will add substantially to readers' confidence in their evaluation of the research. NR

References

CHALMERS, T., SMITH, H., BLACKBURN, B., SILVERMAN, B., SCHROEDER, B, REITMAN, D., & AMBROZ, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials, 2*, 31–49.

COHEN, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

DERSIMONIAN, R., CHARETTE, L., MCPEEK, B., & MOSTELLER, F. (1982). Reporting on methods in clinical trials. *New England Journal of Medicine, 306*, 1332–1337.

DIERS, D. (1979). *Research in nursing practice.* Philadelphia: J. B. Lippincott.

DUMAS, R. G., & LEONARD, R. C. (1963). The effect of nursing on the incidence of postoperative vomiting. *Nursing Research, 12*(1), 12–15.

FEINSTEIN, A. R. (1985). *Clinical epidemiology.* Philadelphia: Philadelphia: W. B. Saunders.

FREIMAN, J., CHALMERS, T., SMITH, H., & KUEBLER, R. (1978). The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *New England Journal of Medicine, 299*, 690–694.

GARFIELD, E. (1984). Journal citation studies. 44. Citation patterns in nursing journals, and their most-cited articles. *Current Contents*, (43), 3–14.

GERBER, R. M., & VAN ORT, S. R. (1979). Topical application of insulin in decubitus ulcers. *Nursing Research, 28*(1),16–19.

GERBER, R. M., & VAN ORT, S. R. (1981). Topical application of insulin to pressure sores: A questionable therapy. *American Journal of Nursing, 81*, 1159.

HILL, A. B. (1971). *Principles of medical statistics* (9th ed.). New York: Oxford University Press.

GOODWIN, L. D. (1984). The use of power estimation in nursing research. *Nursing Research, 33*, 118–120.

MOSTELLER, F., GILBERT, J. P., & MCPEEK, B. (1980). Reporting standards and research strategies for controlled trials: Agenda for the editor. *Controlled Clinical Trials, 1*, 37–58.

POCOCK, S. J. (1983). *Clinical trials: A practical approach.* Chichester, England: John Wiley & Sons.

VAN ORT, S. R., & GERBER, R. M. (1976). Topical application of insulin in the treatment of decubitus ulcers: A pilot study. *Nursing Research, 25*(1),9–12.

Determination and Quantification Of Content Validity

MARY R. LYNN

Validity is a crucial factor in the selection or application of an instrument, for validity is the extent to which that instrument measures what it is intended to measure. Although over 35 terms may be used to connote kinds of validity (Brown, 1980), only three types are in common usage today—content, criterion-related, and construct. There has been little argument about the relative merits of criterion-related and construct validity, but the legitimacy of content validity as a *real* type of validity has been questioned by a number of psychometricians (Carmines & Zeller, 1979; Cronbach, 1970; Messick, 1981). These challenges to the value and merit of content validity have arisen from the confusion of content validity with face validity, the unstandardized approaches to the determination of content validity, and the previously unquantified nature of content validity.

The purpose of this article is to differentiate content and face validity, present the process by which content validity can and should be determined, and demonstrate some means by which aspects of the content validity determination can be quantified.

Face validity, rather than a true psychometric assessment technique, has been defined as validity conferred by the lay person's acceptance that a procedure, statement, or instrument appears to be sound or relevant (Guilford, 1954; Waltz & Bausell, 1981). In addition to this appearance of validity, face validity includes validity by assumption (a nonstatistical assessment of the logical tie between the elements or items of an instrument and its purpose) and validity by definition (the determination by one or more content experts that the elements/items of an instrument represent the content domain being assessed) (Mosier, 1947). Face validity is not quantifiable and long ago fell out of favor in psychometric usage (Mosier, 1947), but the continued confusion between content validity and face validity could also lead to content validity's eventual dismissal as a legitimate form of validity.

Assessment of Content Validity: Content validity is the determination of the content representativeness or content relevance of the elements/items of an instrument by the application of a two-stage (development or judgment) process. When content validity has been viewed as a one-stage process (either development or judgment), it has been challenged most as a form of validity (Jensen, 1980; Messick, 1981). Using a two-stage process to determine and quantify content validity is fundamental to the validation of virtually all instrumentation.

Developmental Stage: The assess-

Accepted for publication November 7, 1985.

The assistance of Jeanne H. Howe, PhD, RN, and Nancy L. McCain, DSN, RN, in the preparation of this manuscript is acknowledged.

MARY R. LYNN, PHD, RN, is a National Research Service Award postdoctoral fellow and research associate at the College of Nursing, University of Arizona, Tucson.

ment of content validity begins in the earliest development of an instrument. The development stage of content validity determination has three steps: domain identification, item generation, and instrument formation (Carmines & Zeller, 1979; Nunnally, 1978; Williamson, 1981). These steps vary in context, depending on the measurement objective of the instrument, i.e., whether the measure is being developed for a cognitive or affective/personality measure (see Table 1).

Table 1. Stages of Content Validity (CV) Determination

Step	Cognitive Instrument	Affective Instrument
	Stage I–Development Stage	
1	Identification of full content domain	Identification of dimension(s) and/or subdimensions of affective variable
2	Sampling and item generation	Generation of items for all dimensions and/or subdimensions
3	Assimilation of items into useable form	Assimilation of items into useable form
	Stage II—Judgment–Quantification Stage	
4	Judgment/quantification of CV of items	Judgment/quantification of CV of items
5	Judgment/quantification of CV of instrument	Judgment/quantification of CV of instrument

For cognitive measures, the full content domain must be identified in Step 1. If, for example, an instrument is being developed to measure clients' knowledge about self-care measures to prevent hypertension, all self-care measures need to be identified and categorized for the next step in this stage to be addressed. Domain identification is often facilitated by the use of a table of specification or a blueprint of the content domain.

The second step in the development stage of a cognitive measure is to sample from the content domain identified and generate the items from these sampled areas. Rarely is it necessary for a cognitive measure to address each and every aspect of the content area to fully represent the scope of the content. Ideally, this sampling should be random, but such is rarely practical. It is sufficient to ensure that all areas or cells of the table of specifications or blueprint have been represented appropriately (see Mehrens & Lehmann, 1984).

In the development of an affective/personality measure the first two steps differ from those just described. In affective measures the domain identification (Step 1) is accomplished through a thorough review of the literature on the topic of the measure so that all dimensions and subdimensions can be identified. This incorporation of the ideas of a variety of experts (perhaps limited to those published) is in contrast to the domain identification in a cognitive instrument which can reasonably be done by a single expert or an arbitrary number of experts.

In the second step for an affective/personality measure, items are generated for all dimensions and subdimensions identified in Step 1. Rarely is a domain for an affective/personality variable so large as to justify or even make possible sampling for item generation, and it is always better to generate too many items than too few (Carmines & Zeller, 1979; Nunnally, 1978).

The third step of the development stage is the same for cognitive and affective/personality instruments. The items generated in Step 2 are now assembled in a usable form, that is, items generated in the second step are refined (reworded as necessary) and arranged in a suitable sequence. Once the items have been finalized and the instrument assembled, the development stage of content validity determination is completed.

Judgment–Quantification Stage: Stage II, the judgment–quantification stage of content validity, has two steps which are the same for both cognitive and affective/personality measures. These steps entail the assertion by a specific number of experts that the items are content valid (Step 4) and that the entire instrument is content valid (Step 5). The number of experts needed in this stage and the proportion of those experts that must agree for content validity to be established can be decided by application of the standard error of the proportion; content validity can be quantified with application of the index of content validity (CVI), as outlined in Waltz and Bausell (1981, p. 71).

Determining the number of experts needed has always been somewhat arbitrary in content validity determination. The number of experts often depends on how many accessible and agreeable persons the instrument developer or user can identify, not on a population estimation principle. Although this practice is widespread, specific guidelines can and should be applied to the selection of experts for content validity determination. A minimum of five experts would provide a sufficient level of control for chance agreement: however, in some content areas it may be difficult to locate this many content/domain experts and to obtain their cooperation. Therefore, in content/domain areas of sufficient restriction to preclude large numbers of experts, a minimum of three experts should be used. The use of only two judges is not only statistically unjustifiable, but also it places the instrument developer at great risk of an erroneous conclusion that content validity has not been achieved when it actually has. The maximum number of judges which might be used has not been established, but it is unlikely to exceed 10.

After (or perhaps as a part of) determining the number of experts, the minimum number of judges who must agree for the items and total instrument to be assessed as content valid must be established. This can be done by calculating the proportion of the number of experts who might agree out of the total number planned for use, and then setting the standard error of the proportion to identify the cut-off for chance versus real agreement. In Table 2 the proportions of experts agreeing on the content validity of an item and the entire instrument and the standard error of those proportions are presented (as a function of the total number of experts). If there are five or fewer experts, all must agree on the content validity for their rating to be considered a reasonable representation of the universe of possible ratings. When six or more experts are used, one or more can be in disagreement with the others and the instrument will be assessed content valid.

Regardless of the number of experts used, a structured procedure for the evaluation of the content validity

Table 2. Proportion of Experts (Above the Line) Whose Endorsement Is Required to Establish Content Validity Beyond the .05 Level of Significance

Number of Experts	Number of Experts Endorsing Item or Instrument as Content Valid								
	2	3	4	5	6	7	8	9	10
2	1.00								
3	.67	1.00							
4	.50	.75	1.00						
5	.40	.60	.80	1.00					
6	.33	.50	.67	.83	1.00				
7	.29	.43	.57	.71	.86	1.00			
8	.25	.38	.50	.63	.75	.88	1.00		
9	.22	.33	.44	.56	.67	.78	.89	1.00	
10	.20	.30	.40	.50	.60	.70	.80	.90	1.00

NOTE: The caution over using the standard error of the proportion when $n \leq 10$ (Downie & Heath, 1974) does not apply in this situation because only when $p > q$ is there significance, and any nonunique p [×] q solutions are irrelevant.

of the instrument must be given to the experts (Kerlinger, 1973). Although the persons selected have already been determined to have expertise in the content/domain area(s) of the instrument, it is not sufficient to supply them with the instrument and simply ask if it is content valid. They must be given a set of specific instructions by which to determine the domain or content relevance of the items and also of the instrument as a whole (see McCain, 1984). The experts should be provided with either a delineation of the full content domain for cognitive instruments or dimensions of the affective variable, with specific questions pertaining to the content relevance of each item.

The most widely used quantification of content validity is the index of content validity (CVI), which is derived from the rating of the content relevance of the items on an instrument using a 4-point ordinal rating scale, where 1 connotes an irrelevant item and 4 an extremely relevant item. The actual CVI is the proportion of items that received a rating of 3 or 4 by the experts (Waltz & Bausell, 1981).

Use of the CVI in both steps of the judgment stage of content validity determination necessitates extension of the CVI to item evaluation as well as entire instrument evaluation. The CVI for each item (Step 4) is determined by the proportion of experts who rate it as content valid (a rating of 3 or 4); the CVI for the entire instrument (Step 5) is the proportion of total items judged content valid.

Waltz and Bausell (1981) posed two limitations to the CVI procedure for the assessment of content validity: The possibility of chance inflation (agreement) of the CVI and the dependence of the CVI on the number of categories used in the rating, in that a four-option scale is not universally used in CVI determinations. These limitations can be controlled by using the methods discussed in this article. The first limitation is addressed in the selection of the number of experts and the minimum number that must agree by using a significant combination from Table 2. The second concern would be controlled by establishment of a 4-option rating scale (1 = not relevant; 2 = unable to assess relevance without item revision or item is in need of such revision that it would no longer be relevant; 3 = relevant but needs minor alteration; 4 = very relevant and succinct) for all content validity assessments. The scale should be both mathematically and conceptually meaningful. Although a 3- or 5-point rating scale might be considered, the 4-point scale is preferable because it does not include the ambivalent middle rating common in odd number rating scales. Using a 4-point rating scale should provide the instrument developer with sufficiently delineated information upon which to calculate a meaningful CVI.

The experts, in addition to judging each item, should identify any area(s) that have been omitted from the instrument. If omissions are identified although the entire instrument is assessed as content valid, further clarification may be necessary. If no omissions are identified and the instrument is judged content valid by the experts, in conjuction with the knowledge of the magnitude of the CVI values, the content validity of the instrument may be asserted. Should an expert identify no areas of omission, evaluate the items positively, and, yet, not assess the instrument as content valid, doubt about the person's expertise should lead to considering replacing that person as an expert. Suggestions for item improvement may be requested from the experts, however, without interfering with the content validity judgment.

Items that do not achieve the required minimum agreement of the experts (Table 2) should be eliminated or further revised. It is certainly possible that several items, and, therefore, the entire instrument, need to be evaluated more than once in order to obtain sufficient content validity. In this case, the instrument developer should determine whether

to use the same experts or obtain new, but equally qualified, replacement experts. Should the same experts be used in repeated assessments of an instrument, a 10- to 14-day period between assessments should be a sufficient time interval (see McCain, 1984).

This process for content validity determination may also be applied to existing instruments, for the validity of many such instruments either has never been reported or has been simply untested. Given that the instrument has been already created, Steps 2 and 3 may not be applicable if, by completion of the content validity process, acceptable content validity is determined. In the event that content validity is not sufficiently achieved, with the instrument developer's permission, Steps 2 and 3 (item and instrument revision) may also be incorporated to eventually obtain content validity.

Summary: The arbitrary assertion of two or three experts does not establish content validity. Application of a two-stage process that incorporates rigorous instrument development practices and quantifies the aspects of content validity is required. In the first stage of this process, the content domain or dimensions are identified and items are generated to reflect the scope of the content domain of a cognitive variable or each of the dimensions of an affective variable. Once generated, the items are assembled in a usable, testable format. The instrument and domain or dimension specifications are then presented to a panel of experts, the size of which is an a priori decision, for their judgment of the items using a 4-point ordinal rating scale. Using the item evaluation, CVI calculations are applied to both the items and the entire instrument. The experts are asked, as a part of the content validity assessment, to identify areas of omission and to suggest areas of item improvement or modification.

Admittedly, there are times when adherence to such rigor may not be feasible. When less stringent methods of determining validity are applied, it should not be said that content validity has been determined. Opponents of the process described in this article might argue that these specifications and expectations exceed practical application and that this process is therefore too rigorous. Content validity, by its nature and definition, demands rigor in its assessment, and its assessment is, in fact, critical. Such a rigorous process for content validity determination is offered because content validity is an inexpendable form of validity which is rapidly losing credibility due to its less than standardized and rigorous assessments. Content validity, different from all other forms of validity, is crucial to the understanding of research findings and their practical or theoretical applications. It is worth the rigor. NR

References

BROWN, F. (1980). Perspectives on validity. *NCME Measurement News*, 23(3), 3-4.

CARMINES, E. G., & ZELLER, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

CRONBACH, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper and Row.

DOWNIE, N. M., & HEATH, R. W. (1974). *Basic statistical methods* (4th ed.). New York: Harper and Row.

GUILFORD, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

JENSEN, A. R. (1980). *Bias in mental testing*. New York: The Free Press.

KERLINGER, F. (1973). *Foundations of behavioral research* (2nd ed.). New York: Holt, Rinehart and Winston.

MCCAIN, N. L. (1984). *A test of Cohen's developmental model for professional socialization with baccalaureate nursing students*. Unpublished doctoral dissertation, The University of Alabama in Birmingham

MEHRENS, W. A., & LEHMANN, I. J. (1984). *Measurement and evaluation in education and psychology* (3rd ed.). New York: Holt, Rinehart and Winston.

MESSICK, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9-20.

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

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

WALTZ, C. W., & BAUSELL, R. B. (1981). *Nursing research: Design, statistics and computer analysis*. Philadelphia: F.A. Davis.

WILLIAMSON, Y. M. (1981). *Research methodology and its application to nursing*. New York: John Wiley and Sons.

1986 American Nurses' Foundation Research Grant Recipients Named

American Nurses' Foundation research grants for 1986 have been awarded to 27 registered nurses. Recipients are:

Donna Lee Algase, MSN, RN, Case Western Reserve University, "Prevalence and Temporal Pattern of Wandering"; Elaine Bagley, PhD, RN, University of California, San Francisco Hospital, "Bagley Support Inventory: Continued Development and Testing"; Janet L. Blenner, PhD, RN, San Diego State University, "Stimulus Intensity Modulation Scale: Development"; Sharon Ann Brown, MN, RN, University of Texas, "Effects of Educational Interventions on Knowledge, Self-Care Behaviors, and Metabolic Control in Diabetic Adults: A Meta-Analysis of Findings"; Karen W. Budd, PhD, RN, Case Western Reserve University, "Preterm Delivery: Predictors Beyond the Socioeconomic Status"; Marci-Lee Catanzaro, PhD, RN, University of Washington, "Role Identity and Chronic Illness During Midlife"; Kathleen M. Costa, MSN, RN, St. Joseph's Hospital, "A Comparison of Colony Counts of Breast Milk Utilizing Two Methods of Breast Cleansing"; Karen E. Dennis, PhD, RN, Francis Scott Key Medical Center, "Control Preference and Intervention Congruence"; Margaret P. Gagne, MSN, RN, Medical Center Hospital of Vermont, "Diabetes and Breast-Feeding: An Exploratory Study."

Marcia M. Grant, MSN, RN, University of California, San Francisco, "Supporting Oral Intake During Radiation Therapy"; Deborah Ann Gross, DNSc, RN, Pace University, "Maternal Confidence: Comparing Pre-Term/Full-Term Groups"; Pamela Hermansdorfer, BSN, RN, University of Florida, "Nursing Administration Research Themes and Methods"; Carol A. Landis, MS, RN, University of California, "Contribution of Pain to Disturbed Sleep in Arthritic Rats"; Jane G. Llewellyn, DNSc, RN, Rush-Presbyterian-St. Luke's Medical Center, "Outcomes of Same Day Admission Surgery"; Nancy K. Lowe, PhD, RN, Northern Illinois University, "Testing the Theoretical Structure of the McGill Pain Questionnaire in Parturient Women"; Jean A. Massey, PhD, RN, University of South Carolina, "Development of a Psychosomatic Self-Regulation Strategy Inventory"; Arlene C. Miller, MSN, RN, University of Cincinnati Medical Center, "Muralvision® as Distraction for Control of Pain"; Judith A. Fitzerald Miller, MSN, RN, Marquette University, "Psychometric Evaluation of an Instrument to Measure Hope."

Ellen F. Olshansky, DNSc, RN, C, University of Washington, "The Meaning of Infertility to a Marital Relationship"; Theresa Overfield, PhD, RN, Brigham Young University, "Lifestyle and Health"; Constance A. Reid, MS, RN, Edward Hines Jr. VA Hospital, "Efficacy of a Group Exercise Program in Long-Term Care"; Donna Lynn Rew, EdD, RN, University of Texas at Austin, "Intuitive Experiences of Nurses in Clinical Nursing Practice"; Barbara Sachs, PhD, RN, University of Kentucky, "Maternal Health Risks and Neonatal Outcomes"; Sharon L. Scandrett, PhD, RN, University of Tennessee, "Endogenous Healing Process Among Elderly"; Sharon Williams Utz, PhD, RN, Medical College of Ohio, "Health State and Self-Care in Subjects with Mitral Valve Prolapse"; William R. Whetstone, PhD, RN, University of Missouri—Columbia, "Health Promotion of Older Adults: Perceived Barriers"; Nancy Margaret Wineman, MSN, RN, University of Rochester, "Correlates of Adaptation in Multiple Sclerosis."

Over $62,000 was awarded in 1986. Support for the program came from members of the ANF Century Club and corporate contributers: the Allstate Foundation, American Journal of Nursing Company, American Nurses' Association, American Organization of Nurse Executives, Bristol-Meyers Fund, Burroughs Wellcome Fund, C.V. Mosby Company, Deluxe Check Printers, March of Dimes Birth Defects Foundation, *Nursing '86*, Pfizer Pharmaceuticals, and Sterling Drug, Inc.

Rosemary Ellis

Rosemary Ellis, PhD, FAAN, professor emerita of nursing at the Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, died October 10, 1986.

Dr. Ellis was a long-time friend of and contributor to *Nursing Research*. She served for many years on the Manuscript Review Panel.

She received the B.S. degree in nursing from the University of California at San Francisco, and the Ph.D. degree from the University of Illinois at Chicago. During World War II she was a 1st lieutenant in the Army Nurse Corps. At Case Western Reserve she taught nursing administration, education, and research; she retired in July 1986.

Dr. Ellis received the American Nurses' Foundation's Distinguished Contribution to Nursing Science Award in 1986, and a Congressional citation for her contribution to nursing was awarded posthumously.