Research in Nursing & Health, 2006, 29, 489–497

The Content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations

Denise F. Polit,[1,2]* Cheryl Tatano Beck[3]**

[1]Humanalysis, Inc., Saratoga Springs, NY
[2]Griffith University School of Nursing, Gold Coast, Australia
[3]University of Connecticut School of Nursing, Storrs, CT
Accepted 16 May 2006

Abstract: Scale developers often provide evidence of content validity by computing a content validity index (CVI), using ratings of item relevance by content experts. We analyzed how nurse researchers have defined and calculated the CVI, and found considerable consistency for item-level CVIs (I-CVIs). However, there are two alternative, but unacknowledged, methods of computing the scale-level index (S-CVI). One method requires universal agreement among experts, but a less conservative method averages the item-level CVIs. Using backward inference with a purposive sample of scale development studies, we found that both methods are being used by nurse researchers, although it was not always possible to infer the calculation method. The two approaches can lead to different values, making it risky to draw conclusions about content validity. Scale developers should indicate which method was used to provide readers with interpretable content validity information. © 2006 Wiley Periodicals, Inc. Res Nurs Health 29:489–497, 2006

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

When a new scale is developed, researchers following rigorous scale development procedures are expected to provide extensive information about the scale's reliability and validity. Although the criterion-related and construct validity of a new instrument are considered especially important, information about the content validity of the measure is also viewed as necessary in drawing conclusions about the scale's quality. Content validity has been defined as follows:

(1) "...the degree to which an instrument has an appropriate sample of items for the construct being measured" (Polit & Beck, 2004, p. 423);
(2) "...whether or not the items sampled for inclusion on the tool adequately represent the

Correspondence to Denise F. Polit, Humanalysis, Inc., 75 Clinton Street, Saratoga Springs, NY 12866 and Griffith University School of Nursing, Gold Coast, Australia.
E-mail: dpolit@rocketmail.com
*President and Adjunct Professor.
**Professor.
Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/nur.20147

domain of content addressed by the instrument'' (Waltz, Strickland, & Lenz, 2005, p. 155); and

(3) ''. . .the extent to which an instrument adequately samples the research domain of interest when attempting to measure phenomena'' (Wynd, Schmidt, & Schaefer, 2003, p. 509).

There is general agreement in these definitions that content validity concerns the degree to which a sample of items, taken together, constitute an adequate operational definition of a construct.

There is also agreement in the methodologic literature that content validity is largely a matter of judgment, involving two distinct phases: a priori efforts by the scale developer to enhance content validity through careful conceptualization and domain analysis prior to item generation, and a posteriori efforts to evaluate the relevance of the scale's content through expert assessment (e.g., Beck & Gable, 2001; Lynn, 1986; Mastaglia, Toye, & Kristjanson, 2003). This article focuses on the second part of this process.

BACKGROUND ON CONTENT VALIDITY APPROACHES

Numerous methods of quantifying experts' degree of agreement regarding the content relevance of an instrument have been proposed. These include, for example, averaging experts' ratings of item relevance and using a pre-established criterion of acceptability (e.g., Beck & Gable, 2001); using coefficient alpha to quantify agreement of item relevance by three or more experts (Waltz et al., 2005, p. 157); and computing a multirater kappa coefficient (Wynd et al., 2003). A variety of other indexes that capture interrater agreement have been proposed and are used mainly in the field of personnel psychology (Lindell & Brandt, 1999).

One approach, recommended several decades ago, has special relevance in this article. This approach involves having a team of experts indicate whether each item on a scale is congruent with (or relevant to) the construct, computing the percentage of items deemed to be relevant for each expert, and then taking an average of the percentages across experts. As an example with two experts, if Expert 1 rated 100% of a set of items to be congruent with the construct, and Expert 2 rated 80% of the items to be congruent, the value of this index would be 90%. This has been referred to as the *average congruency percentage* (ACP) and is attributed to Popham (1978). Waltz et al. (2005, p. 178) advise that an ACP of 90 percent or higher would be considered acceptable.

Among nurse researchers, the most widely reported measure of content validity is the content validity index, or CVI. The CVI (which we define and describe at length later in this article) has been used for many years, and is most often attributed to Martuza (1977), an education specialist. However, researchers who use the CVI to assess the content validity of their scales—regardless of their own disciplinary backgrounds—often cite methodologic work in the nursing literature, most often Davis (1992), Grant and Davis (1997), Lynn (1986), Waltz et al. (2005), or Waltz and Bausell (1981). Lynn's seminal study has been especially influential.

The CVI has had its share of critics, however, even among nurse researchers. For example, Wynd and her colleagues (2003) used both the CVI and a multirater kappa coefficient in their content validation of the Osteoporosis Risk Assessment Tool. They argued that the kappa statistic was an important supplement to (if not substitute for) the CVI because the formula for kappa yields an index of degree of agreement beyond chance agreement, unlike the CVI, which does not adjust for chance agreement. Other concerns are that the CVI throws away information by collapsing experts' multipoint ordinal ratings into two categories (i.e., into relevant/not relevant categories, a common practice), and that the CVI focuses on item relevance of the items reviewed but does not capture whether a scale includes a comprehensive set of items to adequately measure the construct of interest.

Our purpose in this article is not to advocate for or against using the CVI as the standard index of content validity. Rather, because the CVI is used so widely in nursing, our purpose is to clarify what this index is actually capturing and to demonstrate that researchers are not always clear in articulating how they have computed it.

THE CONTENT VALIDITY INDEX FOR ITEMS (I-CVI)

As noted by Lynn (1986), researchers compute two types of CVIs. The first type involves the content validity of individual items and the second involves the content validity of the overall scale.

There is considerable agreement about how to compute the item-level CVI, which we refer to for the purpose of clarity as the I-CVI. A panel of

content experts is asked to rate each scale item in terms of its relevance to the underlying construct. Lynn (1986) advised a minimum of three experts, but indicated that more than 10 was probably unnecessary. By tradition, and based on the advice of early writers such as Lynn, as well as Waltz and Bausell (1981), these item ratings are typically on a 4-point ordinal scale. Lynn acknowledged that 3- or 5-point rating scales might be considered, but she advocated using a 4-point scale to avoid having a neutral and ambivalent midpoint. Several different labels for the four points along the item-rating continuum have appeared in the literature, but the one that was advocated by Davis (1992) appears to be in frequent use: 1 = *not relevant*, 2 = *somewhat relevant*, 3 = *quite relevant*, 4 = *highly relevant*. Then, for each item, the I-CVI is computed as the number of experts giving a rating of either 3 or 4 (thus dichotomizing the ordinal scale into *relevant* and *not relevant*), divided by the total number of experts. For example, an item that was rated as *quite* or *highly* relevant by four out of five judges would have an I-CVI of .80.

One concern that has been raised about the CVI is that it is an index of interrater agreement that simply expresses the proportion of agreement, and agreement can be inflated by chance factors. For example, if two judges rated the relevance versus irrelevance of an item, by chance alone the two judges would be expected to agree on relevance 25 percent of the time. In recognition of this problem, Lynn (1986) developed criteria for item acceptability that incorporated the standard error of the proportion. She recommended that with a panel of "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" (p. 383). In other words, the I-CVI should be 1.00 when there are five or fewer judges. When there are six or more judges, the standard can be relaxed, but Lynn recommended I-CVIs no lower than .78. For example, with six raters, there could be one "not relevant" rating (I-CVI = .83) and with nine raters there could be two *not relevant* ratings (I-CVI = .78).

Researchers use I-CVI information to guide them in revising, deleting, or substituting items. In research reports, however, researchers do not usually provide information about I-CVI values. I-CVIs tend only to be reported in methodological studies that focus on descriptions of the content validation process. What is most often reported in scale development studies is the CVI for the entire scale, and that is where the problems lie.

THE CONTENT VALIDITY INDEX FOR SCALES (S-CVI)

Computational procedures for the scale-level CVI, which we refer to for the sake of clarity as the S-CVI, have been fully explicated in terms of ratings by two experts. Here are two frequently cited definitions: The S-CVI is defined as "the proportion of items given a rating of quite/very relevant by both raters involved" (Waltz et al., 2005, p. 155) and "the proportion of items given a rating of 3 or 4 by both raters involved" (Waltz & Bausell, 1981, p. 71). Both references present tables to illustrate how to compute the S-CVI with two raters using 4-point scales of item relevance. An example similar to that shown in Waltz et al. (p. 155) is presented in Table 1. In this example, 8 out of 10 items were judged to be *quite* or *highly* relevant (i.e., a rating of 3 or 4) by both experts, and so the S-CVI is computed to be .80. Many writers have indicated that an S-CVI of .80 or higher is acceptable (e.g., Davis, 1992; Grant & Davis, 1997; Polit & Beck, 2004).

The key word in the definition of the two-rater S-CVI is *both*. According to the definition, both judges have to agree that any individual item is relevant in order for it to *count* toward the S-CVI.

Now consider the case when there are more than two judges, which is by far the more usual situation—and, indeed, having more than two experts was explicitly recommended by Lynn (1986). Here is how the CVI for scales has been defined for two or more raters: The CVI for the entire scale is (1) "the proportion of total items judged content valid" (Lynn, p. 384); (2) "the proportion of items on an instrument that achieved a rating of 3 or 4 by the content experts" (Beck & Gable, 2001, p. 209); (3) and "the proportion of

Table 1. Computation of an S-CVI for a 10-Item Scale With Two Expert Raters*

	Expert Rater No. 1		
	Items Rated 1 or 2[a]	Items Rated 3 or 4[b]	Total
Expert rater no. 2			
Items rated 1 or 2[a]	2	0	2
Items rated 3 or 4[b]	0	8	8
Total	2	8	10
S-CVI = 8/10 = .80			

S-CVI, content validity index for the scale.
*After Waltz et al. (2005), p. 155.
[a]Ratings of 1 = *not relevant*; 2 = *somewhat relevant*.
[b]Ratings of 3 = *quite relevant*; 4 = *highly relevant*.

experts who score items as relevant or representative with either 3 or 4'' (Grant & Davis, 1997, p. 273). These definitions are more ambiguous than the definition for two raters because there is no analog for the ''both'' specification, which for three or more raters would be ''all.'' An extension of the definition for a two-person S-CVI for multiple raters would be: the proportion of items on an instrument that achieved a rating of 3 or 4 by *all* the content experts. For convenience, we refer to this definition of the CVI for scales as S-CVI/UA (universal agreement).

To illustrate, Table 2 shows the relevance ratings of six experts for a 10-item scale. In this example, all six experts rated 9 out of the 10 items as relevant. However, the item judged not relevant differed for the six experts. Following the definition requiring universally congruent ratings by the experts, the S-CVI/UA in this example would be .40. Only 4 out of the 10 items (items 7–10) received relevance ratings of 3 or 4 by *all* the experts. It is easy to see that when this definition of the S-CVI is used, the more experts are included, the greater the likelihood that the S-CVI will be low: As the number of experts increases, the likelihood of achieving total agreement decreases. For example, if a seventh expert were added who rated only item 7 as not relevant, the S-CVI as thus defined would be .30—despite the fact that all I-CVIs are in an acceptable range. While critics of the CVI have worried that the S-CVI and I-CVI are inflated because of the possibility of chance agreement, there is a corresponding possibility that *disagreement* will be inflated due to chance factors as well. For example, in a two-rater situation, the probability of chance disagreement on a dichotomous rating of relevance is .500; this is analogous to the odds of getting one head and one tail (i.e., disagreement) in a 2-coin toss. In a 6-rater situation, the probability of at least one chance disagreement on dichotomous relevance ratings is .968. This is analogous to the probability of getting at least one head or one tail (disagreement) in a 6-coin toss. The probability that all raters would agree on relevance, and on irrelevance, is $.5^N$, where $N =$ the number of raters.

The three definitions that we cited for the general case of the S-CVI did *not* use the word ''all.'' In fact, there are other ways to interpret the definitions, and that is by inferring that what is meant is the *average* proportion of items rated as 3 or 4 across the various judges. The proportion of items rated as relevant by each of the six experts in Table 2 is .90, and so the average also would be .90. This is clearly a more liberal interpretation of the definition for the S-CVI. We refer to this approach as S-CVI/Ave. Figure 1 summarizes our terms, acronyms, and definitions relating to content validity.

There are three ways to calculate the S-CVI/Ave, which we illustrate with the information in Table 2. The first, as just explained, averages the proportion of items rated relevant across experts. Thus, we can calculate S-CVI/Ave as $(.90+.90+.90+.90+.90+.90)/6 = .90$. Another way is to average the I-CVIs by summing them and dividing by the number of items: $(.83 + .83 + .83 + .83 + .83 + .83 + 1.00 + 1.00 + 1.00 + 1.00)/10 = .90$. A third way is to count the total number of Xs in the table—the number of

Table 2. Fictitious Ratings on a 10-Item Scale by Six Experts: Items Rated 3 or 4 on a 4-Point Relevance Scale

Item	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Number in Agreement	Item CVI
1	—	X	X	X	X	X	5	.83
2	X	—	X	X	X	X	5	.83
3	X	X	—	X	X	X	5	.83
4	X	X	X	—	X	X	5	.83
5	X	X	X	X	—	X	5	.83
6	X	X	X	X	X	—	5	.83
7	X	X	X	X	X	X	6	1.00
8	X	X	X	X	X	X	6	1.00
9	X	X	X	X	X	X	6	1.00
10	X	X	X	X	X	X	6	1.00
Proportion Relevant:	.90	.90	.90	.90	.90	.90	Mean I-CVI = .90 S-CVI/UA = .40 Mean expert proportion = .90	

I-CVI, item-level content validity index.
S-CVI/UA, scale-level content validity index, universal agreement calculation method.

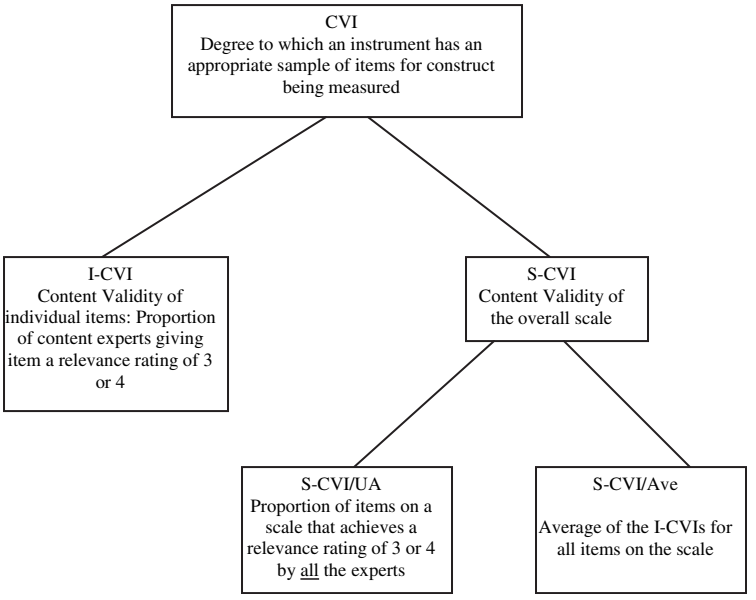

FIGURE 1. Definitions of content validity terms. I-CVI, item-level content validity index; S-CVI, scale-level content validity index; S-CVI/UA, scale-level content validity index, universal agreement calculation method; S-CVI/Ave, scale-level content validity index, averaging calculation method.

items rated relevant by all experts combined, which in this case is 54—and to then divide by the total number of ratings: 54/60 = .90. All three computations will *always* yield the same results. We think, however, that it is best to conceptualize the S-CVI/Ave as the average I-CVI value because this puts the focus on average item quality rather than on average performance by the experts.

One other thing is important to note. The S-CVI/Ave is identical to the index mentioned earlier as the average congruency percentage (ACP). The guideline offered by Waltz et al. (2005, p. 178) is that the ACP should be .90—not .80 as is the standard criterion for acceptability for the S-CVI. It seems reasonable to demand a higher standard for the ACP (or S-CVI/Ave) than for an S-CVI/UA because the former is much more liberal in its definition of congruence. With the fictitious data in Table 2, the S-CVI/UA would be .40 (not even remotely acceptable according to traditional standards), while the S-CVI/Ave (i.e., the ACP) would be .90.

THE CVI IN THE SCALE DEVELOPMENT LITERATURE

An important question is, how are nurse researchers computing the S-CVI for the content validity of a new scale when they use more than two experts? Are they using the conservative requirement of 100% agreement at the item level for at least 80% of the items? Or, are they averaging I-CVIs (or averaging proportions rated relevant across experts, which yields the same results) and using .80 as their standard of acceptability? With one exception discussed below, we could find no explanation in instrument development studies. When information about method of computing the CVI is absent, readers of such studies do not necessarily have a good understanding of the content validity of a new scale. Although Table 2 was admittedly an exaggerated example so that we could highlight possible disparities in computational methods, it is clear that the two approaches can lead to different conclusions.

In searching the literature for information about content validity, we found only one study, which is in a social work journal, that fully specified how the researchers calculated their S-CVI. This was also the only study we found that acknowledged the fact that there are two methods of computing this index. Rubio, Berg-Weger, Tebb, Lee, and Rauch (2003) illustrated the content validation process they used in developing the Caregiver Well-Being Scale. They calculated their S-CVI based on ratings of relevance by six judges, using

the averaging approach. They specifically adopted this approach because of their concern that with so many raters, the content validity would be depressed if they used the S-CVI/UA approach that demanded 100% agreement.

Although we found no reports on scale development in the nursing literature that described fully how the researchers computed their S-CVIs, it is sometimes possible to make inferences by working backward from information on number of items, number of experts, and the CVI value. We conducted an analysis to assess the extent to which inferences about S-CVI calculations were possible, and to determine if both calculation approaches are being used. Unfortunately, there is no well-defined ''population'' of content validation efforts from which to select a random sample, and so we sampled purposively, selecting 10 scale development studies from the recent nursing literature. In as much as the purpose of the analysis was to determine whether there is evidence that both S-CVI computational approaches are being used, rather than to calculate the percentage of studies using one approach or the other, such a purposive sample seems justifiable. By purposive we mean that we deliberately selected 10 psychometric studies from seven different nursing journals, written by nurse researchers from different countries (e.g., the United States, Canada, and China), regarding the development of scales relating to various nursing specialty areas (e.g., maternal and child health, pediatrics, palliative care, education, administration).

The results of our analysis of the 10 scale development studies are shown in Table 3. To make inferences about scale developers' S-CVI computational methods, we first tested whether the published S-CVI value, when multiplied by the number of items on the scale, was close to a whole number. For example, for the first entry (Champion, Skinner, & Menon, 2005), when the S-CVI of .80 is multiplied by 10 items, the result is 8. This means that the S-CVI/UA was plausible: the value of .80 could have been achieved if all five judges universally rated 8 of the 10 items as relevant. Next, we assessed the plausibility of the S-CVI/Ave by first multiplying the number of experts by the number of items. This yields the total number of possible item-by-expert ratings. Recall that in Table 2, there were a total of 60 possible ratings (6 experts times 10 items); 54 of them indicated relevance, and so the S-CVI/Ave was $54/60 = .90$. Then in our analysis for Table 3, the total number of item-by-expert ratings was multiplied by the S-CVI to see if this value was close to a whole number. In the case of the first entry (Champion et al.), the value of 50 ratings (5 experts times 10 items) multiplied by .80 is 40, indicating that 40 out of 50 ratings were judgments of relevance. Thus, the S/CVI for this first study could also have used the averaging approach.

The final column in Table 3 indicates what inferences could be made using this strategy. As the table indicates, for 7 out of the 10 studies our backward calculation inferences were inconclusive—that is, either S-CVI/UA or S-CVI/Ave plausibly could have been used. However, for two of these seven studies, there was sufficient supplementary information in the article to conclude that the S-CVI/UA approach had been

Table 3. S-CVI Calculations in Selected Scale Development Studies in Nursing Journals

Reference	No. of Experts	No. of Items	S-CVI Value	Inferred S-CVI Calculation Method
Champion et al. (2005)	5	10	.80	Could be either
Chen et al. (2003)	10	12	.92	Could be either
Chien & Norman (2004)	15	25	.96	Could be either, but S-CVI/ UA probable[a]
Dobratz (2004)	3	77	.83	Could be either
Fowles & Feucht (2004)	2	18	.72	Could be either
Li and Lopez (2004)	10	20	.98	S-CVI/Ave
Lindgren (2005)	3	28	.83	S-CVI/Ave
McGilton (2003)	5	6	.83	Could be either
Sauls (2004)	6	43	.81	Could be either, but S-CVI/ UA probable[a]
Smith et al. (2004)	7	33	.86	S-CVI/Ave

S-CVI, scale-level content validity index.
S-CVI/UA, scale-level content validity index, universal agreement calculation method.
S-CVI/Ave, scale-level content validity index, averaging calculation method.
Note: [a]Although the backward calculations support an inference that *either* S-CVI calculation method could have been used, information in these articles suggest that the S-CVI/UA approach was used.

adopted. For example, Sauls (2004) specifically mentioned that the denominator of her calculations was items: ''The CVI was .81, meaning that 81% of the total items were judged content valid'' (p. 126).

We were able to rule out the S-CVI/UA approach for 3 of the 10 studies, and thus inferred that the S-CVI/Ave method had been used. To illustrate, consider the study by Li & Lopez (2004). For their 20-item scale, the S-CVI was .98. If there had been universal agreement for 19 of the items, the S-CVI would have been .95 (19/20 = .95), not .98. This suggests that the S-CVI/UA method was not used to determine the scale's content validity. When .98 is multiplied by the total number of expert-by-item ratings (200), the result is the whole number 196—that is, all but 4 of the 200 ratings indicated relevance. Consequently, in this study, we concluded that the S-CVI/Ave method had been used. In this situation, an S-CVI/Ave of .98 could have been achieved if 16 items had I-CVIs of 1.0, and the remaining 4 items had I-CVIs of .90.

In conclusion, our analysis suggests that nurse researchers are using both computational approaches in the calculation of the S-CVI, and that inferences about the method used are not always possible. Even if such inferences *were* possible, we think that readers of psychometric reports should not have to do their own calculations to understand what a reported S-CVI means.

It might be noted that, in addition to the two methods of calculating the S-CVI already discussed, a third approach is possible. The S-CVI could be calculated as the proportion of items on the scale that were judged to be content valid at the item level. As noted previously, Lynn (1986) advocated I-CVIs of 1.0 when there are five or fewer experts, and I-CVIs in the vicinity of .80 when there are six or more experts. For the data in Table 2, *all* of the items meet Lynn's I-CVI criterion, and so this third approach would mean that the S-CVI for the 10-item scale would be 1.0. If this definition were used, *all* S-CVIs should, in theory, be 1.0, because items with lower-than-acceptable levels of content validity should be revised or discarded. Few scale development studies report an S-CVI of 1.0, as suggested by the data in Table 3 and so we conclude that this third approach probably has not been adopted.

RECOMMENDATIONS FOR THE CVI

Our investigation suggests that greater clarity about content validation in scale development studies is needed. We offer several recommendations that we think will improve communication about content validity for researchers using the CVI to measure agreement about item relevance.

First, as this article suggests, it is important to distinguish between content validity at the item level and at the scale level. The acronym CVI has been used for both, following Lynn' (1986) influential study. Sauls (2004), for example, used the same acronym for both item-level computations (''. . .8 items had a CVI of .67,'' p. 126) and scale-level calculations (in her revised scale, ''the CVI of the total instrument was .95'', p. 126). We think that the acronyms introduced in this article to distinguish the two (I-CVI and S-CVI) would be useful, but even if the acronyms are not adopted the distinction should still be made clear (for example, referring to item-level CVIs and scale-level-CVIs).

Second, we recommend that researchers report the range of their I-CVI values for items retained on the scale, in addition to the value of the S-CVI. Rempusheski & O'Hara, 2005, for example, provided such range information about items on their scale, noting that ''CVI ranged from .60 to 1.0'' (p. 421). Providing range information for items is especially important when the S-CVI/UA method has been used because this calculation method ignores I-CVI values for which there was not universal agreement. To give an exaggerated and unlikely example, the Champion et al. (2005) study, listed in Table 3, had an S-CVI value of .80. If the universal agreement method was used, this means that eight items on their 10-item scale had I-CVIs of 1.0—but the I-CVIs on the remaining two items could have been 0.0, .20, .40, .60, or .80—we do not know anything about the two I-CVIs, except that they could not be 1.0.

We also urge scale developers who compute the CVI in their content validation efforts to be clear about how they calculated the S-CVI. As we have shown, the two approaches can yield dramatically different results. Scale users need to have accurate information about the quality of the scales they are considering.

We prefer the S-CVI/Ave method for scale-level CVIs, although there may well be valid reasons to prefer the S-CVI/UA method. Our reasoning is that the universal agreement is overly stringent when there are many experts on the validation panel. It seems excessively conservative to demand 100 percent agreement—what if, for example, one expert did not understand the task or had a biased viewpoint? The example in Table 2 illustrates our rationale: even though 90% of the

overall ratings for the 10-item scale were judged to be relevant, and all of the I-CVI's were higher than .80, the value of the S-CVI/UA was only .40. Perhaps the most informative procedure is to compute the S-CVI both ways, and to report both values.

One final issue concerns an acceptable standard for the S-CVI. Davis (1992) and others have recommended a minimum S-CVI of .80. This may be a reasonable (and even strict) criterion for the S-CVI/UA, but it might be argued that researchers using the more liberal SCVI/ Ave approach should follow Waltz et al.'s (2005) advice of using .90 as the standard for this index of average congruity. If items with unacceptable I-CVIs are revised and re-evaluated, this should not be a difficult standard to meet.

In summary, we recommend that for a scale to be judged as having excellent content validity, it would be composed of items with I-CVIs that meet Lynn's (1986) criteria (I-CVI = 1.00 with 3 to 5 experts and a minimum I-CVI of .78 for 6 to 10 experts) and it would have an SCVI/ Ave of .90 or higher. This requires strong conceptualizations of constructs, good items, judiciously selected experts (Davis, 1992), and clear instructions to the experts regarding the underlying constructs and the rating task (Lynn). The recommended standards may necessitate two rounds of expert review if the initial I-CVIs suggest the need for substantial item improvements, or if the reviewers identify aspects of the construct not adequately covered by the initial pool of items (Lynn). Whichever computation method is used, we urge scale developers to be explicit about how their CVI values were calculated so that potential users of the scale can draw informed conclusions about the scale's content validity, as a supplement to other empirical information about the scale's quality.

REFERENCES

Beck, C.T., & Gable, R.K. (2001). Ensuring content validity: An illustration of the process. Journal of Nursing Measurement, 9, 201–215.

Champion, V., Skinner, C.C., & Menon, U. (2005). Development of a self-efficacy scale for mammography. Research in Nursing & Health, 28, 329–336.

Chen, H.S., Horner, S.D., & Percy, M.S. (2003). Cross-cultural validation of the Stages of the Tobacco Acquisition Questionnaire and the Decisional Balance Scale. Research in Nursing & Health, 26, 233–243.

Chien, W.T., & Norman, I. (2004). The validity and reliability of a Chinese version of the Family Burden Interview Schedule. Nursing Research, 53, 314–322.

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

Dobratz, M.C. (2004). The Life Closure Scale: Additional psychometric testing of a tool to measure psychological adaptation in death and dying. Research in Nursing & Health, 27, 52–62.

Fowles, E.R., & Feucht, J. (2004). Testing the Barriers to Health Eating Scale. Western Journal of Nursing Research, 26, 429–443.

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

Li, H.C.W., & Lopez, V. (2004). Psychometric evaluation of the Chinese version of the State Anxiety Scale for Children. Research in Nursing & Health, 27, 198–207.

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

Lindgren, K. (2005). Testing the Health Practices in Pregnancy Questionnaire-II. Journal of Obstetric, Gynecologic, & Neonatal Nursing, 34, 465–472.

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

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

Mastaglia, B., Toye, C., & Kristjanson, L.J. (2003). Ensuring content validity in instrument development: Challenges and innovative approaches. Contemporary Nurse, 14, 281–291.

McGilton, K.S. (2003). Development and psychometric evaluation of supportive leadership scales. Canadian Journal of Nursing Research, 35, 72–86.

Polit, D.F., & Beck, C.T. (2004). Nursing research: Principles and methods (7th ed.) Philadelphia: Lippincott, Williams, & Wilkins.

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

Rempusheski, V.F., & O'Hara, C.T. (2005). Psychometric properties of the Grandparent Perceptions of Family Scale. Nursing Research, 54, 419–427.

Rubio, D.M., Berg-Weger, M., Tebb, S.S., Lee, E.S., & Rauch, S. (2003), Objectifying content validity: Conducting a content validity study in social work. Social Work Research, 27, 94–104.

Sauls, D.J. (2004). The Labor Support Questionnaire: Development and psychometric analysis. Journal of Nursing Measurement, 12, 123–312.

Smith, A.J., Thurkettle, M.A., & dela Cruz, F.A. (2004). Use of intuition by nursing students: Instrument development and testing. Journal of Advanced Nursing, 47, 614–622.

Waltz, C.F., & Bausell, R.B. (1981). Nursing research: Design, statistics, and computer analysis. Philadelphia: F. A. Davis.

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

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