Objectifying content validity: Conducting a content validity study in social work research

Doris McGartland Rubio, Marla Berg-Weger, Susan S. Tebb, E. Suzanne Lee, and Shannon Rauch

Social scientists frequently study complex constructs. Despite the plethora of measures for these constructs, researchers may need to create their own measure for a particular study. When a measure is created, psychometric testing is required, and the first step is to study the content validity of the measure. The purpose of this article is to demonstrate how to conduct a content validity study, including how to elicit the most from a panel of experts by collecting specific data. Instructions on how to calculate a content validity index, factorial validity index, and an interrater reliability index and guide for interpreting these indices are included. Implications regarding the value of conducting a content validity study for practitioners and researchers are discussed.

Key words: **constructs; content validity; measure; psychometric testing**

***Doris McGartland Rubio, PhD,** is assistant professor of medicine and director, Data Center, Center for Research on Health Care, University of Pittsburgh, 230 McKee Place, Suite 600, Pittsburgh, PA 15213; e-mail: rubiodm@msx.upmc.edu. **Marla Berg-Weger, PhD, LCSW,** is associate professor, and **Susan S. Tebb, PhD,** is associate professor and dean, School of Social Service, Saint Louis University. **E. Suzanne Lee, PhD,** is assistant professor, Saint Xavier University, Chicago, IL. **Shannon Rauch, MS,** is a research assistant, Department of Psychology, Saint Louis University.*

Researchers in the social sciences study complex constructs for which valid and reliable measures are needed. The measures should be brief, clear, and easy to administer. Measures that are too long or difficult to read may result in a lowered response rate or inaccurate responses. In addition, the measure must be appropriate for use in the targeted population. For example, measures designed for use with heterogeneous populations may not be appropriate for a specific population with certain characteristics.

A plethora of measures exist with known psychometric properties, but researchers may need to develop a new measure for a particular construct because no measure exists that operationalizes the construct as the researcher conceptualized it. In these circumstances, a content validity study should be conducted.

VALIDITY

Traditionally, three types of validity may be demonstrated: content, criterion, and construct validity.

Content Validity

Content validity refers to the extent to which the items on a measure assess the same content or how well the content material was sampled in the measure. Content validity can be characterized as face validity or logical validity. Face validity indicates that the measure appears to be valid, "on its face." Logical validity indicates a more rigorous process, such as using a panel of experts to evaluate the content validity of a measure.

Nunnally and Bernstein (1994) did not distinguish among different types of content validity; but presented alternative ways of assessing content validity. They suggested evaluating content

CCC Code:1070-5309/03 $3.00

validity by demonstrating internal consistency through correlating the scores from the measure with another measure of the same construct and by showing change in posttest scores over pretest scores.

Criterion Validity

Criterion validity is demonstrated by finding a statistically significant relationship between a measure and a criterion (Nunnally & Bernstein, 1994). Criterion validity is considered the "gold standard," and usually a correlation is used to assess the statistical relationship. For example, the Graduate Record Examination (GRE) has been found to predict graduate school success (as measured by the first-year grade-point average) for certain disciplines (Rubio, Rubin, & Brennan, 2003). Three types of criterion validity are postdictive, concurrent, and predictive. If the criterion has occurred, the validity is postdictive. The validity is concurrent if the criterion exists at the same time as the construct measured. The GRE example demonstrates predictive validity, because graduate school success (criterion) occurs after taking the GRE (measure). According to Nunnally and Bernstein, a correlation of .30 indicates adequate criterion validity.

Construct Validity

Anastasi and Urbina (1997) described construct validity as "the extent to which the test may be said to measure a theoretical construct or trait" (p. 126). Three kinds of construct validity are factorial, known groups; and convergent and discriminant (or divergent) validity. Factorial validity can be assessed by conducting an exploratory factor analysis such as principal components or a confirmatory analysis using structural equation modeling. Known groups validity is determined by finding statistically significant differences in scores between a group with a known property of a measure and a group that does not have a characteristic. Campbell and Fiske (1959) introduced convergent and discriminant validity. Using multitrait–multimethod matrix, they proposed that the researcher measures different constructs with different methods (such as self-report and observation). The degree of validity can be established by assessing four correlations. To the degree that convergent validity is present, the construct that is measured with different methods should have the highest correlation. At the other end of the continuum, a low correlation between two different constructs that are measured with two different methods demonstrates discriminant validity.

The distinctness of content, criterion, and construct validity dissolves when assessing the different types of validity, and their interconnectedness becomes apparent. As was discussed in the section on content validity, Nunnally and Bernstein (1994) suggested correlating two measures of the same construct or documenting evidence of the content validity by a change in posttest scores. It could be argued that these two methods better approximate construct validity. (Readers are referred to Messick's 1989 influential work on validity. For a discussion on the trinity of validity, see Shepard [1994] and Hubley & Zumbo [1996].)

CONTENT VALIDITY STUDIES

Researchers can receive invaluable information by conducting a content validity study. Using a panel of experts provides constructive feedback about the quality of the newly developed measure and objective criteria with which to evaluate each item.

Without conducting a content validity study, researchers would need to spend resources disseminating an untested measure to a subject pool to obtain analyzable data. These data may indicate needed revisions in the measure. The researcher would then need to conduct another pilot study to evaluate the revised measure. Thus, researchers would be spending numerous resources on evaluating and redeveloping the measure. For a content validity study, the researcher would spend more resources initially but fewer resources in numerous revisions of the measure through evaluations. All measures need to be evaluated repeatedly. However, the measures that have established content validity would need fewer revisions in the evaluation phase.

A content validity study can provide information on the representativeness and clarity of each item and a preliminary analysis of the factorial validity. In addition, the expert panel offers concrete suggestions for improving the measure. The revised measure can then be used in a pilot study to assess other psychometric properties.

Some limitations of content validity studies should be noted. Experts' feedback is subjective; thus, the study is subjected to bias that may exist

among the experts. In addition, this type of study does not eliminate the need for additional psychometric testing, which is critical for the development of a measure. Another potential limitation is that this type of study does not necessarily identify content that might have been omitted from the measure. However, experts are asked to suggest other items for the measure, which may help minimize this limitation.

CONDUCTING A CONTENT VALIDITY STUDY

Select a Panel of Experts

The panel of experts consists of content experts and lay experts. The content experts are professionals who have published or worked in the field. Selecting an expert in measurement or a related field can also be helpful in determining whether the measure is well-constructed and suitable for psychometric testing (Davis, 1992). Criteria for selecting these experts are the number of publications or the work experience.

The lay experts are people for whom the topic is most salient. Using potential research subjects as experts ensures that the population for whom the measure is being developed is represented. The lay group addresses issues such as phrasing and unclear terms and recommends other important or salient items.

The literature is diverse with respect to the number of content experts needed. Lynn (1986) recommended a minimum of three. However, others have suggested a range of two to 20 experts (Gable & Wolf, 1993; Walz, Strickland, & Lenz, 1991). As noted in Grant and Davis (1997), the number of panel experts depends on the desired level of expertise and diversity of knowledge. We recommend using at least three experts for each group (professionals and lay experts) with a range up to 10. This yields a sample size of six to 20. Using a larger number of experts may generate more information about the measure.

Solicit Experts' Participation

After identifying potential panel members, a letter, e-mail, or telephone call soliciting their participation is recommended at least one week in advance to provide the subjects time to respond to your request. An incentive, such as a copy of the revised scale, is recommended to increase the response rate.

Mail Cover Letter and Response Forms

The packets distributed to the experts should include a cover letter, response form, and self-addressed, stamped return envelope. A brief demographic questionnaire could be included to provide demographic information about the panel of experts in subsequent publications or reports.

Cover Letter. The cover letter should include the purpose of study, the reason the expert was selected, a description of the measure and its scoring, and an explanation of the response form. Explaining the purpose and use of the measure clarifies the need for the content validity study.

Finally, a description of the response form is needed. Instruction is critical regarding the response form itself, but the cover letter should explicate the conceptual purpose of the form. For example, this paragraph might state:

> The enclosed survey asks you to evaluate how representative the items are of the content domain of (name construct). That is, to what extent do you think that each question on the survey measures (name construct)? Because (name construct) comprises of several different factors, you are also asked to indicate which factor the item measures. The clarity of each item is another important aspect for you to evaluate. Specifically, indicate how clear you think each item is. Finally, you are asked to evaluate the overall comprehensiveness of the entire measure by either adding or deleting items.

The cover letter for the two groups can reflect the educational level of each group. For example, the lay experts might not be familiar with the expression "how representative the items are of the domain." The phrase can be rewritten as "Does the item seem to address the area of well-being?"

Description of Response Form. Four criteria are used to evaluate the measure: (1) representativeness of the content domain; (2) clarity of the item; (3) factor structure; and (4) comprehensiveness of the measure. Each item is rated on a scale from 1 to 4 for representativeness and clarity. Representativeness is demonstrated by an item's ability to represent the content domain as described in the theoretical definition. The clarity of an item is evaluated on the basis of how clearly an item is worded. Some authors have suggested asking about the clarity of all items in one question at the end of the survey. On the basis of our experience with the content validity study (we found

that this approach was confusing for the experts). We recommend evaluating clarity with each item on the same scale as the representativeness, which allows experts to evaluate each item completely, rather than having to recall each item at the end of the survey. This process yields more useful information for revising the measure.

Anchors are provided for the scale points. A value of one indicates that the item is not representative of the domain or clear; a value of four indicates that the item is representative or clear. Space is provided for the experts to suggest ways to improve the item.

Several factors are listed for the construct, if factors are present in the measure. Experts are asked to assign each item to a factor. The expert can also identify a factor that is not specified. If the measure consists of only one factor, this step would be eliminated. An advantage of having the experts identify to which factor the item corresponds is that a preliminary assessment of the factorial structure can be made. Another option would be to keep the items grouped according to the factor and ask the experts to indicate how well the item measures that factor. This does not allow for any indication that an item may load onto two factors, nor does it assess the congruence between the proposed factor structure and the items.

Finally, the experts address the comprehensiveness of the measure. After evaluating the representativeness, clarity, and factor structure, experts are asked to consider the entire measure and specify the addition or deletion of any item.

A format was modeled after Grant and Davis's (1997) form (see Figure 1).

Analyze the Data

Three types of analyses can be performed.

Reliability or Interrater Agreement. First, interrater agreement (IRA) is assessed to determine the extent to which the experts are reliable in their ratings. Interrater agreement should be calculated for representativeness and clarity. The four-point scale is used to calculate the IRA for representativeness and clarity. The scale is dichotomized, with values one and two combined and values three and four combined. This method is consistent with the literature on conducting content validity studies (for example, Davis, 1992; Grant & Davis, 1997; Lynn, 1986). The data is dichotomized so that the researcher can assess the extent to which the experts agree that the item is representative of the item or not. The original four-point scale provides additional information for the researcher to determine the extent to which the item needs to be modified or deleted. The researcher counts the items that experts rated one or two and the items that are rated three or four. An IRA can be calculated for each item as well as for the scale. To determine the IRA for each item, the agreement among the experts is calculated. The IRA for the scale is computed as follows. The number of items considered 100 percent reliable is divided by the total number of items. For a less conservative approach, the IRA can be calculated by counting the number of items that have an IRA of at least .80 and dividing that number by the total number of items. The less conservative approach is recommended for studies that involve a many experts (that is, a sample of experts that exceeds five, as suggested by Lynn). As the number of experts increases, the chances of all of them agreeing decreases.

Content Validity Index. The content validity index (CVI) of a measure is calculated based on the representativeness of the measure. The CVI can be calculated by one of several methods. We recommend first computing the CVI for each item by counting the number of experts who rated the item as three or four and dividing that number by the total number of experts. This gives you the proportion of experts who deemed the item as content valid. The CVI for the measure is estimated by calculating the average CVI across the items. Davis (1992) recommends a CVI of .80 for new measures.

Some researchers suggest that the CVI for the measure be calculated by counting the number of items rated as a three or four by all the experts and dividing that number by the total number of items (Davis, 1992; Grant & Davis, 1997). A limitation of this method is that as the number of reviewers increases, the CVI is likely to decrease. It is more difficult to obtain agreement on the representativeness of an item with 10 reviewers than it is with two experts. To account for this, Lynn (1986) proposed setting a "standard error of the proportion" (p. 383) to help determine chance agreement versus actual agreement. Lynn provided a table, based on the number of experts, to determine whether an item is content valid. Instead of

FIGURE 1—Instructions for Rating Items in a Measure

INSTRUCTIONS – This measure is designed to evaluate the content validity of a measure. Please rate each item as follows:

- **Please rate the level of representativeness on a scale of 1 – 4, with 4 being the most representative. Space is provided for you to comment on the item or to suggest revisions.**
- **Please indicate the level of clarity for each item, also on a four-point scale. Again, please make comments in the space provided.**
- **Please indicate to which factor the item belongs. The factors are listed along with a definition of each. If you do not think the item belongs with any factor specified, please circle number 3 and write in a factor that may be more suitable.**
- **Finally, evaluate the comprehensiveness of the entire measure by indicating items that should be deleted or added. Thank you for your time.**

Theoretical definition	Representativeness	Clarity	Factors
Specify the construct being measured and provide a definition	1 = item is not respresentative 2 = item needs major revisions to be representative 3 = item needs minor revisions to be representative 4 = item is representative	1 = item is not clear 2 = item needs major revisions to be clear 3 = item needs minor revisions to be clear 4 = item is clear	List and number the factors and provide a definition of each 1 = factor 2 = factor 3 = other, specify
Items			
1. Item 1	1 2 3 4 Comments:	1 2 3 4 Comments:	1 2 3 Comments:
2. Item 2	1 2 3 4 Comments:	1 2 3 4 Comments:	1 2 3 Comments:

using the stringent criteria that all experts must agree, Lynn argued that even if one or more experts disagree, the item may still be content valid. The disagreement is accepted only if six or more experts are used.

Factorial Validity Index. A factorial validity index (FVI) has not been presented in the literature. We created the FVI to determine the degree to which the experts appropriately associated the items with their respective factors. This gives a preliminary indication of the factorial validity of the measure. To calculate the FVI for each item, the number of experts who correctly associated the item with the factor is divided by the total number of experts. Again, the average is taken across the items to compute the FVI for the measure. Because this is a new index, no criteria exist to determine the desired level to achieve. In our work we have found an FVI of at least .80, which is consistent with the recommended level of the CVI. To assess the full degree of FVI, other analyses (for example, a factor analysis) would need to be done.

Revise the Measure

After the data have been analyzed, the researcher may determine whether revisions are necessary to accommodate the experts' feedback. On occasion the researcher may want to contact a panel member for clarification. To do this, the researcher must have specified that the study is not anonymous. The panel may also be contacted

to examine a revised measure. If major revisions are needed for the measure, the researcher may want to repeat the process.

A SOCIAL WORK PRACTICE EXAMPLE

Method

Caregiver well-being and its measurement are established areas of interest for researchers and have been studied and discussed in the literature for some time (that is, George, 1994; George & Gwyther, 1986; Hooker, Monahan, Shifren, & Hutchinson, 1998; Maitland, Dixon, Hultsch, & Hertzog, 2001; Ory, Yee, Tennstedt, & Schulz, 2000). Efforts to measure caregiving experiences, in general, have been criticized for lack of rigor, limited focus of measurement (that is, caregiver stress only), and exclusion of the positive aspects of caregiving (Knight, Lutzky, & Macofsky-Urban, 1993; Kramer, 1997). Tebb (1995) developed the Caregiver Well-Being Scale to encompass positive and negative aspects of caregiving and to provide comparisons of caregivers and noncaregivers. After feedback from using the scale in practice and research settings, we determined that a shortened version of the scale could enhance the applicability and administration of the measurement and designed a shorter measure.

The Caregiver Well-Being Scale (Tebb, 1995) was developed using a health-strengths model (Weick, 1986) and incorporated the work of other researchers to capture basic needs and activities of living that are specific to measurement of the well-being of the family caregiver (Barusch, 1988; George & Gwyther, 1986; Maslow, 1962; McCubbin, 1982; Slivinske & Fitch, 1987). The original scale consists of 45 items with two subscales: (1) Basic Needs (BN; 23 items) and (2) Activities of Living (AOL; 22 items).

The scale has been evaluated and the psychometric properties tested. In the first evaluation, Tebb (1995) found that the scale reliably measured well-being (α = .91 for BN and .92 for AOL). In an exploratory factor analysis, Tebb found that several items loaded onto a factor more highly than other items. The measure was shown to have adequate validity, because it correlated with life satisfaction ($p < .001$). In subsequent work, structural equation modeling was used to evaluate the psychometric properties of the scale (Rubio, Berg-Weger, & Tebb, 1999). Confirmatory factor analysis showed that the measure consists of three factors for each subscale (BN and AOL). Nine items measured BN; each of the indicators had strong validity and reliability as demonstrated by the significant relationship to their respective factors (t values ranged from 1.96 to 8.86) and an R^2 of at least .60. The AOL subscale was measured by 10 items found to be highly reliable and valid. The high R^2 for each of the items (the majority have an R^2 of at least .60) indicates high reliability. The strong path coefficients (standardized values ranged from .49 to .91) demonstrate high validity. Research by Berg-Weger, Rubio, and Tebb (2000) found strong internal consistency of the measure (BN α = .91; AOL α = .81). Several methods of assessing the construct validity were conducted (factorial validity, convergent and discriminant validity, and known-groups validity). Across all three methods, the well-being measure performed as hypothesized; this was evident in the strong correlation between the two subscales and the negative correlation with depression. A weak correlation with strain was also noted. The factor structure for BN was similar to that of Rubio et al. (1999) in that three factors emerged. As also found in Tebb's work, not all of the items had strong factor loadings.

In each of these studies, some items were found to be better indicators than others. For this reason, we decided to use only the most valid and reliable indicators. Maslow's (1962) Hierarchy of Needs was consulted to provide a theoretical framework for the measure. According to Maslow, a pyramid of needs exists, with lower-level needs needing to be met before higher-level needs. The lowest needs, physiological needs, must be met before safety. The need after safety is love and belongingness, which is followed by self-esteem. The highest level is self-actualization.

Sample

Six professionals were identified who have expertise in family caregiving and well-being. Of the six professionals five were in academia, had doctoral degrees, and had engaged in research on family caregiving. One of the academic experts specializes in measurement; he was included to evaluate the psychometric potential of the measure. The one expert who was not in academia worked with family caregivers. She was asked to participate and to distribute 10 response forms to

family caregivers willing to serve as lay experts. These lay experts were caregivers of a family member with Alzheimer's disease or a related disorder. Eight response forms were returned; two of the response forms were not completed, leaving six lay experts for the analyses.

Materials and Procedure

The revised measure was conceptualized to consist of two dimensions, "Needs" and "Activities." This was revised from the original measure that defined the two dimensions as Basic Needs and Activities of Living. Each dimension has nine items that measure three factors. A definition of the dimension with sample items is presented.

> Needs items are meeting the biological, psychological, social needs to sustain life
>
> 1. Eating a well-balanced diet
> 2. Getting enough sleep
> 3. Access to health care

> Activities items are the implementation of the biological, psychological, social needs
>
> 1. Buying food
> 2. Attending to personal daily maintenance activities (meals, hygiene, laundry, and so forth)
> 3. Attending to medical needs

Using the theoretical definition provided, the experts rated each item to determine the item's ability to represent its respective dimension. The response form for this study varied slightly from the one described earlier. A clarity category was not included in the form for each item. When we were analyzing the data, we realized the need for a clarity category for each item. For this study, we asked at the end of the form for the experts to indicate which items were clear and which were not. Many experts did not complete this section. Of those who did, several referred to their comments on the items. Having a section for clarity would have eliminated the ambiguity our experts apparently experienced.

Each dimension (Needs and Activities) was identified a priori as having three factors. Specific items were developed for the three factors of Needs and the three factors of Activities. The experts were asked to identify the item with the appropriate factor. See Figure 2 for an example of the response form for Activities.

Because social work practitioners are the intended administrators of the Caregiver Well-Being Scale for family caregivers, the appropriateness of reading level for caregivers was incorporated in the clarity question. This question for the Activities dimension is presented.

> Clarity: Are the activities items well-written, and at an appropriate reading level for individuals who provide psychosocial and physical care to a family member?
>
> ____ Yes, the following items are clear (in the space below indicate the clear items)
>
> ____ No, some of the items are unclear (in the space below indicate the unclear items)
>
> Suggestions for making the items clear.

Last, the experts were asked to assess how well the items represent the entire conceptual domain and were asked to identify items they would recommend including or deleting. An example follows.

> The measure is designed to assess well-being. Two dimensions of well-being are evaluated, needs and activities. Please evaluate to what extent you think the entire instrument is comprehensive. In other words, are the items sufficient to represent the entire domain of well-being?
>
> What additional items would you recommend including?
>
> What items would you recommend deleting?

RESULTS

Of the 10 experts, eight returned the survey; two of the eight did not rate the items, but provided comments on how to revise the measure. The calculations of the IRA for Needs and Activities dimensions are provided for representativeness. The CVI is then presented for representativeness as well. The IRA and CVI cannot be calculated for clarity as previously advocated, because we did not devise the response form in that manner.

FIGURE 2—Sample Activities Items and Factors

Activities Items	Representativeness	Factors
Theoretical definition: Activities – are the implementation of the bio/psycho/social needs	1 = item is not respresentative of needs 2 = item needs major revisions to be representative of needs 3 = item needs minor revisions to be representative of needs 4 = item is representative of needs	1 = Self Care 2 = Connectedness 3 = Time for Self
1. Buying food	1 2 3 4 Comments:	1 2 3 Comments:
2. Attending to personal daily maintenance activities (meals, hygiene, laundry, etc.)	1 2 3 4 Comments:	1 2 3 Comments:
3. Attending to medical needs	1 2 3 4 Comments:	1 2 3 Comments:

Reliability or Interrater Agreement

Two methods exist for calculating the IRA. Traditionally, the researcher would count the number of items that have 100 percent agreement and divide by the total number of items. For example, of the nine items for the Needs dimension, four were considered to be 100 percent reliable. That is, for the four items, all experts rated the item consistently. This produced an IRA of .44. For the Activities dimension, when the experts rated the representativeness of each item, the experts rated five of the eight items the same, generating an IRA of .62. Although the IRA for the dimensions are not encouraging, if the researcher examines the IRA for each item, all of the items (with the exception of one) have an IRA of .80 or greater. One item has an IRA of .67. That is, four of six experts rated the item the same. Calculating the IRA by the second method of computing the average IRA yields an IRA of 89 percent for Needs dimension and 100 percent for the Activities dimension. We recommend using the latter, less conservative method to calculate IRA when the number of experts exceeds five. This represents the average agreement among the experts. The former method does not allow for any discrepancies among experts.

Content Validity Index

To estimate the CVI for each item, the number of experts who rated the item as either a three or four were counted and divided by the total number of experts. To calculate the CVI for the scale, the average was calculated across all items. See Figure 3 for an illustration of how to calculate the CVI.

FIGURE 3—Needs Items as Rated by Experts for Content Validity

	Experts						
Needs Items	1	2	3	4	5	6	CVI
1. Eating a well-balanced diet	4	4	4	4	3	4	6/6 = 100
2. Getting enough sleep	4	4	4	4	4	4	6/6 = 100
3. Access to health care	4	4	3	4	2	3	5/6 = 83

NOTE: CVI = content validity index.

For the Needs dimension, the CVI for the items range from .67 to 1.00. One item has a CVI of .67, two items have a CVI of .80, two have a CVI of .83, and four have a CVI of 1.00. The average of these is .88, which is the CVI for the scale, clearly above the .80 criteria. More important, every item with the exception of one has a CVI of .80 or greater. The item with the low CVI was subsequently revised.

Factorial Validity Index

The FVI is calculated by counting the number of experts who correctly assigned the item to the factor and dividing that number by the total number of experts (Figure 4). For the Needs dimension, all of the experts were able to assign six items to the correct factor. For two items, only three experts (of four) were able to correctly assign the items to the factor for a FVI of .75. One item has an FVI of 0, because no expert correctly assigned this item to the factor. This was the same item with the CVI of .67. The average FVI for the Needs dimension is .83.

The Activities dimension yielded similar results. All the experts correctly assigned five items to their respective factors (FVI = 1.00). Two items have an FVI of .75; three of four experts correctly connected the item with the respective factor. For one item, the FVI was .50, indicating that only half of the experts were able to discern the correct factor. The average FVI for the Activities scale is .84.

DISCUSSION

The findings from our study enabled us to clarify the measure. The low IRA demonstrates that at least one expert rated the item differently from the other experts. We analyzed all of the experts together and did not distinguish between lay and professional experts. The results might be different if we had. We recommend keeping the study confidential, as opposed to anonymous, so that the experts' responses can be analyzed separately.

The CVI and FVI are considered strong. Two items (one from each dimension) were not correctly assigned to the factor. These two items were revised on the basis of the experts' suggestions.

We made several revisions to improve the measure. The item from the Needs dimension that received a CVI of .67 was the same item that the experts were not able to assign to the correct factor. The item "feeling fulfilled" was revised to "feeling good about yourself." The revised item better represents the domain. This item measures self-security, which the experts were not previously able to discern with the former item.

One item from the Activities dimension also had a low FVI of .50. Although the item was considered to be content valid (CVI = 100 percent), we revised it from "have plans for the future" to "making plans for your financial future." This item measures the factor time for self.

The findings also indicated that some of the wording in the items needed revising. For example, the item, "optimal shelter" was rated reliably and had a CVI of .80. However, the reviewers consistently noted that the word 'optimal' was not well-suited for the measure. This item was revised to "having adequate shelter."

Overall, the content validity study provided direction for revision of the measure. The content validity of the measure is strong, as indicated by the panel of experts. The IRA was low using the conservative method, but this might be improved if the analyses were conducted separately for the professional and lay experts or if the less conservative method were used.

FIGURE 4—Needs Items as Rated by Experts for Factorial Validity

	Experts				
Needs Items	1	2	3	4	FVI
1. Eating a well-balanced diet	1	1	1	1	4/4 = 100
2. Getting enough sleep	1	0	1	1	3/4 = 75
3. Access to health care	0	0	0	0	0/4 = 0

NOTE: FVI = factorial validity index.

CONCLUSION

This article demonstrates how to conduct a content validity study, a crucial step in scale development. An expert panel was used to evaluate a new measure. The experts critiqued the measure to determine the representativeness and clarity of the items, the factor with which the item is associated, and the extent to which the measure is comprehensive.

Although content validity is subjective, using this method can add objectivity. Using a panel of experts provides the researcher with valuable information to revise a measure. Validating a measure is a never-ending process. The first step should be evaluating the content validity of the measure. Subsequent analyses should include evaluating the reliability (such as internal consistency and test–retest) and construct validity through factor analysis, and correlating the measure with the long version and similar measures and constructs, to name a few.

After completing the content validity study and revising the measure, a pilot study can be undertaken to examine additional psychometric properties. A pilot study can identify coding errors, format problems, and ease of administration. After a pilot study, researchers can evaluate the psychometric properties. Testing a measure should be conducted before testing theory.

Understanding the need for and process of conducting a content validity study is important for social work researchers and practitioners. Social work researchers are obligated to critically review measures they use in their research to determine if the measures are relevant for the construct being measured, the sample population, and the cultural, political, and social characteristics of the times in which they are conducting research. In conducting such a review and critique of the available measures, researchers may conclude that new or revised measures are needed, but few have had the opportunity to develop or revise measures. Having a guide to follow can be helpful.

Moreover, practitioners can be a resource for researchers in developing and revising measures, because they are on the frontlines working with the populations who often become research participants. Training in the area of content validity studies can help practitioners understand the measurement issues and be better informed and effective consultants to researchers in the creation of measures that are appropriate and easily administered for the sample populations. ■

REFERENCES

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Barusch, A. S. (1988). Problems and coping strategies of elderly spouse caregivers. *Gerontologist, 28,* 677–685.

Berg-Weger, M., Rubio, D. M., & Tebb, S. S. (2000). The Caregiver Well-Being Scale Revisited. *Health & Social Work, 25,* 255–263.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validiation by the multitrait—multimethod matrix. *Psychological Bulletin, 56,* 81–105.

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

Gable, R. K., & Wolf, J. W. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings.* Boston: Kluwer Academic.

George, L. K. (1994). Caregiver burden and well-being: An elusive distinction. *Gerontologist, 34,* 6–7.

George, L .K., & Gwyther, L. P. (1986). Caregiver well-being: A multidimensional examination of family caregivers of demented adults. *Gerontologist, 26,* 253–259.

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

Hooker, K., Monahan, D., Shifren, K., & Hutchinson, C. (1998). Mental and physical health of spouse caregivers: The role of personality. *Psychology and Aging, 7,* 367–375.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123,* 207–215.

Knight, B. G., Lutzky, S. M., & Macofsky-Urban, F. (1993). A meta-analytic review of interventions for caregiver distress: Recommendations for future caregiver research. *Gerontologist, 33,* 240–248.

Kramer, B. J. (1997). Gain in the caregiving experience: Where are we? *Gerontologist, 37,* 218–232.

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

Maitland, S. B., Dixon, R. A., Hultsch, D. F., & Hertzog, C. (2001). Well-being as a moving target: Measurement equivalence of the Bradburn Affect Balance Scale. *Journal of Gerontology, 56B,* P69–P77.

Maslow, A. (1962). *Toward a psychology of being.* New York: Van Nostrand.

McCubbin, H. (1982). *Systematic assessment of family stress, resources and coping: Tools for research, education and clinical information.* St. Paul: University of Minnesota Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Ory, M. G., Yee, J. L., Tennstedt, S. L., & Schulz, R. (2000). The extent and impact of dementia care: Unique challenges experienced by family caregivers. Reprinted in full from R. Schulz (Ed.), *Handbook on*

dementia caregiving: Evidence-based interventions for family caregivers. New York: Springer. Retrieved from http://www.aoa.dhhs.gov/carenetwork/ory-article.html

Rubio, D. M., Berg-Weger, M., & Tebb, S. S. (1999). Assessing the validity and reliability of well-being and stress in family caregivers. *Social Work Research, 23,* 54–64.

Rubio, D. M., Rubin, R. S., & Brennan, D. G. (2003). How well does the GRE work for your university? An empirical case study of the Graduate Record Examination across multiple disciplines. *College and University Journal, 79*(4), 11–17.

Shepard, L. A. (1994). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Slivinske, L. R., & Fitch, V. L. (1987). The effect of control enhancing interventions on the well-being of elderly individuals living in retirement communities. *Gerontologist, 27,* 176–181.

Tebb, S. S. (1995). An aid to empowerment: A caregiver well-being scale. *Health & Social Work, 20,* 87–92.

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

Weick, A. N. (1986). The philosophical context of a health model of social work. *Health & Social Work, 20,* 87–92.

This research was supported by the Saint Louis University Faculty Development Fund. The authors would like to thank all the experts who so graciously gave their time and shared their expertise for this study.

Original manuscript received May 18, 2001
Final revision received March 22, 2002
Accepted August 6, 2002