary of this literature and its measurement problems.

Consider briefly the measurement problems of the laboratory. The dependent variable in psycholinguistic research may be reaction time. Perceptual research may call for accuracy of discriminations between physical stimuli. Operant studies may count the number of bar presses.

Content validity is considered irrelevant in such physical measurement—at least, I presume that Ebel would say so given his view that not all measures must be valid (Ebel, 1961). Nevertheless, the experimenters I know are very careful to specify the conditions under which they measure the passage of time, or the number of bar presses, or the accuracy of discriminations. That is, for psychological measurement they do in fact establish behavioral content domains, and samples from these domains constitute their measurements.

The only thing is, they don't talk about content validity! They speak of operational definitions. But do those who talk about content validity mean anything more than the adequacy of operational definition?

I think not. I think a sample of behavior within a defined content domain, with its standardized set of "admissible operations" (Cronbach, 1971), constitutes the operational definition of a construct.

Ebel (1975) has made me uneasy about this. He decries loose use of the term *construct*. He wants the word restricted to postulated attributes underlying and determining overt behavior:

> If the behavior can be directly observed, or if the trait can be operationally defined, it is not a construct in this sense . . . Most of what we teach in educational institutions, and most of what we test for in employee selection, are knowledges, skills and abilities. These can all be defined operationally. They are not hypothetical constructs. Ability to type, to spell, to weld, to solve problems with algebra, calculus, or computers; these are not the kind of latent traits Cronbach and Meehl had in mind. We would speak more sensibly, I think, if we did not call them constructs.

I *do* like to speak sensibly, and I dislike further corrupting the language, but I don't know what else to call them.

Suppose we need a test of driving skill to be used in selecting cab drivers. We must define a content domain in order to define operationally what is meant by driving skill. We identify stimuli and the possible responses to them such as driving on ice, swerving to avoid another car, and so on. Identification of these kinds of stimulus conditions, and the responses to them that distinguish the skilled driver from the less skilled, conceptually defines an attribute of drivers. When a standardized sample is drawn from that domain and called a test, it operationally defines that attribute. The attribute in question *is* a construct. It is not directly observed; it is inferred from specific observable behaviors (and its lack is inferred from other behaviors), and it enters into a nomological network with other variables such as maintenance costs, accidents, and the like.

The reading test described earlier is also a measure of a construct. One does not really observe people reading. One observes that printed material is in place before them, and certain kinds of eye movements, and infers that they are reading. The nomological net is less obvious; but reading, as an inferred psychological process, is also a construct. Inferred attributes and processes include assertiveness in man and aggressiveness in mice, and the constructs so inferred are conceptually defined when the boundaries of a behavioral content domain are established. They are operationally defined when a standardized sample from that domain is used for measurement.

It seems obvious, then, that people who talk about content validity are either (a) talking about a special case of construct validity or (b) not talking about validity at all, but simply about the operational definitions of their constructs. This latter conclusion enjoys some distinguished company. Listen to Messick:

> Content coverage is an important consideration in test construction and interpretation, to be sure, but in itself it does not provide validity. Call it "content relevance," if you would, or "content representativeness," but don't call it content validity . . . (Messick, 1974).

Or consider Tenopyr:

> . . . there should be no real conflict about whether content or construct validation is appropriate in a given situation. The question instead is one of for which *class* of constructs is evidence of traditional views of content validity alone enough to justify the contention that these constructs are being measured . . . (Tenopyr, 1975).

Or hear Ebel:

> Only when one variable is measured in order to make inferences about some presumably related variable, or about some underlying, and hence unobservable determinant of a particular kind of behavior, do real questions of validity arise . . . content validity is not really a kind of validity at all. (Ebel, 1975).

The little man of content validity isn't there again today.

Content Validity as a Defense

But he still won't go away. He won't go away because people who talk about content validity are talking about important *evaluations* of operational definitions. They invoke the concept of content validity primarily justifying the use of a measuring instrument. Under this term, they are asking: "When do content-oriented considerations alone justify the acceptance of an operational definition of a variable without further empirical data?"

I do not have a satisfactory answer, but I would like to propose five conditions that may be necessary to the acceptance of a measure on the basis of its content. This answer is to be recognized as tentative, one among possible

others, and not very profound.[2]

First: The content domain must be rooted in behavior with a generally accepted meaning.

A good illustration of generally accepted meaning is a driver's license examination. To pass the test, one must make a right turn without veering into the left lane or climbing the curb. One must stop the car without producing whiplash in the examiner. One must park the car between flags without knocking them down. I did *not* say, "One must be able to . . . " One must *do* these things. If one *does* them, the *ability* to do them is accepted without question by the doer and by the observer. The *meaning* of the score is directly and unequivocally related to the doing of them.

Of the three examples I offered earlier, only the measurement of aggression in mice involves such directly observable physical activity. Yet, it is the one for which judgments of content validity are least useful. The difficulty is in the meaning of the observations. We can observe one mouse place paws on the other, but perhaps this could be interpreted as affection as well as aggression; the meaning of the observed behavior depends on empirical investigations.

"Behavior" in the other examples is answering questions. The assertiveness scale may ask, "Do you initiate conversations?" Initiating a conversation or failing to do so is observable behavior, but the answer-behavior poses problems of interpretation. A "yes" answer may mean acquiescence, assertiveness, or verbal diarrhea. We need empirical data to decide. So-called content validity is a helpful but not a sufficient basis for interpretation.

The reading test asks questions, not about behavior, but about ideas. From these questions and their answers, we infer a cognitive rather than a physical or social process: reading with comprehension.

In both assertiveness and reading ability, we infer process variables—constructs—from the answers to questions. There is, however, an important qualitative difference between the inference of assertiveness and the inference of a reading ability. The inference of assertiveness depends on a theoretical structure. The inference of reading with comprehension depends on simple introspection; i.e., we know by a kind of introspective reasoning that we can answer the questions only if we have read them and the material they cover.

Introspection, of course, is not enough. Introspectively, one may report that one is assertive if he answers a particular question in a certain way, but there is a strong likelihood of encountering objections. If the reading comprehension test can be taken by itself as an acceptable operational definition, it is because the behavior sampled has a generally accepted meaning. Reading passages, recalling factual information from a history book, driving an automobile skillfully, solving complex problems in arithmetic are all processes, but they are *not* especially hypothetical. Their meaning derives from their action and outcome, not from a theoretical, nomological network—even though such a network doubtless exists.

Second: The content domain must be defined unambiguously. The boundaries of a domain should be clear enough that different people understanding the measurement problem at hand should be able to recognize reasonably well whether a particular item or basis for observation is in or outside those boundaries. Those people need not agree on the correctness of the boundaries. Two people developing achievement tests may have entirely different ideas of what the content domain ought to be. But if either of them describes the boundaries of his domain to the other, both should understand those boundaries and be able to judge whether specific test items fit within them.

Third: The content domain must be relevant to the purposes of measurement.

---

[2]A totally different sort of answer could be developed and be at least as useful using as a point of departure the notion of intrinsic validity (Gulliksen, 1950). I've taken content sampling as the theme simply as it is currently an orthodox line needing development.

In my first example, the reading comprehension domain was relevant to the selection purpose insofar as the content was necessary for training. In the second example, the content domain was relevant insofar as it assisted in the development of the measure to fit the theory. In the third example, the content domain was relevant to the agonistic behavior being studied insofar as empirical data demonstrate a probability of attack associated with the included behavior.

Most discussions of content validity are, in fact, discussions of content relevance. To illustrate content relevance, assume for the moment that good sales behavior is assertive. If we develop those scales of assertiveness that fit well the theoretical content of assertive behavior, we will not, nevertheless, have a sufficient basis for using the scales to identify good salesmen. The considerations that declare the scales acceptable for scientific research about the construct of assertiveness do not also declare them sufficient for selecting salesmen or for counseling people on the appropriateness of a sales career.

Content relevance is perhaps best understood by thinking of a test content domain independently of some external content domain. In educational testing, the external domain might be called a curriculum content domain. In employment testing, it might be called a job content domain. The content relevance of the educational achievement test is a function of how well the test content matches the curriculum content domain, as defined. The content relevance of the employment test is a function of the excellence of the match of the test content and the job content domain, as defined.

This third condition has several implications: (a) Changes in the definition of the external content domain change the degree of relevance of the measure. (b) The more major features of the external content domain duplicated in the measure, the more relevant it is for that domain. (c) The more closely the proportion of elements in the content of the measure matches the proportion in the external domain, the greater the relevance. (d) The more the measurement content includes behaviors not within the defined external domain, the less relevant it is. (e) The notion of content relevance is a quantitative one, even if we currently lack the means of measuring it.

Fourth: Qualified judges must agree that the domain has been adequately sampled. I may judge a measurement procedure an adequate sample of a defined domain. If someone else disagrees, we may argue. If one of us is better qualified to make the judgment, however, the weight of the argument swings in his favor. Further weight is given to that argument if most other qualified judges in a group agree.

Who is a qualified judge? A psychologist? Perhaps—for some things. The important qualification, however, is the degree of one's knowledge of the external content domain. The relevance of content sampling in a test to the content domain of a job is better judged by people who have performed that job, or supervised its performance, or have done a careful analysis of it, than by those whose main qualifications are degrees in psychology. People who have, over a period of time, worked and compromised and fought to build a certain curriculum are better qualified to judge the relevance of a test to the curriculum content domain than are professors of education with abstract, generalized curriculum ideas but no first-hand knowledge of the curriculum in that specific school system.

Fifth: The response content must be reliably observed and evaluated.

Among the things that tempt me to toss a tantrum is the affirmation that statistical considerations are irrelevant to discussions of content validity. Since my evaluation of this nonsense lacks the dignity this forum deserves, I'll simply express an opposing view. Reliability is essential.

This does not refer to internal consistency, of course, nor to retesting after long intervals. It does refer to standardization that allows at least some assurance that the stimulus content is presented in the same way for all examinees and the response content is evaluated according to the same rules by all observers.

It annoys me that we have no adequate means for statistical assessment of how well each of these five conditions has been satisfied in the particular case. Lawshe has presented his content validity ratio (Lawshe, 1975, in press), and it is a useful tool for being sure that each item in a test or proposed for inclusion is judged relevant to a job content domain by a panel of qualified judges. Cronbach (1971) has suggested correlating tests built to the same sampling specifications. Beyond these, indices of content relevance, distance of inferential leaps, clarity of boundaries, and deficiencies in content sampling are wanting.

Nevertheless, a content-oriented test development can and should utilize statistical approaches, even where content-referenced interpretations are anticipated. Questions of item difficulty, functional unity and reliability are not irrelevant.

Further Sources of Discontent

If I could stop here, I could pretend to have relieved the discontent created by discussions of content validity. The fact is that, even if these five conditions are satisfied, I would still be uneasy.

For one thing, I am uneasy because I anticipate a runaway use of the notion of content validity in employment testing. Judgments of content validity have been too swiftly, glibly and easily reached in accepting tests that otherwise would never be deemed acceptable. My fear is that the result will be a stupid use of tests which will further erode what is of value in the content validity discussions. Messick and Ebel have decried the use of the term, but ill-advised practitioners may well cause us to lose more than the use of a word.

The term gave rise to these rules for the sufficiency of content-oriented operational definitions. That idea is exceedingly important, particularly in the measurement of dependent variables in all areas of psychology. In organizational psychology, in educational psychology, in program evaluation, and in many other areas, the dependent variables are the criteria used in criterion-related validation of other instruments or treatments. It has often been pointed out that unless we engage in an infinite regress in criterion-related validation of criteria, there comes a point when one simply must accept the measure at hand. Operational definitions of variables deemed sufficient on content considerations provide the set of circumstances by which we, singly or collectively, can decide on such acceptance. But if these notions are eroded by misuse, their potential values will never be realized.

I am uneasy also about the untouched problem of fairness. Cronbach's example describes the sort of situation that distresses me: "A dictated spelling test is a measure of hearing *and* spelling, vocabulary *and* ability to write. In terms of content, the spelling test tests ability to spell from dictation whether the pupil is deaf or had normal hearing" (Cronbach, 1971, p. 453). He's right, of course, but that doesn't quiet my discontent in an era of equal employment opportunity for the deaf.

Another example comes from Content Validity II. Schoenfeldt, Schoenfeldt, Acker, & Perlson (1975) had described the development of an industrial reading test; a question from the floor asked whether blacks had been represented in the panel of experts he used. Everything in the litany on content validity says this is an irrelevant question. If in fact the jobs require the reading of material written in standard English, then the color of those who judge whether the samples chosen adequately represent that material is totally unimportant. But that doesn't quiet my discontent.

Existing fairness literature, from Guion (1966) and Cleary (1968) to McNemar (1975), is based on criterion-related regression. It is as if the problem of fair test use exists only when the test is justified on the basis of its correlation with an external criterion. But the fairness literature also speaks in terms of cutting scores, and cutting scores are applied also to tests defended as content samples.

Cronbach's content validation, he says, "looks on the test as an instrument of absolute measurement . . ." (Cronbach, 1971, p. 453). But in all candor, I do not see how one can assign an absolute interpretation to a score based on a specific sample from a behavior domain. We all know of the capriciousness of the psychometric properties of questions, formats, instructions, or testing conditions. We know we can change the difficulty of an item by changing its wording or the position of a distractor.[3] Personally, I consider this important even if one intends content-referenced interpretations of the scores. If the difficulty level can be changed within a group of subjects, is it unreasonable to ask whether difficulty levels might differ in different ethnic or sex groups—and for reasons unrelated to the defined content domain but to unrecognized content of the test? And if so, is it unreasonable to ask whether the same cutting score means the same thing in differing groups? More fundamentally, is it unreasonable to inquire into the fairness of the defined content domain?

Finally, I am uneasy about the lack of basic psychological thought in these discussions. I am particularly uneasy about our tendency to treat defined content domains as discrete, independent territories, all unrelated to each other. It perpetuates unnecessarily the situation-bound nature of most mental measurement. It seems to me that we could go far in developing more generalizable interpretations of measures if we could instead apply concepts of the transfer of training, that is, if we were to identify transferable domains. Certainly the selection of candidates for special educational opportunities or treatments or employment would be improved if tested performance in one domain could be experimentally demonstrated to transfer to performance in certain other domains.

[3]Such considerations may not matter for people who have really mastered the content. They cannot be ignored, however, for degrees of achievement at less than the level of genuine mastery.

Summary

Where has my discontent led me? It has led me to retreat somewhat from the use of the term validity but to hold more strongly than ever to the ideas represented by the notion of content domain sampling. My conclusions can be summarized in a set of dogmatic statements:

1. The ideas implied by discussion of content validity apply to *all* psychological measurement, including that of the laboratory.
2. The definition of a content domain requires specification of both stimulus and response domains. It is, therefore, a *behavioral* domain, not simply a domain of free-floating acts or facts.
3. A standard sample from a behavior content domain is the operational definition of a construct.
4. That operational definition is, by itself, a sufficient justification for the use of the resulting measurement if: (a) the behavioral content has a generally accepted meaning; (b) the domain is unambiguously defined; (c) it is relevant to the purposes of measurement; (d) it is, according to consensus, adequately sampled; and (e) it is sampled reliably.
5. We have no established set of rules for demonstrating, other than those that are matters of consensus, that these conditions have or have not been met.
6. The glibness with which many people have invoked the idea of content validity threatens the fruitful development of the more precise idea of content sampling as a sufficient justification for an operational definition.
7. The question of fairness in the use of such operational definitions has not been faced.
8. Discussions under the heading of content validity fail to apply the principle of transfer of training that might permit knowledge of an individual's performance in one domain to be used to estimate or predict his performance in another.

Clearly, the ideas associated with discussions under the heading of content validity are not as simple or trivial as I once thought.

References

American Psychological Association, American Educational Research Association & National Council on Measurements Used in Education. Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin,* 1954, *51,* 201-238.

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

Cairns, R. B., & Nakelski, J. S. On fighting in mice: Ontogenetic and experimental determinants. *Journal of Comparative and Physiological Psychology,* 1971, *74,* 354-364.

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

Division of Industrial and Organizational Psychology, American Psychological Association. *Principles for the validation and use of personnel selection procedures.* Hamilton, Ohio: Author, 1975.

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

Ebel, R. L. *Prediction? Validation? Construct validity?* Paper presented at Content Validity II, a conference at Bowling Green State University, July 18, 1975.

Equal Employment Opportunity Commission. Guidelines on employee selection procedures. *Federal Register,* August 1, 1970, *35* (No. 149), 12333-12336.

Guion, R. M. *Personnel testing.* New York: McGraw-Hill, 1965.

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

Lawshe, C. H. *A quantitative approach to content validity.* Paper presented at Content Validity II, a conference at Bowling Green State University, July 18, 1975.

Lawshe, C. H. A quantitative approach to content validity. *Personnel Psychology,* 1975, in press.

Lazarus, A. A. Assertive training: A brief note. *Behavior Therapy,* 1973, *4,* 697-699.

Messick, S. *The standard problem: Meaning and values in measurement and evaluation.* Presidential address to Division of Measurement and Evaluation, Meeting of the American Psychological Association, New Orleans, August, 1974.

Office of Federal Contract Compliance. Validation of tests by contractors and sub-contractors subject to the provisions of Executive Order 11246. *Federal Register,* September 24, 1968, *33* (No. 186), 14392-14394.

Office of Federal Contract Compliance. Employee testing and other selection procedures. *Federal Register,* October 2, 1971, *36* (No. 192), 19307-19310.

Schimmel, D. J. *Subscale analysis and appropriate content domain sampling in the initial development of a measure of assertive behavior.* Unpublished M.A. thesis, Bowling Green State University, 1975.

Schoenfeldt, L. F., Schoenfeldt, B. B., Acker, S. R., & Perlson, M. R. *Content validity revisited: The development of a content validated test of industrial reading.* Paper presented at Content Validity II, a conference at Bowling Green State University, July 18, 1975.

Scott, J. P. Agonistic behavior of mice and rats: A review. *American Zoologist,* 1966, *6,* 683-701.

Tenopyr, M. L. *Content - construct confusion.* Paper presented at Content Validity II, a conference at Bowling Green State University, July 18, 1975.

Acknowledgements

This paper was an invited address to the Division of Measurement and Evaluation (Division 5) at the annual convention of the American Psychological Association, Chicago, 1975.

Author's Address

Robert M. Guion, Department of Psychology, Bowling Green State University, Bowling Green, Ohio, 43403.