Retaining qualitative validity while gaining quantitative reliability and validity: Development of the Transition to Parenthood Concerns Scale

This article raises issues about the retention of qualitative validity while establishing psychometric estimates of reliability and validity for a scale that was developed from inductively generated concepts to assess and evaluate the learning needs and concerns of expectant parents in the trimester before their baby's birth. A set of practical procedures for pilot testing qualitatively based scales is described. The three procedures, which provide estimates of clarity, apparent internal consistency, and content validity, preserve the assumptions underlying qualitative methods. The ratings from these procedures have provided a base for item and scale revisions and formal quantitative testing.

Margaret A. Imle, PhD, RN
Associate Professor
Department of Family Nursing
Research Facilitator
Office of Research
Oregon Health Sciences University
Portland, Oregon

Jan R. Atwood, PhD, RN
Professor
College of Nursing
University of Arizona
Behavioral Sciences Coordinator
Cancer Prevention and Control
Arizona Cancer Center
Tucson, Arizona

A PRIORITY TASK in clinical nursing research is the measurement of phenomena of interest to nursing in the arena of health promotion and protection.[1,2] The nursing care of childbearing couples is often cited as promoting the healthy transition of prospective parents to parenthood. Thinking about and dealing with concerns about one's transition from adult to adult-plus-parent is a phenomenon that is apparently unique to first-time prospective parents during this developmental period.

Holistic nursing care can make a significant impact on the parenthood transition experience if concerns can be identified and addressed. Essential to such nursing

An earlier version of this article was presented at the Western Society for Research in Nursing Conference, Albuquerque, New Mexico, April 30, 1981.

The research on which this article is based was supported in part by US Department of Health and Human Services National Research Award No. 1F31-NU-05347; by Beta Mu Chapter of Sigma Theta Tau; and by the University of Arizona College of Nursing.

The authors acknowledge prepublication critiques of this manuscript by Drs Jo Anne Horsley and Virginia Tilden.

Adv Nurs Sci 1988;11(1):61-75

care are client assessments and/or evaluations of nursing care that are reliable and valid measures and are sensitive indices not only to the presence of the phenomenon but also to changes in its status. However, knowing the nature of phenomena and developing measures that capture them may be challenging tasks for nurse researchers,[3] especially when the relevant concepts are not fully known or adequately described.

The learning needs and concerns of expectant parents in the transition to parenthood provide an example of a common clinical phenomenon not being described or indexed conceptually, most specifically from the viewpoint of the parents themselves. Using qualitative research procedures, researchers[4] inductively developed and defined concepts that were valid representations of the needs and concerns normally experienced by the parents themselves. The purpose of this article is to identify strategies for retaining qualitative validity while gaining quantitative reliability and validity with inductively generated concepts that were then used to generate a scale that would assess and evaluate the learning needs and concerns of prospective parents, the Transition to Parenthood Concerns (TPC) Scale.[4]

So many of the phenomena that are significant problems to nursing have yet to be isolated, named, and described that they are only now being conceptualized; once they are identified, the need to measure them is pressing.[5,6] This explication process is inherent in the conduct of nursing research on clinical practice phenomena where conceptual specifications are absent or incomplete.[7]

The TPC Scale had been developed to assess quantitatively the level of expectant parents' learning needs and concerns,[4] in order to plan and evaluate their nursing care. The challenge became one of creating scales and scale items to index the concepts in such a way that they would retain their validity for the parents themselves and yet would meet the criteria for adequate psychometric performance. While inductive qualitative methods are appropriate to discover and delineate such empirically grounded concepts, the transformations necessary to produce quantitative items for use with psychometric procedures may not preserve the meaning of the concepts. When there has been a high investment of resources to capture qualitatively the meaning of a phenomenon, it is essential that this meaning be preserved in any quantitative instruments developed to measure the concepts. In this article, issues of scale development will be raised and practical approaches will be proposed.

QUALITATIVE VERSUS QUANTITATIVE CONCEPTS

A serious challenge to the nurse researcher who is developing instruments from qualitatively generated concepts is the evaluation of whether the meaning inherent in qualitatively generated concepts has been retained in scales constructed for a quantitative instrument. Easy, yet sound procedures are needed to pilot test a scale and its quantitative items in a timely and practical way while preserving the assumptions underlying the qualitative mode of inquiry. While triangulation using qualitative and quantitative methods has been described as a means of validating findings,[8] the interface between qualitative

and quantitative methods for scale development is less clear.

At issue is the interface between qualitative and psychometric methods. The means of quantitatively estimating the reliability and validity of scales are well-documented in psychometric literature.[9-13] While careful assessment of newly developed instrument items is recommended in those texts, the operational procedures for preliminary assessments of the meaning of items and scales are less clearly described. Furthermore, the psychometric literature is silent about sound procedures to assess clarity, apparent internal consistency, and content validity for items and scales developed from new domains that are generated inductively. Yet, these pilot-testing procedures are crucial to the later performance of these scales, as estimated by internal consistency reliability, criterion-related validity, and construct validity.[11]

Quantitative testing of qualitatively generated material was identified by Glaser and Strauss[6] as a viable means of meeting the need for new nursing measures. Knafl and Howard[14] have more recently supported this idea by identifying conceptualization and instrumentation as two purposes for qualitative research. The problems of reliability and validity thus present a double challenge for the nurse researcher, who must (1) accurately conceptualize the phenomenon and (2) develop the scale prior to doing quantitative psychometric estimates of the reliability and validity of the scale. These challenges may be expressed by one question: How does the researcher retain qualitative meaning and yet demonstrate evidence of meeting psychometric criteria for an adequate scale?

Of the psychometric criteria for an adequate scale, one criterion that is well accepted in the scholarly communities of both qualitative and psychometric research is that of the adequate depiction of the essential components of a phenomenon.

An answer to the question is that the researcher uses procedures that draw upon operations of qualitative methods to preserve meaningful definitions of inductively derived concepts at the same time that he or she is developing items that are amenable to psychometric assessment procedures. Of the psychometric criteria for an adequate scale, one criterion that is well accepted in the scholarly communities of both qualitative and psychometric research is that of the adequate depiction of the essential components of a phenomenon. Qualitative researchers would satisfy this criterion through the saturation of the conceptual category,[15] the preservation of the context within which the concept was developed,[16] and perhaps the preservation of an emic perspective.[17] Psychometric researchers would meet this criterion through the domain sampling model.[11] Domain sampling, a term well known in psychometric circles, is the expectation that the theoretical universe of all items about a conceptual domain have been sampled in a way that makes an instrument representative of the conceptual domain.[11]

PILOT-TESTING STRUCTURE

Miles and Huberman[18] issued a call for researchers to contribute to a scholarly

repertoire of analysis methods by which the validation of qualitative data may be more apparent. They concluded that there was a need for "widespread experimentation, documentation, and sharing of methodological advances among qualitative researchers."[18(p20)] In answering their call to share, this article will describe the application of a practical procedure for assessing the validity and internal consistency of inductively generated domains in pilot testing the TPC Scale, which was constructed from qualitatively generated concepts.[4] The researcher who can obtain an indicator of scale performance through small-sample pilot procedures is being both logical and thrifty.

The basic strategy described here involves quantitative pilot-testing of a newly generated instrument and its items in the same way that the qualitative data supporting the inductively generated concepts have been tested for meaning (validity) and reliability. The original TPC Scale consisted of subscales, each representing a concept, and a set of items per subscale. Pilot testing the TPC Scale included assessments of clarity, apparent internal consistency, and content validity. The pilot-testing procedure was an adaptation of a concept assessment procedure, first recommended by Atwood,[19] and operationalized subsequently in several studies,[20-22] to apply criteria for reliability and validity to concepts generated through grounded theory analysis. Applying the recommendations of Aamodt,[16] the strategy involved preserving and elaborating the same context as that used in generating the qualitative data for the pilot testing of a newly generated quantitative instrument and its items. Context is the crucial information describing the situation within which the qualitatively induced concepts have been developed or will apply.

Field and Morse[5] suggested several strategies by which to assess qualitative data for reliability and validity, as a way of preserving the assumptions inherent in inductive methodology. One of the strategies that can be applied to instrument development is participant review of findings to ensure that both researcher and informant view the data in a consistent way.[18] Informants need to agree not only on definitions, but also on the clusters of data. This step ensures that other observers besides the researcher can agree about the internal cohesion of each conceptual category, even without the aid of a definition. Essentially this is a criterion of domain, used to check whether the constant comparative analysis process has worked, in a way detectable by others, to produce conceptual categories that are both mutually exclusive and internally consistent.[15]

The ratings used here for these pilot procedures relied on the raters' degree of agreement with the researcher's conclusion. Agreement, defined as percentage agreement, is the proportion of raters on the panel who agree with the researcher about a particular item or scale rating. Thus, agreement of five of six raters on a panel for any single rating task, or item rating, gave an 83% agreement; four out of six achieved a 67% agreement, and eight out of ten raters achieved an 80% agreement. Mean ratings for a scale or subscale were obtained by averaging the item ratings across all items in the respective scale or subscale.

Rating panel

To preserve the context of the data and to retain the accuracy of meaning—and thus the qualitative validity—the raters should be drawn from the context within which the original qualitative data were generated.[16] Consistent with the recommendations of Aamodt[16] and Field and Morse,[5] the pilot-testing procedure described here accomplished this aim by having the same type of informants who had provided data for the initial concept definition give their views about the consistency between the inductively generated concepts and the scales developed. That is, the raters used were expectant parents similar to those who had generated the qualitative data for concept generation. Expectant parents were used for all phases of the pilot testing, and a sample of doctoral students in clinical nursing research was used to screen out scale items that had a double meaning or were ambiguous.

Psychometric procedures

Judgments about clarity, internal consistency, and content validity of the scale items and subscales were sought through a systematic procedure, which had three parts, each one a pilot-rating procedure for scales and/or items. The three psychometric assessments are described here and illustrated. Item clarity, consistency, and content validity have the same measurement connotation, regardless of how the concept is derived.[22]

Item clarity is desired to convey a single message or portion of the inductively generated concept. Since the success of subsequent ratings depended on how clearly the scale items communicated a portion of the concept, clarity assessment was the first judgment requested of a panel of informants. Rating apparent internal consistency, a preliminary indicator for homogeneity of content, was the next psychometric assessment to be made. An estimate of the internal consistency of a scale, the degree to which items are grouped together, is typically made by a quantitative test using alpha, theta, or split-half coefficients for deductively derived instruments.

Because no psychometric term existed to describe the preliminary assessment of homogeneity of content, the researchers chose the phrase "apparent internal consistency" to describe the nonquantitative assessment of that quality—homogeneity. Homogeneity of content, indicated by apparent internal consistency, was assessed for qualitatively generated scale content; the informants were asked to make judgments about whether or not each item fit and whether or not they all fit together. Following this judgment of apparent internal consistency with definition-free groupings of items, the panel members were given the necessary definitions and concept labels and were asked to make content judgments about each respective set of items. Panel judgments for content validity were similar to those typically used for instruments measuring deductively derived concepts; the informants were asked to consider the labels, definitions, and item content to see if they matched.

MATERIALS AND PROCEDURES

The authors developed a format for applying the three pilot-testing procedures

to scales that index qualitatively derived concepts. Each set of materials consisted of

- instructions for the raters;
- the list of numbered scale items; and
- a response sheet with corresponding numbers (Figs 1-3).

This three-part format was specified for the three procedures: clarity, apparent internal consistency, and content validity. The instructions specified how the raters were to evaluate each set of items. The list of items and the response section were each printed on a vertical half-page; thus they could be placed side by side, the list to the left and response to the right, for ease of completion.

Each rating procedure required a separate set of instructions to help raters respond to the issue of clarity, the apparent internal consistency of a group of items, and the content validity of the items within a subscale indexing a defined concept.

Each rating procedure required a separate set of instructions to help raters respond to the issue of clarity, the apparent internal consistency of a group of items, and the content validity of the items within a subscale indexing a defined concept.

The sequence of rating procedures was important to the study. For example, the apparent internal consistency of items needed to be assessed before content validity, because, once the panel became aware of the definitional content, they would be unable to make content-free reliability estimates.[23] The selection of items for a particular rating procedure also was important. The first rating procedure—that including clarity of items—was done on groups of items that were mixed from all subscales, to prevent the meaning of any one item from cuing the rater about the meaning of any other items (see Fig 1). Each group of items was rated, until all scale items had been rated, before the raters moved on to the next procedure.

The second rating procedure was apparent internal consistency, for which each set consisted only of items from a particular subscale. Each list to be assessed for apparent internal consistency was presented to the rater as a set of items that was unnamed and undefined for the purposes of that rating procedure. Each set, one for each subscale, was rated for apparent internal consistency prior to the third rating procedure, which was done by the same panel. In this third rating procedure, the same sets of subscale items were presented to the panel with new response sheets, new instructions specified for content validity, labels, and defined concepts to help determine content validity and domain sampling.[11]

In the following sections, each of the three variations in the set of rating materials will be presented, with discussion of the psychometric purpose of the test-rating steps, rater activities, and examples of how the test-rating procedure has worked in developing a new instrument, the TPC Scale.[4] The untested draft of the original scale had eight subscales, each with a set of items. At the completion of the following procedures for pilot testing, the TPC comprised seven subscales and their respective items.

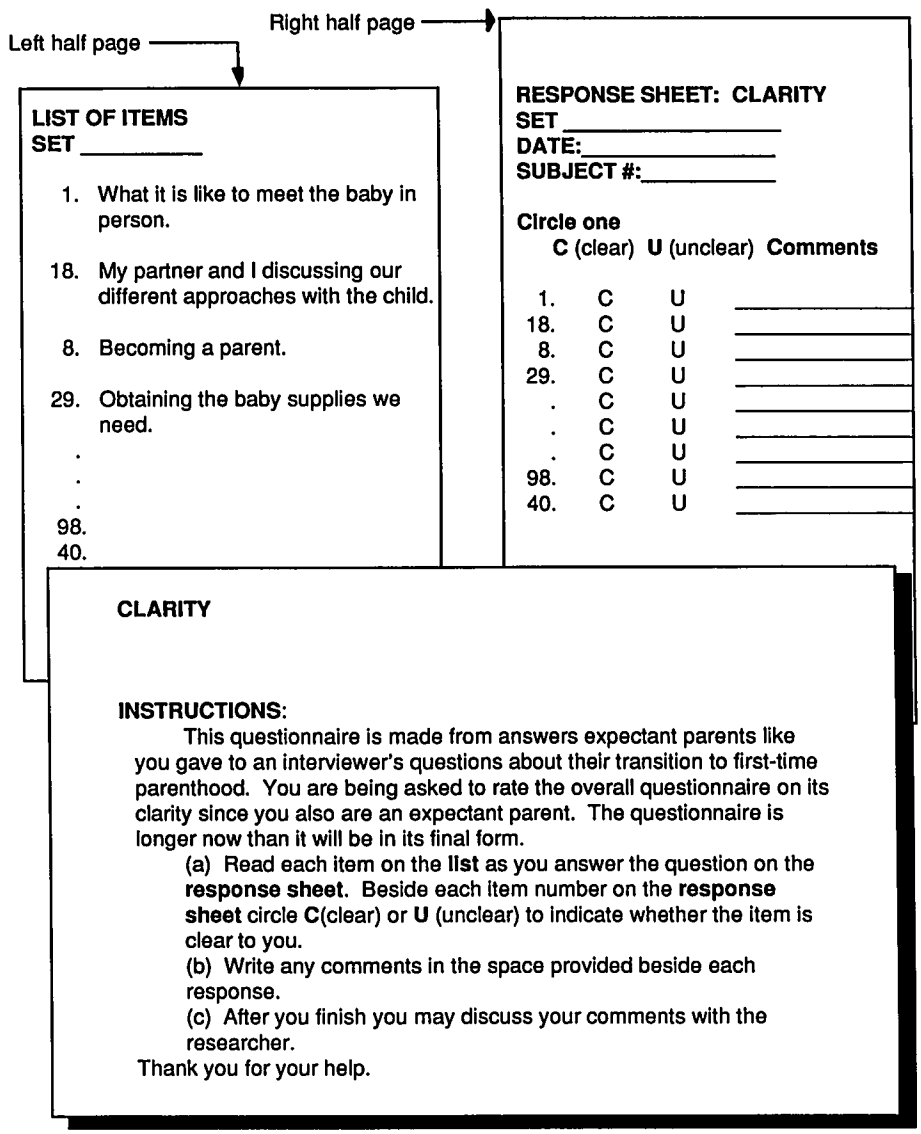

Fig 1. A set of rating materials; this example rates clarity.

Procedure for clarity rating

The first rating procedure, clarity, had two parts: clarity of scale items and clarity of scale directions. Raters were to determine if each item was successful at conveying a single portion of the inductively generated concept and if the directions were successful in guiding respondents to complete the scale. Raters received sets of materials, each with identical instructions but with different sets of items. In the early steps of scale development, there is often a

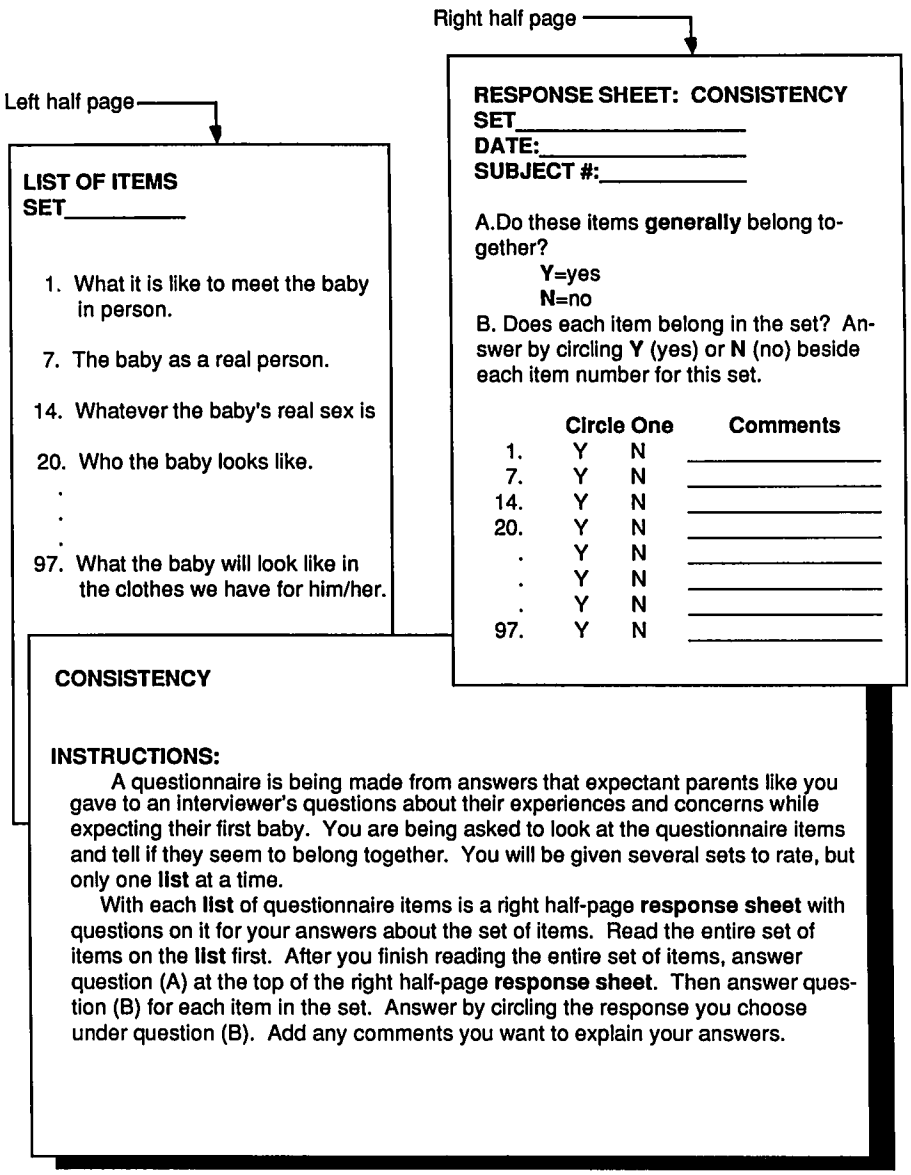

Fig 2. A set of rating materials; this example rates apparent internal consistency.

larger-than-needed pool of items, which, if presented all at once, would discourage most test-rating panels. Thus, the eight sets contained items selected randomly from the eight subscales and indexing the eight concepts, so that all subscales were equally likely to have one or more items appearing within a given set. Raters were asked to take each set in the order presented and then to proceed to the next set.

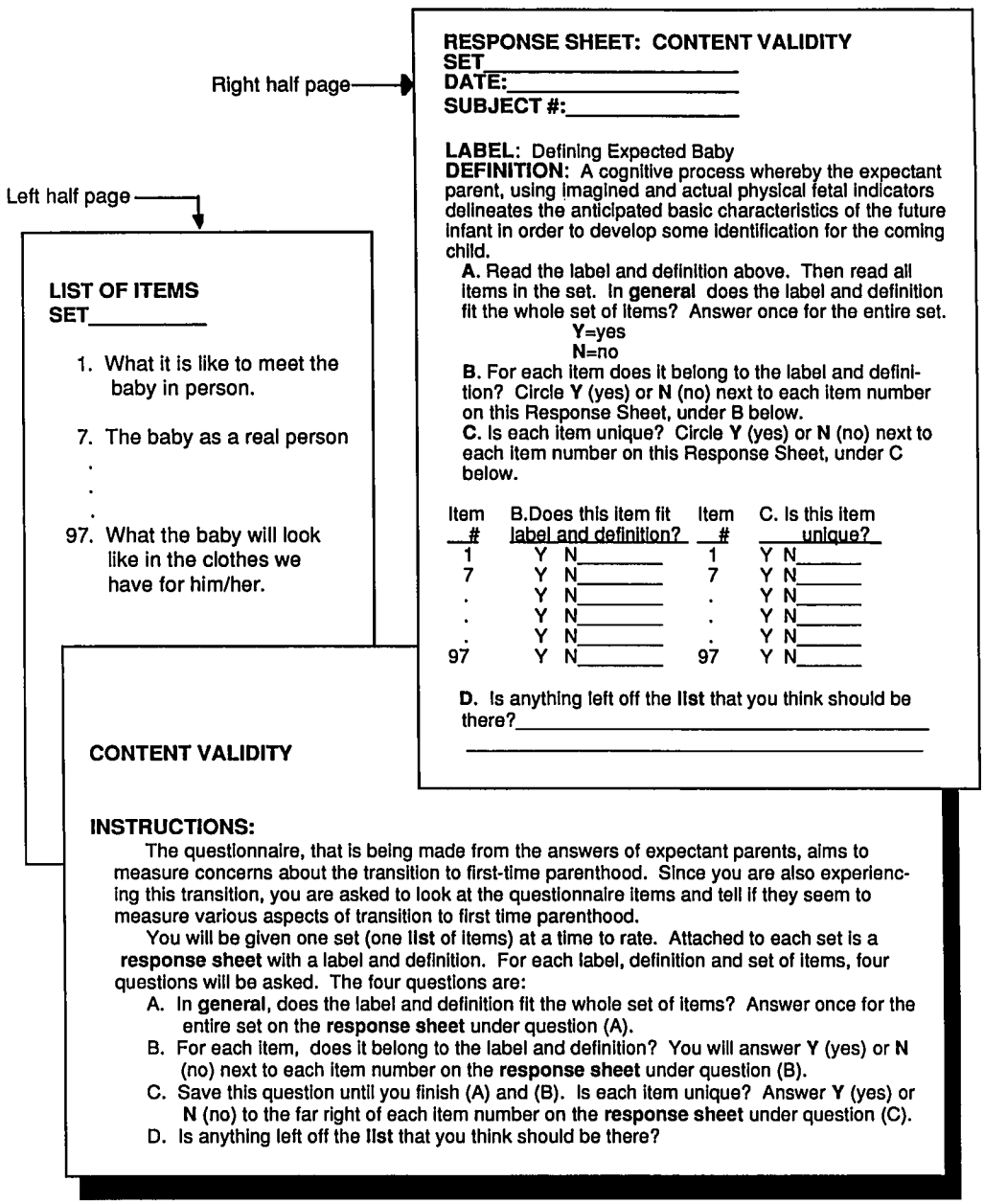

Fig 3. A set of rating materials; this example rates content validity.

The set of materials for rating clarity is shown in Fig 1, and only that psychometric quality was addressed by this panel of raters. Space for comments was provided beside each item response, and the rater was encouraged to discuss comments with

the researcher after he or she finished rating all items in the eight sets. The scale directions were also examined for format, clarity, and reading level by having the raters read them, after they had finished clarity ratings on the eight sets, and note any unclear portions. Raters tried out several scale items, using the directions, to see if the directions were adequate and if scale items could be responded to with the Likert[24] set of response options. Before the rater assessment, an a priori criterion of 66% agreement had been set for clarity of each scale item and of scale instructions, with 80% the criterion for the overall scale. The average of the ratings per item yielded the average rating for a subscale.

In the TPC instrument development study,[4] the eight sets of items were rated, as were the directions, first by ten doctoral students in clinical nursing research and then by five expectant parents in the last trimester before the birth of their babies. Revisions were made after the doctoral student panel's ratings, and again after the expectant parents' ratings. The doctoral student panel was helpful in pointing out items that conveyed more than a single idea. The expectant parents' ratings and comments helped to determine items with unclear wording or ideas too idiosyncratic to be readily understood. The qualitative data that had helped to generate the concept had then been used to guide the researcher in wording the scale items to retain language and expression used by expectant parent informants. Use of such data can be helpful, because the expressions used may be particularly clear to other subjects like the original informants; however, the trade-off is that idiosyncracies of the original informants may also be retained.

Having doctoral students rate items for ability to express a single idea and having expectant parents rate items for clarity of ideas made it possible for items to be both clear and applicable to other expectant parents. These groups rated 114 items; 98 were retained, either as originally stated or after one or more revisions, by having met the criterion of 66% agreement among members of the rating panel, with the mean percentage agreement for the total scale being 77%. This provided the nearest approximation to the 80% criterion for the five raters' responses to all acceptable items. Through clarity-driven revisions, 18 additional items were developed, for a total of 116 items to be rated for apparent internal consistency, the next pilot procedure.

Procedure for rating apparent internal consistency

Adequate internal consistency, the degree to which all scale items group together, is a baseline requirement for both reliability and content validity, according to the domain sampling model.[11] Domain theory is based on the idea that there is a hypothetical population of items for which there is an infinitely large correlation matrix with hypothetical correlations among all items. The average correlation of all hypothetical items in a given domain is the extent to which a common core, or internal homogeneity, exists for the items in the domain. This homogeneity within a domain serves as the basis for later estimates of both internal consistency reliabil-

ity and content validity. While the sampling of a domain must be representative of various facets of the domain, it must also reflect a consistency within the domain such that items are sufficiently homogeneous. Assessment of apparent internal consistency, a preliminary indicator of the level of internal consistency that was obtained by later quantitative testing, is the purpose of the second test-rating procedure. Content validity will be addressed later by the third test-rating procedure.

Pilot procedures that are based on the central ideas behind the domain sampling model yield preliminary indicators of internal consistency and enable the researcher to revise scales before going to the greater expense and effort for large-sample quantitative testing of internal consistency reliability. Thus the purpose of the second test-rating procedure was to obtain a rough indicator of the internal consistency of a scale, so that problems related to the homogeneity of the domain sampling could be addressed and the scales could be revised. The set of materials the raters received for this task is shown in Fig 2. Using the same set of instructions for each subscale, the raters proceeded through the items. For this apparent internal consistency pilot procedure, raters were not given the names of the subscales or any information by which they could identify the subscale name or purpose, since this was an estimate of reliability and not validity.

The rating activity consisted of answering two questions: "Do these items generally belong together?" and "Does each item belong in the set?" After rating

The rating activity consisted of answering two questions: "Do these items generally belong together?" and "Does each item belong in the set?"

all eight subscales, raters were encouraged to jot down comments in the space provided and to share them with the researcher. The a priori criterion for an item or subscale to be retained in the scale was 70% agreement among raters per item and an average of 70% agreement per subscale. Because only six raters were used, 67% agreement was the nearest approximation to the criterion of 70%. Using this test-rating procedure, 116 items, in 8 subscales, were assessed for apparent internal consistency. All 8 subscales and 99 of the 116 items met or exceeded the criteria, with a mean percentage agreement of 82%.

Procedure for rating content validity

The third of the test-rating procedures, that for the content validity of the subscales, assesses the capability of a set of items truly and adequately to sample the content in each qualitatively generated domain. In essence, it tests how well the scale fits the domain. This validity assessment is based on the domain sampling model.[11] While internal consistency is the necessary basis for content validity, it alone is not sufficient. Content validation rests upon how well and how adequately items tap the meaning of the conceptual domain,

as it is named and defined for the reader, and how well they avoid redundancy in tapping the domain.

For ratings of content validity, the raters received a set of materials (Fig 3) with a format slightly expanded over the previous sets. This format included

- an instruction page with four questions;
- the list of items, one list per subscale; and
- an item-numbered response sheet with the label of the subscales and the concept definition.

Each rater responded to one complete set, ie, the items of a given subscale, before moving on to the next set. The questions to which the rater must respond "yes" or "no" are shown in the figure. Raters were to respond to the first two questions prior to answering the last two questions and were asked for their comments. After rating all subscale sets, the raters shared their comments with the researcher.

In the present study,[4] the 116 items that had been assessed for apparent internal consistency were also rated for content validity by the same six-member panel. The a priori criteria for acceptance were 80% agreement per subscale set and per item for the first two questions and 85% agreement on uniqueness for the third question; however, with six raters, the practical criterion was 83% agreement for both procedures, the nearest approximation to the a priori criteria. Eighty-seven of the 116 items, with a mean percentage agreement of 86%, met both the 83% criteria, but only seven of the eight subscales met the 83% criterion for fit between the subscale and/or items and the label and definition. For uniqueness, most items and all subscales met the criterion of at least 83% agreement. However, only 3 out of 15 items on the Becoming Mentally Ready Subscale met the 83% criterion for fit, and the subscale was judged by one third of the panel not to fit generally with the label and definition. Ten out of 15 items on this subscale met criteria for apparent internal consistency (with a subscale mean of 84% agreement), but the subscale did not validly fit the domain once a label and definition were apparent to the raters.

The rating of this subscale illustrates one reason why the same panel of raters who respond to the apparent internal consistency ratings also do the content validity ratings. Both types of ratings are needed to make scale revisions related to domain membership and consistency. If two panels had been used and the reliability and validity ratings were different, the researchers could not tell if the differences were due to differences between panels or to genuine differences between reliability and validity ratings. However, by having the same panel, the researchers could rule out differences between panels as a cause for different ratings on domain sampling issues. In this case, the sampling of items apparently was internally consistent, but it did not fit well with the conceptual labeling and definition, which was a validity issue.

ISSUES RELATED TO CRITERIA AND RATING PROCEDURES

In the pilot study,[4] several issues became apparent. One was the use of terms that may be unclear to raters with lower verbal skills. In the clarity procedure, the term "ambiguous" had to be replaced with "un-

clear" for many raters to understand the directions and response sheet. In the content validity procedure, the term "unique" triggered questions about its meaning, requiring the researcher to explain that it meant that items did not duplicate each other.

A second issue was the choice of rating procedure. Topf[25] describes the uses and relative advantages of three types of estimates of interrater reliability for dichotomous data. The present pilot-testing study used percentage agreement among raters. Topf conceptualizes interrater agreement as a 2 x 2, or larger matrix, wherein agreement between two raters about occurrence and nonoccurrence of an event are considered. The present pilot study[4] did not fit this matrix, because there were not multiple opportunities to rate a given item by any rater. Instead, multiple raters assessed a given item only once for agreement with the researcher about the placement or wording of an item.

Topf[25] has noted that percentage agreement may be very different from the correlation-based interrater procedures, which are Cohen's kappa and phi. Percentage agreement has the relative advantages of ease of calculation and of being more informative to the reader, but it is less related to the percentage of variance explained than a correlation-based coefficient would be. Because the purpose of this study was pilot testing a scale for psychometric attributes rather than explaining outcomes in terms of another variable, obtaining the percentage of explained variance would not have been a meaningful procedure. Thus, the researchers chose to use percentage agreement.

The likelihood of chance agreement between two raters increases as the number of occurrences of an event is increased, which is a major criticism of percentage agreement. The use of percentage agreement where there is only one occurrence of an event (per item) seems to avoid this pitfall. While percentage agreement is less related to correlation-based estimates such as Cronbach's alpha,[26] the percentage agreement was chosen as a screening procedure, rather than as a substitute for quantitative estimates. Quantitative estimates are based on the homogeneity of the items within a scale, while the interrater reliability issue is one of agreement across raters. Thus the choice for the pilot study was to use percentage agreement because of its ease of calculation, intuitive clarity, and the single opportunity for more than two raters to agree/disagree with the researcher. As Topf recommended, the type of percentage agreement is reported so that readers may be clear about the procedure used.

A third issue was the criteria to be used for the pilot study and their relationship, if any, to later quantitative estimates. Topf reports that criteria for percentage agreement are based on a general consensus among behavioral scientists: "70% is necessary, 80% is adequate, and 90% is good."[25(p254)] These percentage agreement criteria roughly parallel the ranking of criteria for quantitative testing of scales. For example, Nunnally[11] states that a Cronbach's alpha of at least 0.80 is adequate for a scale, but 0.70 is adequate for an immature scale. Nunnally further recommends that a coefficient of 0.90 is best for scales that will influence clinical decisions.

The criteria chosen for validity were higher than those for reliability, because

The criteria chosen for validity were higher than those for reliability, because reliability is necessary, but not sufficient, for validity.

reliability is necessary, but not sufficient, for validity. Thus a criterion of 80% for validity was selected, in relation to the 70% agreement criterion used for assessing apparent internal consistency. To screen out duplicated items, an even higher criterion, that of 85% agreement on uniqueness, was used. Only the best items, those items receiving at least adequate ratings for clarity, apparent internal consistency, and validity of fit between item and concepts, that were also unique, were desired for a scale that could be useful in a clinical area for assessment of parents' concerns. The subsequent quantitative testing of the TPC with expectant parents in the last trimester before the birth of their babies yielded Cronbach's alphas, estimates of internal consistency, ranging from 0.79 to 0.92 for 45 mothers-to-be and from 0.84 to 0.91 for 36 fathers-to-be on the seven subscales.

• • •

The rationale for practical pilot procedures that can be applied without loss of meaning to scales developed from inductively developed concepts has been discussed. This article described a set of pilot procedures that are related both to the qualitative method and to the quantitative psychometric estimates needed for scale testing. The three procedures have been illustrated as used with the TPC Scale,[4] which was developed from inductively generated concepts to provide estimates of clarity, apparent internal consistency, and content validity, while preserving the assumptions underlying the qualitative methods. Issues related to the use of various rating procedures were considered; however, the ease and intuitive clarity of percentage agreement ratings made them the ideal choice to summarize quick and easy pilot assessment procedures. Further study may show the relative usefulness of alternative estimates of interrater agreement, but the basis for such ratings needs to be examined conceptually in a methodological study. The set of percentage agreement ratings provided an economical basis for item and subscale revisions of the TPC Scale prior to quantitative testing. Subsequent quantitative testing of the TPC Scale produced Cronbach's alpha estimates of internal consistency that were adequate for an immature scale. This evidence supported claims about qualitative validity (both mutually exclusive and internally homogeneous) provided by the constant comparative analysis process.

REFERENCES

1. American Nurses' Association: *Social Policy Statement.* Kansas City, Mo, American Nurses' Association, 1983.
2. American Academy of Nursing: *Priorities Within the Health Care System: A Delphi Survey.* Kansas City, Mo, American Nurses' Association, 1981.
3. Oberst M: Nursing in the year 2000: Setting the agenda for knowledge generation and utilization, in Sorensen, G (ed): *Setting Agendas for the Year 2000: Knowledge Development in Nursing.* Kansas City, Mo, American Academy of Nursing, 1986, pp 29-37.
4. Imle MA: *Indices to measure concerns of expectant*

parents in transition to parenthood, dissertation. University of Arizona, Tucson, Ariz, 1983.
5. Field PA, Morse JM: *Nursing Research: The Application of Qualitative Approaches.* Rockville, Md, Aspen Publishers, 1985.
6. Glaser BG, Strauss AL: The purpose and credibility of qualitative research. *Nurs Res* 1966; 15:56-61.
7. Stern PN: Grounded theory methodology: Its uses and processes. *Image* 1980; 12:20-23.
8. Jick TD: Mixing qualitative and quantitative methods: Triangulation in action. *Admin Sci Q* 1979; 24:602-611.
9. Anastasi A: *Psychological Testing,* ed 4. New York, Macmillan, 1982.
10. Kerlinger FN: *Foundations of Behavioral Research,* ed 3. New York, Holt, Rinehart & Winston, 1986.
11. Nunnally J: *Psychometric Theory.* New York, McGraw-Hill, 1978.
12. Waltz C, Strickland O, Lenz E: *Measurement in Nursing Research.* Philadelphia, F.A. Davis, 1984.
13. Zeller RA, Carmines EG: *Measurement in the Social Sciences.* Cambridge, Mass, Cambridge University Press, 1980.
14. Knafl KA, Howard MJ: Interpreting and reporting qualitative research. *Res Nurs Health* 1984; 7:17-24.
15. Glaser BG, Strauss AL: *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago, Aldine, 1967.
16. Aamodt AM: Problems in doing research: Developing criteria for evaluating qualitative research. *West J Nurs Res* 1983; 5:398-401.
17. Pelto PF, Pelto GH: *Anthropological Research: The Structure of Inquiry.* Cambridge, Mass, Cambridge University Press, 1978.
18. Miles MB, Huberman AM: Drawing valid meaning from qualitative data: Toward a shared craft. *Educ Res* 1984; 13:20-30.
19. Atwood JR: *The Application of Grounded Theory to Nursing Research.* Abstract, Faculty Theory Update, University of Arizona, Tucson, Arizona, Jan 12, 1975.
20. Atwood JR: The phenomenon of selective neglect, in Bauwers E (ed): *The Anthropology of Health.* St Louis, Mosby, 1978, pp 192-200.
21. Imle MA, Atwood JR: *Inductively identified expectant parents' concerns.* Presented at the 14th Annual Communicating Nursing Research Conference, Western Society for Research in Nursing, Albuquerque, NM, April 30, 1981.
22. Atwood JR, Hinds P: Heuristic heresy: Application of reliability and validity criteria to products of grounded theory. *West J Nurs Res* 1986; 8:135-147.
23. Atwood JR: Strategy for theory development: Grounded theory. *Proceedings of the First Annual Nursing Sciences Colloquium.* Boston, Boston University School of Nursing, 1985, pp 37-55.
24. Likert RA: A technique for the measurement of attitudes. *Arch Psychol* 1932; 140:5-55.
25. Topf M: Three estimates of interrater reliability for nominal data. *Nurs Res* 1986; 35:253-255.
26. Cronbach LJ: Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; 16:297-334.