Two Quantitative Approaches for Estimating Content Validity

Christine A. Wynd, Bruce Schmidt and Michelle Atkins Schaefer

West J Nurs Res 2003 25: 508
DOI: 10.1177/0193945903252998

Western Journal of Nursing Research, 2003, 25(5), 508-518

Two Quantitative Approaches for Estimating Content Validity[1]

Christine A. Wynd
Bruce Schmidt
Michelle Atkins Schaefer

Instrument content validity is often established through qualitative expert reviews, yet quantitative analysis of reviewer agreements is also advocated in the literature. Two quantitative approaches to content validity estimations were compared and contrasted using a newly developed instrument called the Osteoporosis Risk Assessment Tool (ORAT). Data obtained from a panel of eight expert judges were analyzed. A Content Validity Index (CVI) initially determined that only one item lacked interrater proportion agreement about its relevance to the instrument as a whole (CVI = 0.57). Concern that higher proportion agreement ratings might be due to random chance stimulated further analysis using a multirater kappa coefficient of agreement. An additional seven items had low kappas, ranging from 0.29 to 0.48 and indicating poor agreement among the experts. The findings supported the elimination or revision of eight items. Pros and cons to using both proportion agreement and kappa coefficient analysis are examined.

Keywords: *content validity; instrument development; osteoporosis risk assessment; kappa coefficients; proportion agreement*

Empirical research is based on systematic examination of conceptual abstractions through measurable and observable responses. This process is used to identify and explicate phenomena of interest to a discipline. Content validity is an essential step in the development of new empirical measuring devices because it represents a beginning mechanism for linking abstract concepts with observable and measurable indicators. Content validity is

Christine A. Wynd, Ph.D., R.N., C.N.A.A., Professor of Nursing, University of Akron College of Nursing; *Bruce Schmidt*, Ph.D., R.N., Director of Nursing Staff Development and Research, Akron General Medical Center; *Michelle Atkins Schaefer*, M.S., R.N., Chief, Nursing Education and Staff Development, 914th Combat Support Hospital, Upper Arlington, Ohio.

DOI: 10.1177/0193945903252998

defined as the extent to which an instrument adequately samples the research domain of interest when attempting to measure phenomena (Carmines & Zeller, 1979; Waltz, Strickland, & Lenz, 1991).

Carmines and Zeller (1979) identified two interrelated steps in this process: (a) identifying the entire domain of content related to the phenomena of interest beginning with a thorough review of literature and (b) developing instrument items associated with the identified domain of content. These authors go on to state that there is “no agreed upon criterion for determining the extent to which a measure has attained content validity” (p. 22), indicating the absence of rigorous and objective measures for achieving content validity. The resulting instrument content validity is based mainly on the judgment, logic, and reasoning of the researcher with validation from a panel of judges holding expertise in the domain of content.

Over the years, researchers began to identify a need to test the “fit” of qualitatively derived items with their domains of content, and efforts were made to find more quantifiable methods for determining content validity. The purpose of this article is to compare and contrast two quantitative approaches for estimating content validity using the development of a new instrument as an example. The Content Validity Index (CVI), or proportion agreement method, is analyzed and compared to the multirater kappa coefficient of agreement (Brennan & Hays, 1992; Cohen, 1960; Fleiss, 1971; Topf, 1986). Both proportion agreement and the kappa coefficient are examined for utility and accuracy in estimating multirater agreement about content validity of the Atkins Osteoporosis Risk Assessment Tool (ORAT) (Atkins, 1996). A full description of the ORAT’s content validity and use of a panel of experts is provided elsewhere (Wynd & Atkins Schaefer, 2002).

Lynn (1986) advocated a two-stage process for estimating content validity in new instruments. The first stage, or “Developmental Stage,” identifies the domain of content through a comprehensive literature review followed by generation of the instrument items. Construction of the entire instrument then occurs including instructions to respondents and scoring mechanisms.

An objective method for quantitatively measuring content validity is then incorporated into the second stage, or “Judgment/Quantification Stage,” when a select panel of content experts evaluates the instrument and rates item relevance to the domain of content. During this evaluation, the experts may use a Likert-type rating scale. The proportion of experts who are in agreement about item relevance provides a quantitative measure of content validity, the CVI, which has become very popular with nurse researchers (Anders, Tomai, Clute, & Olson, 1997; Summers, 1993).

PROPORTION AGREEMENT AS AN INDEX OF INTERRATER AGREEMENT ABOUT CONTENT VALIDITY

The CVI, a proportion agreement procedure, allows two or more raters to independently review and evaluate the relevance of a sample of items to the domain of content represented in an instrument. A researcher then tallies the proportion of cases in which the raters agree and determines the stability of their agreement (Lynn, 1986). A Likert-type, ordinal scale with four possible responses is used. The responses include a rating of 1 = *not relevant*, 2 = *somewhat relevant*, 3 = *quite relevant*, and 4 = *very relevant*. Researchers advocating the use of this approach specify that ratings of 1 and 2 are considered "content invalid," whereas ratings of 3 and 4 are considered to be "content valid" (Lynn, 1986; Waltz & Bausell, 1983; Waltz et al., 1991). Waltz and Bausell (1983) indicated "the actual CVI is the proportion of items that received a rating of 3 or 4 by the experts" (p. 384). Researchers are then instructed to collapse four ordinal response rankings into two dichotomous categories of responses ("content invalid" and "content valid") and the CVI becomes a two-category nominal scale (Lynn, 1986; Waltz & Bausell, 1983; Waltz et al., 1991).

Many nurse researchers promote use of the CVI for estimating quantitative evidence of content validity (Davis, 1992; Lynn, 1986; Summers, 1993; Waltz et al., 1991); however, the CVI utilizes proportion agreement, which has been criticized by investigators and statisticians over the past three decades. Cohen (1960) was the first to identify the disadvantages of proportion agreement and described this technique as a "most primitive approach" (p. 38). Proportion agreement lacks a value indicating "no agreement," thereby creating the potential for inflation of agreement due to chance (Garvin, Kennedy, & Cissna, 1988; Suen & Ary, 1989; Topf, 1986; Waltz et al., 1991).

The limitations of CVI, or proportion agreement, are further analyzed by Waltz and Bausell (1983) who reiterate Cohen's (1960) concern about chance inflation of agreement and discuss the dependence of agreement on the number and combinations of categories used in the rating scheme. Because only two categories are examined, random chance agreement could be high.

Lynn (1986) argues that limitations identified by Waltz and Bausell (1983) are overcome by employing larger numbers of experts (a minimum of five) and establishing a four-level, Likert-type rating scheme. An adequate number of experts is determined by applying a standard error of the proportion. Lynn indicates that this approach decreases the likelihood for chance agreement because it brings the expert ratings closer to a normal distribution. Lynn also advocates use of a 4-point scale as superior to smaller or larger scales that include an indecisive middle score (e.g., neutral). Lynn's argument becomes moot because researchers using the CVI are instructed to examine the four Likert-type responses as two nominal, dichotomous categories (content invalid and content valid). The four ordinal responses disappear and do not discriminate among varying levels of agreement. Collapsing the four rating levels into dichotomous categories increases the possibility that the judges will agree by chance alone 50% of the time, no matter how many judges are used (Garvin et al., 1988; Topf, 1986), and there may also be a potential loss of important information when the original ordinal scale is no longer available.

Tinsley and Weiss (1975) further criticize the use of proportion agreement with rationale that is directly in contrast to Lynn's (1986) arguments. These statisticians claim that proportion agreement "overestimates the true absolute agreement by an amount related to the number of raters and the number of points on the scale" (Tinsley & Weiss, 1975, p. 366). An increased number of experts (raters, observers, or judges) and a larger number of categories for data assignment yield greater absolute agreement and increase the risk of chance agreement. The use of more experts and a 4-point scale, as advocated by Lynn (1986), may therefore contribute directly to chance agreement. When there are frequent, similar ratings, proportion agreement is often an inappropriate index because the number of "hidden" disagreements influences the total proportion of agreement and creates spurious inflation (Topf, 1986; Wakefield, 1980).

THE MULTIRATER KAPPA STATISTIC AS AN INDEX OF INTERRATER AGREEMENT

Concerns about proportion agreement, as outlined above, lead many statisticians to recommend Cohen's (1960) coefficient kappa (*k*) for examining interrater agreement. The kappa statistic represents the proportion of agreement remaining after chance agreement is removed. Rather than comparing the total proportion of agreements (P_o) to a maximum value of 100%, the total is compared to a maximum possible value that accounts for agreements occurring by chance alone ($1 – P_e$), given the marginal distribution of item ratings assigned by each expert panelist (Musch, Landis, Higgins, Gilson, &

Jones, 1984). P_e is the proportion of agreements expected to occur by chance alone, and (P_o – P_e) represents the observations for which there are "real" agreements versus chance agreements.

$$k = \frac{P_o - P_e}{1 - P_e}$$

The literature is replete with debates about extensions of kappa beyond Cohen's original intentions, but kappa is mainly used to test interrater agreement among observers who rate dichotomous categories of data (Landis & Koch, 1977; Suen & Ary, 1989). The use of kappa with polytomous categorical data or ordinal data is not recommended because kappa measures the frequency of exact agreement versus approximate agreement, and the value of kappa is highly reliant on definitions of the categories. If there are more than two categories of nominal data, differences among the pairs of data will cause varying levels of disagreement between observers or judges (Hutchinson, 1993; Maclure & Willett, 1987). When ordinal data are used, the distances between categories may contribute to disagreements between judges. Some statisticians advocate the use of a weighted kappa with ordinal or polytomous nominal data; however, the weights are often randomly assigned thus creating an arbitrary statistic. The solution is to assign standard weights, but this solution closely approximates intraclass correlations used to measure associations found in interrater reliability, not interrater agreement. Maclure and Willett (1987) make the distinction between these two procedures by stating that interrater reliability is the amount of proportion that deviates from the means as different experts rate an item, whereas interrater agreement constitutes exact agreement in the ratings made by different experts. Interrater agreement is the indicated procedure for quantitatively estimating the content validity of new instruments.

In nursing research, studies often include two or more judges categorizing items or observations. A recommended modification of kappa is the multirater kappa, which pairs the raters' scores and sums the pairs of agreements and disagreements. An overall measure of agreement is provided based on an average of the pairwise agreements (Antonakos & Colling, 2001; Bishop, Feinberg, & Holland, 1975; Fleiss, 1971; Siegel & Castellan, 1988).

Kappa values range from +1.00 to –1.00, with a positive kappa indicating interrater agreement occurring more frequently than would be expected by chance. A +1.00 demonstrates complete agreement across raters. A zero kappa indicates that agreements are no more than can be expected by

TABLE 1: A Comparison of Magnitude Parameters for Kappa Coefficients

Landis & Koch (1977)		*Cicchetti (1984); Fleiss (1971)*	
Strength of Agreement	*Kappa Statistic*	*Strength of Agreement*	*Kappa Statistic*
Poor	< .00	Poor	< .40
Slight	.00-.20	Fair	.40-.59
Fair	.21-.40	Good	.60-.74
Moderate	.41-.60	Excellent	.75-1.00
Substantial	.61-.80		
Almost perfect	.81-1.00		

chance. Negative kappas reveal that raters agree less frequently than can be expected by chance, and indeed, raters may even disagree more frequently than expected in a random fashion. A coefficient of –1.00 indicates total disagreement (Suen & Ary, 1989). A minimally acceptable kappa of 0.60 for interrater agreement was recommended by Gelfand and Hartmann (1975), and many researchers use this as their measurement rule (Phillips, Castorr, Prescott, & Soeken, 1992). Landis and Koch (1977) also provided benchmarks for various levels of kappa magnitude and strength of agreement, whereas Cicchetti (1984) and Fleiss (1971) assigned a separate set of parameters. Table 1 illustrates the two major variations of kappa ranges.

Although kappa appears to be an improved measure of agreement over proportion agreement, it too can be difficult to interpret. Kappa is sensitive to the number of observations made, the distribution of the data, and the presence of bias among observers. For these reasons, a kappa may be low despite higher values of proportion agreement (Banerjee & Fielding, 1997; Brennan & Hays, 1992).

CONTENT VALIDITY ESTIMATIONS FOR THE ORAT

Lynn's (1986) two-stage approach was used to estimate the content validity of the paper-and-pencil screening tool ORAT. The original tool is provided for review at the end of this manuscript. The ORAT was designed as a simple screening mechanism for determining preliminary risk for osteoporosis. Levels of risk, as measured by the ORAT, assist health care providers and patients in determining the need for education, intervention, or more extensive types of definitive diagnosis, such as bone mineral density testing, or bone densitometry. A full description of the ORAT's content validation process is provided by Wynd and Atkins Schaefer (2002).

During the initial, developmental stage of content validity (Lynn, 1986), a thorough review of the literature established a domain of content about osteoporosis risk. Twenty-three items for the ORAT were generated to assess osteoporosis risk factors such as age, race, gender, previous diagnosis of osteoporosis, past fractures of the hip, spine, or wrist, onset of menopause, diet, alcohol and tobacco consumption, prescription medication usage, estrogen replacement therapy, and calcium supplements.

The Judgment/Quantification stage (Lynn, 1986) then required that a panel of experts review the ORAT. Eight experts were selected from nationally known clinicians and researchers holding well-respected reputations in the area of osteoporosis risk prevention and treatment (Wynd & Atkins Schaefer, 2002).

Procedures for content validation were adapted from those described by several researchers (Lynn, 1986; Martuza, 1977; Waltz & Bausell, 1983; Waltz et al., 1991). In addition to the ORAT itself, experts were provided with a relevance rating scale to quantitatively rate instrument item relevance to the domain of content about osteoporosis risk factors. The rating scheme for content relevance of the overall ORAT instrument and its individual items included the 4-point ordinal scale, described earlier. Experts were also asked to share qualitative comments regarding the ORAT items and the overall tool (Wynd & Atkins Schaefer, 2002).

FINDINGS

In the literature, several authors provide information about the magnitude or the amount of proportion that is sufficient for indicating higher levels of interrater proportion agreement. An average agreement of 70% (0.70) is "necessary" for agreement, 80% (0.80) for "adequate" agreement, and 90% (0.90) for "good" agreement (Hartmann, 1977; House, House, & Campbell, 1981).

Fifteen out of 23 items received expert panelist ratings of 1 or 2 on the Likert-type scale indicating high content validity and establishing an overall instrument CVI of 0.65 (prior to elimination or revision of items). Eight items, assessing medication and caffeine intake, received several expert panelist ratings of 3 or 4 on the Likert-type scale and had CVIs equal to 0.86. These items were either eliminated or revised. Only one ORAT item, which assessed a past diagnosis of kidney stones, received a CVI below the acceptable level of relevance (0.57) and was immediately eliminated.

A majority of the experts rated each instrument item as "relevant." Statisticians point out that higher category frequencies are generally associated with higher interrater agreements as the number of rated observations is increased. As a result, agreement due to chance is enhanced, particularly if rater variability is low (Hartmann, 1977; Soeken & Prescott, 1986; Wakefield, 1980; Yelton, Wildman, & Erickson, 1977). Consistent ratings by experts are therefore often due to chance, and these ratings incorrectly indicate higher levels of agreement (Suen & Ary, 1989).

Due to concerns about the risk of chance agreement among the experts, a second analysis of interrater agreement was undertaken. The multirater kappa statistic was used and interrater agreement was reexamined.

Kappa has statistical properties that reflect formal reliability theory regarding the stability of measures. Based on the statistic obtained, significance levels can be determined for lower levels of agreement and are useful for hypothesis testing (Landis & Koch, 1977; Topf, 1986). Results of the analysis revealed a multirater kappa equal to 0.039 (p = .50), representing very little agreement among the expert panelists. The null hypothesis, that the observed value of rater agreement was greater than the expected chance agreement, was not rejected. The experts' interrater agreement about the relevance of ORAT items was low and most likely due to chance.

Individual, adjusted, multirater kappas were examined for each of the 23 items and ranged from k = 0.29 to 0.71. A decision was made to reword or eliminate items scoring below k = 0.48 because they were identified by the experts as having very low relevance to osteoporosis risk. Items receiving lower kappa coefficients were consistent with items having lower CVI ratings; therefore, items eliminated from the ORAT included questions about previous kidney stones, caffeine intake, and the use of medications such as heparin, cyclosporines, antacids, and barbiturates. Items retained in the original ORAT included age, gender, body mass index (height and weight), previous fractures, diagnosis of thyroid disease, use of thyroid replacement medication, estrogen replacement therapy, weight-bearing exercise, family history of fractures, age at onset of menopause, use of calcium supplements, diet of calcium-rich foods, alcohol and tobacco consumption.

DISCUSSION

Quantitative methods, used to confirm the content validity of new instruments, increase the amount of information available for examining psychometrics. Proportion agreement and the kappa coefficient of agreement provide quantifiable methods for evaluating the judgments of content experts. Kappa offers additional information beyond proportion agreement because it removes random chance agreement. For a better understanding of interrater agreement in general, and to increase confidence in the content validity of new instruments, researchers should report both the proportion agreement, as an indication of data variability, and the kappa as a measure of agreement beyond chance (Brennan & Hays, 1992).

In the current study, these quantitative methods led to an examination of instrument items that were relevant to the domain of content about osteoporosis risk, and out of 23 items, 15 items remained quantitatively valid (8 items were eliminated). The next step in the development of the instrument requires a qualitative examination of comments from the expert panelists with revision of the final 15 items according to the experts' responses.

NOTE

1. The authors acknowledge the Ohio Nurses Association for producing the original Atkins Osteoporosis Risk Assessment Tool, and Merck & Company, Inc., for providing unrestricted educational grants to support this research.

REFERENCES

Anders, R. L., Tomai, J. S., Clute, R. M., & Olson, T. (1997). Development of a scientifically valid coordinated care path. *Journal of Nursing Administration*, *27*, 45-52.

Antonakos, C. L., & Colling, K. B. (2001). Using measures of agreement to develop a taxonomy of passivity in dementia. *Research in Nursing and Health*, *24*, 336-343.

Atkins, M. (1996). *Atkins Osteoporosis Risk Assessment Tool*. Columbus: Ohio Nurses Association.

Banerjee, M., & Fielding, J. (1997). Interpreting kappa values for two-observer nursing diagnosis data. *Research in Nursing and Health*, *20*, 465-470.

Bishop, Y. M. M., Feinberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Brennan, P. F., & Hays, B. J. (1992). The kappa statistic for establishing interrater reliability in the secondary analysis of qualitative clinical data. *Research in Nursing and Health*, *15*, 153-158.

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

Cicchetti, D. V. (1984). On a model for assessing the security of infantile attachment: Issues of observer reliability and validity. *Behavioral and Brain Sciences*, *7*, 149-150.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, *5*, 194-197.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378-382.

Garvin, B. J., Kennedy, C. W., & Cissna, K. N. (1988). Reliability in category coding systems. *Nursing Research*, *37*, 52-55.

Gelfand, D. M., & Hartmann, D. P. (1975). *Child behavior analysis and therapy*. New York: Pergamon.

Hartmann, D. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, *10*, 103-116.

House, A., House, B., & Campbell, M. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. *Journal of Behavioral Assessment*, *3*, 37-57.

Hutchinson, T. P. (1993). Kappa muddles together two sources of disagreement: Tetrachoric correlation is preferable. *Research in Nursing and Health*, *16*, 313-315.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 1159-1174.

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

Maclure, M., & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, *126*, 161-169.

Martuza, V. R. (1977). *Applying norm-referenced and criterion-referenced measurement in education*. Boston: Allyn & Bacon.

Musch, D. C., Landis, J. R., Higgins, I. T. T., Gilson, J. C., & Jones, R. N. (1984). An application of kappa-type analysis to interobserver variation in classifying chest radiographs for pneumoconiosis. *Statistics in Medicine*, *3*, 73-83.

Phillips, C. Y., Castorr, A., Prescott, P. A., & Soeken, K. (1992). Nursing intensity: Going beyond patient classification. *Journal of Nursing Administration*, *22*, 46-52.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Soeken, K. L., & Prescott, P. A. (1986). Issues in the use of kappa to estimate reliability. *Medical Care*, *8*, 733-741.

Suen, H., & Ary, D. (1989). *Analyzing quantitative observation data*. Hillsdale, NJ: Lawrence Erlbaum.

Summers, S. (1993). Establishing the reliability and validity of a new instrument: Pilot testing. *Journal of Post Anesthesia Nursing*, *8*, 124-127.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*, 358-376.

Topf, M. (1986). Three estimates of interrater reliability for nominal data. *Nursing Research*, *35*, 253-255.

Wakefield, J. (1980). Relationship between two expressions of reliability: Percentage agreement and phi. *Educational and Psychological Measurement*, *40*, 593-597.

Waltz, C., & Bausell, R. B. (1983). *Nursing research: Design, statistics, and computer analysis*. Philadelphia: F. A. Davis.

Waltz, C. F., Strickland, O., & Lenz, E. (1991). *Measurement in nursing research* (2nd ed.). Philadelphia: F. A. Davis.

Wynd, C. A., & Atkins Schaefer, M. (2002). The Osteoporosis Risk Assessment Tool: Establishing content validity through a panel of experts. *Applied Nursing Research*, *16*, 184-188.

Yelton, A., Wildman, B., & Erickson, M. (1977). A probability-based formula for calculating interobserver agreement. *Journal of Applied Behavior Analysis*, 10, 127-131.