*Research in Nursing & Health*, 2007, 30, 459–467

# *Focus on Research Methods*
# Is the CVI an Acceptable Indicator of Content Validity? Appraisal and Recommendations

**Denise F. Polit,[1,2†‡] Cheryl Tatano Beck,[3§] Steven V. Owen[4§]**

[1]Humanalysis, Inc., 75 Clinton Street, Saratoga Springs, NY 12866
[2]Griffith University School of Nursing, Gold Coast, Australia
[3]University of Connecticut School of Nursing, Storrs, CT
[4]School of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, TX
Accepted 9 January 2007

**Abstract:** Nurse researchers typically provide evidence of content validity for instruments by computing a content validity index (CVI), based on experts' ratings of item relevance. We compared the CVI to alternative indexes and concluded that the widely-used CVI has advantages with regard to ease of computation, understandability, focus on agreement of *relevance* rather than agreement per se, focus on consensus rather than consistency, and provision of both item and scale information. One weakness is its failure to adjust for chance agreement. We solved this by translating item-level CVIs (I-CVIs) into values of a modified kappa statistic. Our translation suggests that items with an I-CVI of .78 or higher for three or more experts could be considered evidence of good content validity.  Res Nurs Health 30:459–467, 2007

Keywords: instrument development and validation; methodological research; content validity

Evaluating a scale's content validity (S-CVI) is a critical early step in enhancing the construct validity of an instrument (Haynes, Richard, & Kubany, 1995), and so content validation is an important topic for clinicians and researchers who require high-quality measurements. Content validity concerns the degree to which a scale has an appropriate sample of items to represent the construct of interest—that is, whether the domain of content for the construct is adequately represented by the items (e.g., Waltz, Strickland, & Lenz, 2005). Developers of new scales are increasingly expected to provide evidence that their scale and the items on it are content valid. For example, an editorial in *Research in Nursing & Health* indicated that authors submitting an instrument development manuscript to that journal should include a content validity assessment (Froman & Schmitt, 2003).

Among nurse researchers, the most widely used method of quantifying content validity for multi-item scales is the content validity index (CVI)

Correspondence to Denise F. Polit.
[†]President.
[‡]Adjunct Professor.
[§]Professor.
Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/nur.20199

based on expert ratings of relevance. In a previous aricle on content validity, we critiqued inadequacies in reporting CVI values and offered recommendations about computing the CVI for scales, but noted that our purpose was ''not to advocate for or against using the CVI as the standard index of content validity'' (Polit & Beck, 2006, p. 490). In this paper, we *do* recommend the CVI as an appropriate indicator, based on a comparative evaluation of the CVI and alternative indexes. We consider some of the criticisms that have been made about the CVI and offer a solution to the most persistent complaint about the CVI, its failure to adjust for chance agreement.

## CURRENT STATUS OF THE CVI

A CVI value can be computed for each item on a scale (which we refer to as I-CVI) as well as for the overall scale (which we call an S-CVI). To calculate an item-level CVI (I-CVI), experts are asked to rate the relevance of each item, usually on a 4-point scale. There are several variations of labeling the 4 ordinal points, but the scale that seems to be used most often is 1 = *not relevant*, 2 = *somewhat relevant*, 3 = *quite relevant*, 4 = *highly relevant* (e.g., Davis, 1992). Then, for each item, the I-CVI is computed as the number of experts giving a rating of either 3 or 4, divided by the number of experts—that is, the proportion in agreement about relevance. For example, an item rated as ''quite'' or ''highly'' relevant by four out of five judges would have an I-CVI of .80.

Lynn (1986) provided widely cited guidelines for what an acceptable I-CVI should be in relation to the number of experts. She advocated that when there are five or fewer experts, the I-CVI must be 1.00—that is, all experts must agree that the item is content valid. When there are more than five experts, there can be a modest amount of disagreement (e.g., when there are six experts, the I-CVI must be at least .83, reflecting one disagreement).

As for scale-level content validity, there are alternative ways to compute the S-CVI when there are more than two experts, as is most often the case. Unfortunately, scale developers almost never report which procedure they used (Polit & Beck, 2006). One approach is to require universal agreement among experts, defining the S-CVI as the proportion of items on an instrument that achieved a rating of 3 or 4 by all the content experts. Table 1, which presents fictitious ratings of relevance for three experts on a 10-item scale, can be used to illustrate. According to this definition, the S-CVI for these data would be .70—the three experts agreed universally that 7 out of the 10 items (items 1–7) were content valid. We refer to this approach as S-CVI/UA (universal agreement).

Another approach for the S-CVI is to compute the I-CVI for each item on the scale, and then calculate the average I-CVI across items. In Table 1, all three experts agreed that the first seven items were relevant (I-CVI = 1.00), but had divergent opinions about the last three items, which had I-CVIs of .67. Averaging across the 10 I-CVIs, an approach we refer to as S-CVI/Ave, yields a value of .90. As Polit and Beck (2006) noted, authors of scale development papers almost never indicate which method they used to compute S-CVI. Inasmuch as S-CVI/UA and S-CVI/Ave can yield vastly different values—as demonstrated

**Table 1. Fictitious Ratings on a 10-Item Scale by Three Experts: Items Rated 3 or 4 on 4-Point Relevance Scale**

| Item | Expert 1 | Expert 2 | Expert 3 | Experts in Agreement | Item CVI |
|---|---|---|---|---|---|
| 1 | ✓ | ✓ | ✓ | 3 | 1.00 |
| 2 | ✓ | ✓ | ✓ | 3 | 1.00 |
| 3 | ✓ | ✓ | ✓ | 3 | 1.00 |
| 4 | ✓ | ✓ | ✓ | 3 | 1.00 |
| 5 | ✓ | ✓ | ✓ | 3 | 1.00 |
| 6 | ✓ | ✓ | ✓ | 3 | 1.00 |
| 7 | ✓ | ✓ | ✓ | 3 | 1.00 |
| 8 | — | ✓ | ✓ | 2 | .67 |
| 9 | — | ✓ | ✓ | 2 | .67 |
| 10 | ✓ | — | ✓ | 2 | .67 |
| | | | | Average I-CVI = | .90 |
| Proportion relevant | .80 | .90 | 1.00 | | |

I-CVI, item-level content validity index; scale-level content validity index, universal agreement method (S-CVI/UA) = .70; scale-level content validity index, averaging method (I-CVI/Ave) = .90; average proportion of items judged relevant across the three experts = .90.

in Table 1—it is important to make clear which method was used.

Scale developers often use a criterion of .80 as the lower limit of acceptability for an S-CVI. Most writers use this standard either without attribution, or they cite Davis (1992), who stated, without an acknowledged rationale, ''For new instruments, investigators should seek 80% or better agreement among reviewers'' (p. 197). If the standard of .80 were applied to the two methods of computing an S-CVI for the data in Table 1, it would be concluded that the content validity of the10-item scale was *not* adequate using the S-CVI/UA approach (.70), but that it *was* adequate using the S-CVI/Ave approach (.90).

In addition to the confusion about calculating the S-CVI and establishing a defensible standard of acceptability, the CVI as an approach to content validation has not been without critics. Various alternatives to the CVI have been discussed, all of which use expert judgments and the calculation of an index of inter-rater agreement or reliability.

## SOME ALTERNATIVES TO THE CVI

The CVI is an index of inter-rater agreement. There are literally dozens of methods of calculating the degree to which two or more raters are consistent or congruent in their ratings (Stemler, 2004). As described by Stemler, the various methods can be classified into one of three categories: consistency estimates, consensus estimates, and measurement estimates. Proposed alternatives to the CVI have all been indexes that yield consistency or consensus estimates.

Consistency estimates focus on the extent to which the experts are consistent (reliable) in their application of the rating scale for item relevance. One approach to quantifying consistency is coefficient alpha, an approach suggested by Waltz et al. (2005) as a method of calculating content validity when there are multiple experts. In its favor, alpha considers all the variation in expert responses, whereas the CVI shrinks variation by collapsing four rating categories into a dichotomy of relevant/not relevant. There are, however, problems with using coefficient alpha (or other mathematically similar consistency indexes such as the intraclass correlation coefficient) for content validity purposes. For one thing, alpha is computed across all items and experts, and thus provides limited information for evaluating individual items or individual judges. A second limitation of alpha is that a high value of alpha can be obtained even in situations in which agreement about content validity is low. As noted by Waltz and colleagues, a high alpha value ''does not mean that the same rating was assigned by all experts, but rather that the *relative ordering* or ranking of scores assigned by one expert matches the relative order assigned by other experts'' (p. 156; emphasis added). As an extreme example, consider six experts rating four items on the 4-point relevancy scale. Suppose three experts gave ratings of 1, 1, 2, and 2 to the items, and the other three experts gave ratings of 3, 3, 4, and 4 to the same items. In this situation, the value of alpha would be 1.00, despite noteworthy disagreement across experts, and ratings of low-relevancy of all items by three experts. For the ratings just described, the S-CVI/Ave would be .50, which most would consider unacceptably low. Consistency indexes such as coefficient alpha focus on internal consistency (i.e., inter-rater reliability) and not on agreement across experts.

The CVI, and most other indexes that have been proposed for content validity, fall into the category that Stemler (2004) called consensus estimates. Consensus estimates are concerned with the extent to which the experts ''share a common interpretation of the construct'' (Stemler, p. 2) and are able to agree on how to apply the rating scale to the items. A low level of agreement presumably reflects a failure to demonstrate a shared understanding of the construct.

A widely used approach for computing a consensus estimate is to calculate the proportion in agreement, as in the CVI. Critics of the CVI (and of the simple proportion in agreement) are most likely to be concerned about the possibility of inflated values because of the risk of chance agreement, an issue discussed more fully later in this article. The kappa statistic, a consensus index of inter-rater agreement that adjusts for chance agreement, has been suggested as a measure of content validity. For example, Wynd, Schmidt, and Schaefer (2003) used both the CVI and a multi-rater kappa coefficient in the content validation of their Osteoporosis Risk Assessment Tool. They argued that the kappa statistic was an important supplement to (if not substitute for) the CVI because kappa provides information about degree of agreement beyond chance.

Several writers in the field of personnel psychology have proposed other consensus methods to evaluate inter-rater agreement for content validity purposes, as summarized by Lindell and Brandt (1999). These include the indexes referred to as $T$ (Tinsley & Weiss, 1975); $r_{WG}$ (James, Demaree, & Wolfe, 1984), and $r^*_{WG}$ (Lindell,

Brandt, & Whitney, 1999). Another index was suggested by Lawshe (1975), who proposed an index of inter-rater agreement for scale items called the content validity ratio (CVR), used with dichotomous ratings on items. Lawshe recommended averaging CVRs (which use a different formula than the I-CVI formula discussed earlier) to yield an overall *content validity index*. All of these indexes make adjustments for chance agreement, and all but one (the *T* index) provide diagnostic information about both items and the overall scale. It does not appear that researchers in nursing or other health-related fields have adopted any of these coefficients as indicators of content validity.

## CRITERIA FOR SELECTING A CONTENT VALIDITY INDEX

In selecting an index of expert agreement from the various available alternatives, scale developers might be guided by various criteria, including a focus on consensus rather than consistency estimates, ease of computation, understandability and ease of communication, provision of both item diagnostic information and scale validity information, and adjustments for chance agreement.

The CVI holds up well against most of these criteria—except for adjustments for chance agreement. In terms of computational effort, many of the alternative content validity indexes are more computationally complex. For example, the formula for $r_{WG}$ for dichotomous ratings of one item is: $(2p_A - 1)^2$, where $p_A$ is the proportion of experts in agreement (Lindell & Brandt, 1999). Although ease of computation should not be a primary consideration, it is not an irrelevant one. Topf (1986), for example, cited a study in which alternative ways of computing inter-rater agreement were explained to college students, who were then asked to compute various indexes with some data. For these students, computing the proportion agreement (as in the I-CVI) resulted in significantly fewer computational errors than when computing kappa.

With respect to the criterion of understandability, the CVI is again a reasonable choice because it has strong intuitive appeal. Proportion agreement on the relevance of items is both easily computed and easily understood, and averaging the I-CVIs to compute the scale-level CVI is also easy to communicate to others. By contrast, Lawshe's (1975) CVR, for example, is easy to compute, but not so easy to interpret: its values can range from −1.0 to +1.0, with CVR = 0 when half the experts judge an item to be relevant.

Another desirable quality of a content validity method is for it to yield item-level information that can be used to refine or discard items, and a summary of the content validity of the overall scale. Most consensus-type content validity coefficients, including the CVI, do this, although the index *T* (Tinsley & Weiss, 1975) does not.

Unlike the CVI, the alternative consensus estimates reviewed for this article make adjustments for chance agreement. As we discuss in the next section, however, other indexes base their adjustments on the risk of chance agreement about both relevance and non-relevance—which, in our opinion, is problematic.

## CONTENT VALIDITY AND CHANCE AGREEMENT

The CVI typically used by nurse researchers differs from other content validity indexes in one crucial respect—it captures inter-rater agreement, but not *full* inter-rater agreement. This can be illustrated with an example (Table 2). In this table, which depicts ratings of two judges on the relevance of 10 items, assume that both experts rated the same 5 items as relevant (ratings of 3 or 4), and that both also rated the remaining 5 items as not relevant (ratings of 1 or 2). In this example, the S-CVI (calculated by either the universal agreement or averaging method) is .50: Both judges agreed that 5 of the 10 items are relevant. Inter-rater agreement, however, is 1.0—both experts were in complete agreement regarding the relevance of all 10 items. Thus, general indexes of inter-rater agreement (such as most of those that have been used to estimate content validity) are problematic as measures of content validity because they capture agreement *of any type*,

**Table 2. Fictitious Relevance Ratings for a 10-Item Scale with Two Expert Raters**

| | Expert Rater No. 1 | | |
|---|---|---|---|
| | Items Rated | Items Rated | |
| Expert Rater No. 2 | 1 or 2[a] | 3 or 4[b] | Total |
| Items Rated 1 or 2[a] | 5 | 0 | 5 |
| Items Rated 3 or 4[b] | 0 | 5 | 5 |
| Total | 5 | 5 | 10 |

[a]Ratings of 1 = not relevant; 2 = somewhat relevant.
[b]Ratings of 3 = quite relevant; 4 = highly relevant.

including agreement about the *low* relevance of an item.

Wynd et al. (2003), as previously noted, advocated the use of the multi-rater kappa statistic because, unlike the CVI, it adjusts for chance agreements. Chance agreement is an issue of concern in evaluating indexes of inter-rater agreement, especially when the choices are dichotomous, as is the case when 4-point ratings are collapsed into the two categories of relevant and not relevant. Indeed, Cohen (1960), who developed the kappa statistic, criticized using simple proportion agreement as "primitive" (p. 37). But, like most consensus indexes of inter-rater agreement, kappa captures agreement of all types—that is, consensus about relevance *or* non-relevance of an item. Wynd et al. did not provide actual examples of their computations of multi-rater kappa (which is an extension of Cohen's kappa developed by Fleiss, 1971), but they indicated that they used the basic formula for kappa, which is:

$$k = \frac{\text{Proportion}_{\text{Agreement}} - \text{Proportion}_{\text{Chance agreement}}}{1 - \text{Proportion}_{\text{Chance agreement}}}$$

In this equation, the denominator represents maximum possible agreement over and above what would be predicted by chance. The numerator indicates the agreement actually attained *in excess* of chance. Using this formula, the kappa for the data presented in Table 2 would be 1.0, because the proportion in agreement is 1.0 (both raters agreed on all 10 items), and the proportion for chance agreement of any type is .50, as we discuss subsequently. Thus, kappa $= (1 - .5)/(1 - .5) = 1.0$. Clearly, this calculation obscures the content validity focus on item relevance—it only indicates that there was perfect concurrence among the experts.

We might be closer to achieving a useful index of content validity if we consider the proportion of experts among whom there is agreement of relevance, which is exactly what the CVI does. So, in the above formula, if .5 were used as the proportion in agreement for the data in Table 2 (half the items were judged to be of high relevance by both experts), and if the proportion for chance agreement were left at .50, then kappa would be .0—suggesting that the amount of agreement by these two experts is exactly what one would expect by chance.

Chance agreement could, however, be conceptualized differently in assessing content validity, as illustrated in the following example. Suppose that an infinite number of judges rated the relevance of an item (relevant versus not relevant) totally at random, and that these chance ratings were summarized. This is analogous to flipping two coins simultaneously and determining the probability of heads and tails for the two coins. The result is shown in Figure 1. Two judges would agree by chance half the time (cells A and D) and would disagree by chance half the time (cells B and C). In other words, the likelihood of chance *agreement* is .50. *But chance agreement on relevance occurs only one out of four times* (.25 in cell D). Thus, the likelihood of two experts agreeing by chance alone that the item is relevant is .25, *not* .50. (In the coin analogy, the probability would be .50 that the two coins would both come up on the same side, but the probability would be .25 that both coins would be tails.) This logic can be extended to the 10-item situation shown in Table 2. By chance alone, the experts would agree on 5 of the 10 items and disagree on the other 5—but they would agree on high ratings of relevance for one item out of every four, that is, .25. If we used this value as the probability of chance agreement in the calculation of kappa for the data presented in Table 2, we would conclude that $k = .50$, that is, $(1 - .5)/(1 - .25) = .50$.

It is straightforward to compute the probability of chance agreement by multiple experts, extending our two-judge situation. When the ratings are dichotomous (relevant/not relevant), and when

| | Rater (Coin) 1 | |
|---|---|---|
| Rater (Coin) 2 | Not Relevant (Heads) | Relevant (Tails) |
| Not Relevant (Heads) | A .25 | B .25 |
| Relevant (Tails) | C .25 | D .25 |

**FIGURE 1.** Probabilities for random ratings of item relevance by two raters (toss of two coins).

the focus is on agreement *only for relevance*, the formula for the probability of chance universal agreement is:

$$p_c = .5^N$$

where $p_c$ is the probability of chance universal agreement on relevance and $N$ is the number of judges. When there are two judges, $p_c = .25$, when there are three judges $p_c = .125$. Table 3 summarizes the likelihood of chance universal agreement on relevance for up to nine judges. When there are nine experts, the probability that all nine would agree on good relevance by chance alone is only 1 out of 512, or .002.

Critics of the CVI have worried about adjusting for the possibility of chance agreement, *but not about the possibility of chance disagreements*. Table 3 shows that when there are, for example, five experts, the probability is .938 that there will be at least one disagreement on relevance by chance alone. Of course, in content validation the ratings are not random—we expect ratings to be informed judgments of experts. Nevertheless, it is sobering to realize that achieving total consensus becomes increasingly difficult (and unlikely) as the number of experts increases.

## THE CVI AND CORRECTING FOR CHANCE AGREEMENT

Of the various indexes for assessing content validity, the CVI has a number of attractive features. In addition to performing well in terms of computation ease, understandability, and its provision of both item- and scale-level information, the CVI is an index that provides information that decision-makers need when constructing a scale—the extent to which there is a consensus about the *relevance* of the item to the target construct. Other indexes, except for Lawshe's (1975) CVR, indicate the extent to which the experts reached *any* type of agreement. The CVI has an additional virtue—it is already widely used and accepted, at least among nurse researchers.

The most important improvement to the CVI would be to adapt it to adjust for chance agreement. Lynn (1986) addressed the issue of chance agreement for the CVI. She calculated a standard error of proportions, used the standard error to establish a 95% confidence interval (CI) around each I-CVI value, and then compared the lower limit of the 95% CI to .50, her criterion for chance agreement (M.R. Lynn, personal communication, May 23, 2006). As an example, consider an I-CVI of .71, the value for five out of seven experts agreeing on item relevance. The 95% CI around .71 is (.38, 1.00), using the approximation formula for the standard error of proportions: the square root of $(p^* [1-p]/N)$. The lower confidence limit of .38 is below .50, and therefore deemed not acceptable. When five out of six experts agree on item relevance, the I-CVI is .83. The lower confidence limit for this value is .54, which is deemed acceptable as a greater-than-chance level of agreement.

We disagree with Lynn's approach for two reasons. First, by using .50 as the criterion for chance agreement, she was focusing on the risk of chance agreement of any type, rather than on the risk of chance agreement on relevance. The second issue concerns Lynn's overall method. By building a CI around the CVI proportions, Lynn was estimating a 95% likelihood of capturing the unknown population value within some interval, for a specified number of experts on the panel. For example, in the case in which five out of six experts agreed on an item's relevance, it can be inferred with 95% confidence that the true population proportion lies between .54 and 1.00. Knowing that the ''true'' population proportion could be as low as .54 is hardly reassuring about the item's relevance, even if this value is greater than ratings at random.

**Table 3. Probability of Chance Agreements and Disagreements for Dichotomous Ratings of Relevance, Two to Nine Experts**

| | Number of Experts | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Probability that raters would all agree on relevance[a] | .250 | .125 | .063 | .031 | .016 | .008 | .004 | .002 |
| Probability that raters would all agree on non-relevance[a] | .250 | .125 | .063 | .031 | .016 | .008 | .004 | .002 |
| Probability that raters would have at least one disagreement | .500 | .750 | .875 | .938 | .968 | .984 | .992 | .996 |

[a]Probability of chance universal agreement $= .5^N$, where $N$ = number of experts.

In most measurement situations, researchers are guided not by the criterion of statistical significance, but rather by standards relating to magnitude of coefficients. For example, there are rough standards, or guidelines for acceptability, for coefficient alpha, factor loadings, correlations between items and total scores, and so on. While we agree with Knapp and Brown (1995) that these standards should not be considered sacred, we think researchers are aided in their decision-making by determining whether a value is within a range deemed acceptable by those who have given the matter some thought.

Our approach to addressing the issue of chance agreements for the CVI links values of the CVI to a new kappa-like index that adjusts for chance agreements on relevance, not chance agreements of any type. Unfortunately, different authors have proposed different standards for evaluating kappa. For example, Landis and Koch (1977) suggested that kappa values above .60 are *substantial*, while both Fleiss (1981) and Cicchetti and Sparrow (1981) considered values of .75 or higher to be *excellent*. We used these latter, more conservative, standards in our recommendations for I-CVIs.

Table 4, which summarizes our approach, lists several scenarios for rating the relevance of an item, with varying number of experts and agreements. The table excludes two experts because we agree with Lynn (1986) that three experts should be the minimal acceptable number for a content validation effort. The third column of the table indicates the I-CVIs for each scenario. Scenarios in which the I-CVI would be .50 or less are excluded because these would always be unacceptable.

We then computed a modified kappa statistic (which we call $k^*$) that adjusts each I-CVI in the table for chance agreement. The index is called a *modified* kappa because it is an index of agreement *of a certain type*, namely agreement among the judges that the item is relevant. Agreement about non-relevance is not counted, because such agreement does not inform judgments about the content validity of an item. Then, the standards described in Fleiss (1981) and Cicchetti and Sparrow (1981) were applied to evaluate whether the value for each $k^*$ is fair, good, or excellent. Scale developers can compare their I-CVIs to these standards, without needing to actually calculate the modified kappa.

**Table 4. Evaluation of I-CVIs with Different Numbers of Experts and Agreement**

| (1) | (2) | (3)[a] | (4)[b] | (5)[c] | (6)[d] |
|---|---|---|---|---|---|
| Number of Experts | Number Giving Rating of 3 or 4 | I-CVI | $p_c$ | $k^*$ | Evaluation |
| 3 | 3 | 1.00 | .125 | 1.00 | Excellent |
| 3 | 2 | .67 | .375 | .47 | Fair |
| 4 | 4 | 1.00 | .063 | 1.00 | Excellent |
| 4 | 3 | .75 | .25 | .67 | Good |
| 5 | 5 | 1.00 | .041 | 1.00 | Excellent |
| 5 | 4 | .80 | .156 | .76 | Excellent |
| 6 | 6 | 1.00 | .016 | 1.00 | Excellent |
| 6 | 5 | .83 | .094 | .81 | Excellent |
| 6 | 4 | .67 | .234 | .57 | Fair |
| 7 | 7 | 1.00 | .008 | 1.00 | Excellent |
| 7 | 6 | .86 | .055 | .85 | Excellent |
| 7 | 5 | .71 | .164 | .65 | Good |
| 8 | 8 | 1.00 | .004 | 1.00 | Excellent |
| 8 | 7 | .88 | .031 | .88 | Excellent |
| 8 | 6 | .75 | .109 | .72 | Good |
| 9 | 9 | 1.00 | .002 | 1.00 | Excellent |
| 9 | 8 | .89 | .014 | .89 | Excellent |
| 9 | 7 | .78 | .070 | .76 | Excellent |

[a] I-CVI, item-level content validity index.
[b] $p_c$ (probability of a chance occurrence) was computed using the formula for a binomial random variable, with one specific outcome: $p_c = [N!/A!(N-A)!] * .5^N$ where $N$ = number of experts and $A$ = Number agreeing on good relevance.
[c] $k^*$ = kappa designating agreement on relevance: $k^* = (\text{I-CVI} - \text{pc})/(1 - \text{pc})$.
[d] Evaluation criteria for kappa, using guidelines described in Cicchetti and Sparrow (1981) and Fleiss (1981): Fair = $k$ of .40 to .59; Good = $k$ of .60–.74; and Excellent = $k > .74$.

To compute $k^*$, the probability of chance agreement was first computed. The formula for a binomial random variable was used:

$$p_c = \left[\frac{N!}{A!(N-A)!}\right].5^N$$

where $N$ = number of experts and $A$ = Number agreeing on good relevance. Next, $k^*$ was computed using the proportion of agreements *on relevance* (in other words, the I-CVI) and the probability of chance agreement:

$$k^* = \frac{\text{I-CVI} - p_c}{1 - p_c}$$

The calculations of $p_c$ and $k^*$ for the various rating scenarios are presented in columns 4 and 5, respectively. The final column indicates whether the $k^*$ value is fair, good, or excellent. When there is perfect agreement among the experts, the I-CVI is always considered excellent. After adjustments for chance, any I-CVI greater than or equal to .78 (e.g., seven out of nine experts) would also be considered excellent, which we think is an appropriate goal. If we built a 95% CI around any of the I-CVIs deemed *excellent* in Table 4 (even using the exact formula for the standard error of proportions rather than the previously described approximation), the lower confidence limits would all be greater than .25, that is, better than chance agreement on relevance.

As it turns out, our conclusions are consonant with those of Lynn (1986), with one exception. Lynn's guidelines would have caused scale developers to eliminate or revise an item in which four out of five experts gave a rating of *relevant*, but our guidelines do not. Lynn recommended 100% agreement with fewer than six experts, but our guidelines require perfect agreement only when there are three or four experts.

Scale developers could compute values for $k^*$ with 10 or more experts using the formulas provided, but there is little need to do so. An inspection of column 4 indicates that, as the number of experts grows larger, the probability of chance agreement diminishes. As a result, the values of I-CVI and $k^*$ converge with an increasing number of experts. With 10 or more experts, any I-CVI value greater than .75 yields a $k^*$ greater than .75. This means that an I-CVI of .75 would be considered "excellent" with 16 experts, but not with 4 or 8. The safest generalization, then, is that any I-CVI greater than .78 would fall into the range considered excellent, regardless of the number of experts.

## RECOMMENDATIONS

The CVI is a plausible method of estimating the content validity of a new (or revised) scale. Scale developers should recognize, however, that when using a measure of inter-rater agreement such as the CVI, all aspects of the situation are being evaluated. If the value of the CVI is low, it could mean that the items were not good operationalizations of the underlying construct, that the construct specifications or directions to the experts were inadequate, or that the experts themselves were biased, erratic, or not sufficiently proficient. This implies that, at the beginning of the content validation process, scale developers must work hard to develop good items and construct specifications and to select a strong panel of experts.

Like Lynn (1986) and Haynes et al. (1995), we support the concept of multiple iterations in a content validity effort, following a rigorous domain analysis and development process that might involve input from the target population. The first iteration of expert content validation would ideally involve review by a large panel of experts—perhaps 8–12 of them. Experts should be selected with care, using well-defined criteria such as those proposed by Grant and Davis (1997). The focus of the first round would be on the items—discovering which ones needed to be revised or discarded, getting advice about whether additional items are needed to adequately tap the domain of interest, and making efforts to determine if aspects of the construct are represented by items in correct proportions. I-CVI values would then guide decisions about item revisions or rejections, and the experts' comments would guide the development of any new items. Based on the information in Table 4, items with an I-CVI somewhat lower than .78 would be considered candidates for revision, and those with very low values would be candidates for deletion.

Unless only minor item revisions are needed based on the first round results, a second round of expert review should be conducted. In the second round, a smaller group of experts (perhaps 3–5) can be used to evaluate the relevance of the revised set of items and to compute the S-CVI. Lynn (1986) noted that the raters can be drawn from the same pool of experts as in the first round, or they can also be a new panel. Using a subset of experts from the first round has distinct advantages, however, because then information from the first round can be used to select the most capable judges. For example, data from round 1 could be used to eliminate experts who were consistently lenient (e.g., who gave ratings of 4 to all items) or

consistently harsh, or whose ratings were incongruent with those of most other experts. Qualitative feedback from an expert in round 1 in the form of useful comments about the items might indicate both content capability and a strong commitment to the project.

Once the second-round panel is selected and new ratings of relevance are obtained for the revised set of items, the S-CVI can be computed. As noted in our other article (Polit & Beck, 2006), we think the S-CVI/UA calculation method is overly stringent. Acceptable values for S-CVI/UA become more difficult to achieve as the number of experts increases. This approach ignores the risk of chance disagreements (see Table 3)—not to mention non-chance disagreements if an expert is biased or has misunderstood the construct specifications. The S-CVI/Ave is attractive not only because it avoids these problems but also because it inherently embodies information about the performance of each item through the averaging feature.

We think that the goal for the S-CVI/Ave should be .90, consistent with recommendations made by Waltz et al. (2005) for the mathematically identical index called the average congruency percentage or ACP. With a standard of .90 for the S-CVI/Ave, the scale would be composed of some items on which there was complete agreement (I-CVI = 1.00) and a few items on which there was a modest amount of disagreement (i.e., I-CVIs of at least .78). Thus, our recommendation falls somewhere in between the conservatism of a minimum S-CVI of .80 for the universal agreement approach and the liberalism of a minimum S-CVI of .80 for the averaging approach.

In summary, we recommend that for a scale to be judged as having excellent content validity, it would be composed of items that had I-CVIs of .78 or higher and an S-CVI/Ave of .90 or higher. This requires strong conceptual and developmental work, good items, outstanding experts, and clear instructions to the experts regarding the underlying constructs and the rating task. The effort is worth it.

## REFERENCES


Cicchetti, D.V., & Sparrow, S. (1981). Developing criteria for establishing interrater reliability of specific items: Application to assessment of adaptive behavior. American Journal of Mental Deficiency, 86, 127–137.

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.

Fleiss, J. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley.

Froman, R.D., & (2003). Schmitt, M.H. Thinking both inside and outside the box on measurement articles. Research in Nursing & Health, 26, 335–336.

Grant, J.S., & Davis, L.L. (1997). Selection and use of content experts in instrument development. Research in Nursing & Health, 20, 269–274.

Haynes, S., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. Psychological Assessment, 7, 238–247.

James, L., Demaree, R., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology, 69, 85–98.

Knapp, T.R., & Brown, J. (1995). Ten measurement commandments that often should be broken. Research in Nursing & Health, 18, 465–469.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. Biometrics, 53, 159–174.

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

Lindell, M.K., & Brandt, C.J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of the CVI, T, $r_{WG(J)}$, and $r^*_{WG(J)}$ indexes. Journal of Applied Psychology, 84, 640–647.

Lindell, M.K., Brandt, C.J., & Whitney, D.J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. Applied Psychological Measurement, 23, 127–135.

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

Polit, D.F., & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29, 489–497.

Stemler, S. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research, and Evaluation, 9(4). Retrieved October 5, 2006 from http://PAREonline.net/getvn.asp?v= 9&n=4.

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.

Waltz, C.F., Strickland, O.L., & Lenz, E.R. (2005). Measurement in nursing and health research (3rd ed.) New York: Springer.

Wynd, C.A., Schmidt, B., & Schaefer, M.A. (2003). Two quantitative approaches for estimating content validity. Western Journal of Nursing Research, 25, 508–518.