

NINTH EDITION

# NURSING RESEARCH



Generating and Assessing  
Evidence *for* Nursing Practice


Denise F. Polit • Cheryl Tatano Beck

## Developing and Testing Self-Report Scales

**R**esearchers sometimes are unable to identify an appropriate instrument to operationalize a construct. This may occur when the construct is new, but often it is due to limitations of existing instruments. Because this situation occurs fairly often, this chapter provides an overview of the steps involved in the development of high-quality self-report scales.

The scope of this chapter is fairly narrow, but it covers instruments that nurse researchers often use. First, we focus on structured *self-report* measures rather than observational ones (although many steps would apply to observational scales). Second, we describe methods of developing *multi-item* scales (i.e., not 1-item visual analog scales). Third, we exclude infrequently used scale types, such as semantic differentials. Fourth, we focus on scales rooted in classical measurement theory rather than on item response theory. We use examples of scales to measure the *affective domain* (e.g., measures of attitudes, psychological traits, and so on) rather than scales to measure the *cognitive domain* (e.g., achievement, knowledge), but many principles apply to both domains.

---

 **TIP:** The development of high-quality scales is a lengthy, labor-intensive process that requires some statistical sophistication. We urge you to think carefully about embarking on a scale-development endeavor and to consider involving a psychometric consultant if you proceed.

---

### BEGINNING STEPS: CONCEPTUALIZATION AND ITEM GENERATION

#### Conceptualizing the Construct

The importance of a sound, thorough conceptualization of the construct to be measured cannot be overemphasized. You will not be able to quantify an attribute adequately unless you thoroughly understand the **latent variable** (the underlying construct) you wish to capture. In measurement theory, the latent variable, which is not directly observable, is the *cause* of the scores on the measure. The strength of the latent variable is presumed to trigger a certain numeric value on the scale. You cannot develop items to produce the right score, and you cannot expect good content and construct validity if you are unclear about the construct, its dimensions, and its nuances.

Thus, the first step in scale development is to become an *expert* on the construct. This means being knowledgeable about relevant theory, research relating to the construct, and existing (albeit imperfect) instruments. Scale developers usually begin with a thorough review of relevant literature, on which they can base their conceptual definitions.

Most complex constructs have a number of different facets or dimensions, and it is important to

identify and understand each one. In part, this is a content validity consideration: For the overall scale to be content valid, there must be items representing all facets of the construct. Identifying dimensions also has methodologic implications. All scales—or subscales of a broader scale—need to be unidimensional and internally homogeneous, so an adequate number of items (operational definitions) of each dimension needs to be developed.

During the early conceptualization, you also need to think about related constructs that should be differentiated from the target construct. If you are measuring, say, self-esteem, you have to be sure you can differentiate it from similar but distinct constructs, such as self-confidence. In thinking about the dimensions of the target construct, you should be sure that they are truly aspects of the construct and not a different construct altogether.

Before you begin, you should also have an explicit conceptualization of the population for whom the scale is intended. For example, an anxiety scale for a general population may not be suitable if your interest is in measuring childbearing anxiety in pregnant women. There are arguments for developing patient-specific scales, particularly with respect to the relevancy of items. On the other hand, developing a highly focused scale with low “bandwidth,” while possibly enhancing “fidelity” (Cronbach, 1990), reduces the scale’s generalizability and researchers’ ability to make comparisons across populations. The point is that you should have a clear view of how and with whom the scale will be used.

Understanding the population for whom the scale is intended is critical for developing good items. Without a good grasp of the population, it will be difficult to consider such issues as reading levels and cultural appropriateness in wording the items.

For instruments that are being developed for use by others, it is advisable to establish an expert panel to review domain specifications in an early effort to ensure the content validity of the scale (AERA, APA, & NCME Joint Committee, 1999). An iterative, Delphi survey-type approach with opportunities for refinement by the expert panel is often useful (Berk, 1990).

## Deciding on the Type of Scale

Before items can be generated, you need to decide on the type of scale you wish to create because item characteristics vary by scale type. Our focus is restricted to the most widely used scale types because this is not a textbook on psychometrics. For those interested in such scaling approaches as semantic differentials, Guttman or Thurstone scaling, multidimensional scaling, ipsative (forced choice) scaling, or other approaches, consult other references (e.g., Gable & Wolfe, 1993; Nunnally & Bernstein, 1994; Waltz, et al., 2010).

In this chapter, we concentrate on multi-item summated rating scales, which are also the focus of several other books on scale development that can be consulted for greater elaboration (DeVellis, 2003; Streiner & Norman, 2008). Two broad categories of scales fall into this category: traditional Likert scales and latent trait scales.

Traditional Likert scales (Chapter 13) are based in classical measurement theory (CMT). Items on Likert scales, it may be recalled, are declarative statements with a bipolar response scale that is often on an agree/disagree continuum. In CMT, the scale developer selects items that are presumed to be roughly comparable indicators of the underlying construct. The items gain strength in approximating a hypothetical true score through their aggregation. Traditional Likert scales, then, rely on items that are deliberately redundant, in the hope that multiple indicators of the construct will converge on the true score and balance out error.

Item response theory (IRT), an alternative to CMT, is widely used in creating cognitive tests, and its use in developing affective measures is growing. IRT methods differentiate error more finely than CMT methods, particularly with respect to item characteristics. The goal of IRT is to allow researchers to determine the characteristics of items independent of who completes them. **Latent trait scales** are developed using an IRT framework, and although it is beyond the scope of this book to elaborate on the complex statistical procedures involved in testing latent trait scales, we can provide a few brief comments and references for those who wish further guidance.

Latent trait scales can use items like the ones used in CMT, such as items with a Likert-type format—in fact, a person completing a Likert scale would likely not know whether it had been developed within the CMT or IRT framework. But a person *developing* a Likert-type scale must decide in advance which measurement approach is being used because the scale items would be different. Whereas the items on a CMT Likert scale are designed to be similar to each other to tap the underlying construct in a comparable manner, items on a latent-trait IRT scale are carefully chosen and refined to tap different degrees of the attribute being measured.

As an example, suppose we were developing a scale to measure risk-taking behavior in adolescents. In a CMT scale, the items might include statements about risk-taking of similar intensity, with which respondents would agree or disagree. The aggregate of responses would array respondents along a continuum indicating varying propensity to take risks. In an IRT scale, the items themselves would be chosen to reflect different levels of risk-taking (e.g., smoking cigarettes, using drugs, driving a car at 80 miles an hour while text messaging). Each item could be described as having a different *difficulty*. It is “easier” to agree with or admit to lower-risk items than higher-risk items. Item difficulty is one of several parameters that can be analyzed in IRT scale development. When item difficulty is the only parameter being considered in an IRT analysis, researchers often say that they are using a **Rasch model**.

IRT is a more sophisticated approach than CMT for assessing the strengths and weaknesses of individual items, but it is more complex and uses software that is not as readily available. DeVellis (2003) believes that CMT scaling approaches will continue to prevail for affective measures, but suggests that IRT scaling is especially appropriate when the scale involves items that are inherently hierarchical. Those interested in latent trait scales and IRT should consult Hambleton and colleagues (1991) or Embretson and Reise (2000).

**Example of an IRT analysis:** Gómez and colleagues (2007) analyzed the 20-item Death Anxiety Inventory within an IRT framework.

## Developing an Item Pool: Getting Started

The next step is to develop a pool of possible items for the scale. Items—which collectively constitute the operational definition of the construct—need to be carefully crafted to reflect the latent variable they are designed to measure. This is often easier to do as a team effort, because different people articulate a similar idea in diverse ways. Regardless of whether you are doing this alone or with a team, you may be asking: Where do scale items come from? Here are some possible sources:

1. *Existing instruments.* Sometimes it is possible to adapt an existing instrument rather than starting from scratch. Adaptations often require adding and deleting items, but may involve rewording items—for example, to make them more culturally appropriate, or to simplify wording for a population with low reading skills. Permission from the author of the original scale should be sought because published scales are copyright protected.
2. *The literature.* Ideas for item content often come from a thorough understanding of the literature. Since at this point you would already be an “expert” on the construct, this is an obvious source of ideas for items.
3. *Concept analysis.* A related source of ideas is a concept analysis—which you may already have undertaken as a preliminary step. Walker and Avant (2004) offer concept analysis strategies that could be used to develop items for a scale.
4. *In-depth qualitative research.* In-depth inquiry relating to the key construct is a particularly rich source for scale items. A qualitative study can help you to understand the dimensions of a phenomenon, and can also give you actual words for items. Tilden and her colleagues (1990), Beck and Gable (2001), and Gilgun (2004) offer guidance on using qualitative research to enhance the content validity of a new scale. If you are unable to undertake an in-depth study yourself, be sure to pay particular attention to the verbatim quotes in published qualitative reports about your construct.

5. *Clinical Observations.* Patients in clinical settings may be an excellent source of items. Ideas for items may come from direct observation of patients' behaviors in relevant situations, or from listening to their comments and conversations.

**Examples of sources of items:** Jones and Gulick (2009) developed new items for a revised version of the Sexual Pressure Scale, using qualitative data from seven focus groups. Bu and Wu (2008) derived items for the Attitude Toward Patient Advocacy Scale from a literature review and consultation with experts.

DeVellis (2003) urged scale developers to get started writing scale items without a lot of editing and critical review in the early stages. Perhaps a good way to begin if you are struggling is to develop a simple statement with the key construct mentioned in it. For example, if the construct is test anxiety, you might start with, "I get anxious when I take a test." This could be followed by similar statement worded differently (e.g., "Taking tests makes me nervous").

## Making Decisions about Item Features

In preparing to write items, you need to make decisions about such issues as the number of items to develop, the number and form of the response options, whether to include positively and negatively worded items, and how to deal with time.

### Number of Items

In the CMT framework, a **domain sampling model** is assumed, which involves the random sampling of a homogeneous set of items from a hypothetical universe of items relating to the construct. Of course, sampling from a *universe* of all possible items does not happen in reality, but it is a principle worth keeping in mind. The idea is to generate a fairly exhaustive set of item possibilities, given the construct's theoretical demands. For a traditional Likert scale, redundancy (except for trivial word substitutions) is a good thing—the goal is to measure the construct of interest with a set of items that capture the central theme in slightly different ways so that irrelevant idiosyncrasies of individual items will cancel each other out.

There is no magic formula for how many items should be developed, but our advice is to generate a very large pool of items. As you proceed, many items will be discarded. Longer scales tend to be more reliable, so starting with a large number of items promotes the likelihood that you will eventually have an internally consistent scale. DeVellis (2003) recommended starting with 3 to 4 times as many items as you will have in your final scale (e.g., 30 to 40 items for a 10-item scale), but at a minimum there should be 50% more (e.g., 15 items for a 10-item scale).

### Response Options


Scale items involve both a stem (usually a declarative statement), and a set of response options. Traditional Likert scales often involve response options on a continuum of agreement, but other continua are also possible, such as frequency (never/always), importance (very important/unimportant), quality (excellent/very poor), and likelihood (definitely/impossible).

How many response options should there be? There is no simple answer, but keep in mind the goal is to array people on a continuum, and so variability is essential. Variability can be enhanced by including a lot of items, by offering numerous response options, or both. However, there is not much merit in creating the illusion of precision when it does not exist. With a 0–100 range of scores, for example, the difference between a 96 and a 98 might not be meaningful. Also, it has been found that too many options can be confusing to people with limited education.

Most Likert scales have 5 to 7 options, with verbal descriptors attached to each option and—often—with numbers placed under the descriptors to facilitate coding and to further help respondents find an appropriate place on the continuum. An odd number of items gives respondents an opportunity to be neutral or ambivalent (i.e., to choose a midpoint), and so some scale developers prefer an even number (e.g., 4 or 6) to force even slight tendencies and to avoid equivocation. However, some respondents may actually *be* neutral or ambivalent, so a midpoint option allows them to express it. The midpoint can



be labeled with such phrases as “neither agree nor disagree,” “undecided,” “agree and disagree equally,” or simply “?”.

 **TIP:** Here are some frequently used words for response options, with the midpoint term not listed:

- Strongly disagree, disagree, agree, strongly agree
- Disagree strongly, disagree moderately, disagree slightly, agree slightly, agree moderately, agree strongly
- Never, rarely (or seldom), occasionally (or sometimes), frequently (or usually), always
- Very important, important, somewhat important, of little importance, unimportant
- Definitely not, probably not, possibly, probably, very probably, definitely

### Positive and Negative Stems

A generation ago, leading psychometricians advised scale developers to deliberately include both positively and negatively worded statements and to reverse-score negative items. As an example, consider these two items for a scale of depression: “I frequently feel blue,” and “I am happy most of the time.” The objective was to include items that would minimize the possibility of an acquiescence response set—the tendency to agree with statements regardless of their content.

There is now ample evidence that it is not prudent to include both types of items on a scale. Some respondents are confused by reversing polarities. Answering negative item stems appears to be an especially difficult cognitive task for younger respondents. Some research suggests that acquiescence can be minimized by putting the most positive response options (e.g., strongly agree) at the end of the list rather than at the beginning.

### Item Intensity

In a traditional Likert scale, the intensity of the statements (stems) should be similar and fairly strongly worded. If items are worded such that almost anyone would agree with them, the scale will not be able to discriminate between people with different amounts of the underlying latent variable. For exam-

ple, an item such as “Good health is important” would generate almost universal agreement. On the other hand, statements should not be so extremely worded as to result in universal rejection. For example, “Nurses who do not have a Bachelor’s degree should be fired,” is obviously a poor measure of people’s attitudes toward nursing credentials.

For a latent trait scale, scale developers seek a range of item intensities. Yet, even on an IRT-based scale there is no point in including items with which almost everyone would either agree or disagree.

### Item Time Frames

Some items make an explicit reference to a time frame (e.g., “In the past few days, I have had trouble falling asleep”), but others do not (e.g., “I have trouble falling asleep”). Sometimes, instructions to a scale can designate a temporal frame of reference (e.g., “In answering the following questions, please indicate how you have felt in the past week”). And yet other scales ask respondents to respond in terms of a time frame: “In the past week, I have had trouble falling asleep: Every day, 5 to 6 days . . . Never”.

A time frame should not emerge as a consequence of item development. You should decide in advance, based on your conceptual understanding of the construct and the needs for which the scale is being constructed, how to deal with time.

**Example of handling time in a scale:** The Postpartum Depression Screening Scale asks respondents to indicate their emotional state in the past 2 weeks—for example, over the last 2 weeks I: “. . . felt so all alone” or “. . . cried a lot for no reason” (Beck and Gable, 2000, 2001). The 2-week period was chosen because it parallels the duration of symptoms required for a diagnosis of major depressive episode according to the DSM-IV criteria.

### Wording the Items

Items should be worded in such a manner that every respondent is answering the same question. Guidance in wording good items is offered by Fowler (1995) and Streiner and Norman (2008). In addition to the

suggestions on question wording we provided in Chapter 13, some additional tips specific to scale items are as follows:

1. *Clarity.* Scale developers should strive for clear, unambiguous items. Words should be carefully chosen with the educational and reading level of the target population in mind. In most cases, this will mean developing a scale at the 6th- to 7th-grade reading level. But even beyond reading level, you should strive to select words that everyone understands, and to have everyone reach the same conclusion about what the words mean.
2. *Jargon.* Jargon should be avoided. Be especially cautious about using terms that might be well-known in healthcare circles (e.g., lesion) but not familiar to the average person.
3. *Length.* Avoid long sentences or phrases. Simple sentences are the easiest to comprehend. In particular, eliminate unnecessary words. For example, “It is fair to say that in the scheme of things I do not get enough sleep,” could more simply be worded, “I usually do not get enough sleep.”
4. *Double negatives.* It is often preferable to word things affirmatively (“I am usually happy”) than negatively (“I am not usually sad”), but double negatives should always be avoided (“I am not usually unhappy”).
5. *Double-Barreled Items.* Avoid putting two or more ideas in a single item. For example, “I am afraid of insects and snakes” is a bad item because a person who is afraid of insects but not snakes (or vice versa) would not know how to respond.

**Examples of well-worded items:** Ellenbecker and colleagues (2008) revised a scale to measure job satisfaction among home healthcare nurses. Here are two items from their revised scale: “I am able to meet the demands of my job” and “I am satisfied with the amount of control I have over my work.” Respondents indicate agreement or disagreement with items on a 5-point scale.

## PRELIMINARY EVALUATION OF ITEMS

### Internal Review

Once a large pool of items has been generated, it is time for critical appraisal. Care should be devoted to such issues as whether individual items capture the construct, and are grammatical and well worded. The initial review should also consider whether the items taken together adequately embrace the full nuances of the construct—that is, whether additional items need to be generated to enhance the scale’s content validity.

It is also imperative to assess the scale’s **readability**, unless the scale is intended for a population with known high literacy, such as people with advanced degrees. There are different approaches for assessing the reading level of written documents, but many methods are either time-consuming or require several hundreds of words of text, and thus are not suited to evaluating scale items (Streiner & Norman, 2008).

Many word-processing programs provide some information about readability. In Microsoft Word, for example, you could type your items on a list and then get readability statistics for the items as a whole or for individual items, as described in Chapter 7. For example, take the following two sets of items for tapping fatigue:

Set A	Set B
I am frequently exhausted.	I am often tired.
I invariably get insufficient sleep.	I don’t get enough sleep.

The software tells us that the items in Set A have a *Flesch-Kincaid grade level* of 12.0 and a *Flesch reading ease score* of 4.8. (Reading ease scores rate text on a 100-point scale, with higher values associated with greater ease, using a formula that considers average sentence length and average number of syllables). Set B, by contrast, has a grade level of 1.8 and a reading ease score of 89.4. Streiner and Norman (2008) warn that word-processing-based

readability scores be interpreted cautiously, but it is clear from the foregoing analysis that the second set of items would be superior for a population that includes people with limited education. A general principle is to avoid long sentences and words with four or more syllables.

**Example of assessing readability:** Schilling and colleagues (2009) developed the Self-Management of Type 1 Diabetes in Adolescents (SMOD-A) scale. The scale's readability was assessed using the Flesch-Kincaid grade level score, which was found to be at the 5.9 grade level.

## Input from the Target Population

It is often productive to pretest the initial set of items with a sample of 10 to 20 people from the target population. These respondents can be asked some simple questions (e.g., Are there statements that confused you? Did you understand the meaning of each question? Were the directions clear?). *Cognitive questioning* is an excellent technique for discovering how others process the words and ideas presented to them in structured questions. (The Toolkit offers suggestions for cognitive questioning ☒.) Streiner and Norman (2008) describe several other techniques that can be used to detect ambiguities and language problems in a pretest.

**Example of cognitive questioning:** Hamilton and colleagues (2009) developed a measure of preferred coping strategies for older African American cancer survivors. Cognitive questioning methods were used with a small sample to assess how each question was understood.



**TIP:** When questioning pretest respondents about the clarity or meaning of the items, avoid using the word “item,” which is research jargon (e.g., do not say, “Are there *items* that confused you?”).

Additionally, it is a good idea to peruse the pretest answers to see if response patterns suggest the need for item revisions. For example, items with no variability (e.g., everyone agrees or disagrees) should be revised or omitted because they cannot contribute to

the scale's ability to discriminate among people with varying amounts of the underlying construct. Streiner and Norman (2008) warned that if the pretest fails to suggest any changes, this probably indicates a flaw in the pretest rather than problem-free items.

As an alternative or supplement to pretests, *focus groups* can also be used at this stage in scale development. Two or three groups can be convened to discuss whether, from the respondents' perspective, the items are understandable, linguistically and culturally appropriate, inoffensive, and relevant to the construct.

## External Review by Experts

External review of the revised items by a panel of experts should be undertaken to assess the scale's content validity. It is advisable to undertake two rounds of review, if feasible—the first to refine or weed out faulty items or to add new items to cover the domain adequately, and the second to formally assess the content validity of the items and scale. We discuss some procedures in such a two-step strategy, although the two steps are sometimes combined.

### Selecting and Recruiting the Experts

The panel of experts needs to include people with strong credentials with regard to the construct being measured. Criteria such as the following can be used in selecting substantive experts: clinical or personal experience, published papers in refereed journals, or an ongoing program of research on the topic. Experts should be knowledgeable about the key construct *and* the target population. In the first review, it is also desirable to include experts on scale construction.


In the initial phase of a two-part review, we advise having an expert panel of 8 to 12 members, with a good mix in terms of roles (e.g., clinicians, faculty, researchers) and disciplines. For example, for a scale designed to measure fear of dying in the elderly, the experts might include nurses, gerontologists, and psychiatrists. If the scale is intended for broad use, it might also be advantageous to recruit experts from various countries or areas of a country, because of possible regional variations in language.




The second panel for formally assessing the content validity of a more refined set of items should consist of 3 to 5 experts in the content area.


**Example of an expert panel:** Lin and colleagues (2008) developed and tested the Diabetes Self-Management Instrument in Taiwan. A panel of seven experts in diabetes and instrument development was assembled. The experts included three diabetes educators with doctorates, two physicians specializing in diabetes, and two nurse practitioners who worked in a diabetes clinic.

Experts are typically sent a packet of materials, including a strong cover letter, background information about the construct and target population, reviewer instructions, and a questionnaire soliciting their opinion (Grant & Davis, 1997). A critical component of the packet is a careful explanation of the conceptual underpinnings of the construct, including an explication of the various dimensions encompassed by the construct to be captured in subscales. The panel may also be given a brief overview of the literature, as well as a bibliography.

**TIP:** The Toolkit section of the *Resource Manual* includes a sample cover letter and other material relating to expert review, as Word documents that can be adapted.



Preliminary Expert Review: Content Validation of Items

The experts’ job is to evaluate individual items and the overall scale (and any subscales), using guidelines established by the scale developer. The first panel of experts is usually invited to rate each item along several dimensions. Among the dimensions often used are the following: clarity of wording, relevance of the item to the construct or to one of its dimensions, and appropriateness for the target population (e.g., developmental or cultural appropriateness). Experts could either be asked to make judgments dichotomously (e.g., ambiguous/clear) or along a continuum. As noted in the previous chapter, relevance is most often rated as follows: 1 = *not relevant*, 2 = *somewhat relevant*, 3 = *quite relevant*, 4 = *highly relevant*. Figure 15.1  shows a possible format for a content validation assessment of relevance.

The scale items shown below have been developed to measure one dimension of the construct of safe sexual behaviors among adolescents, namely **assertiveness**. Please read each item and score it for its relevance in representing this concept.

**Assertiveness** is defined as the use of verbal and interpersonal skills to negotiate protection during sexual activities.

Item	Relevance Rating			
	Not Relevant	Somewhat Relevant	Quite Relevant	Highly Relevant
1. I ask my partner about his/her sexual history before having intercourse.	1	2	3	4
2. I don't have sex without asking the person if he/she has been tested for HIV/AIDS.	1	2	3	4
3. When I am having sex with someone for the first time, I insist that we use a condom.	1	2	3	4
4. I don't let my partner talk me into having sex without knowing something about how risky it would be.	1	2	3	4

Please comment on any of these items, including possible revisions or substitutions, or your thoughts about why an item is not relevant to the concept of assertiveness. Please suggest any additional items you feel would improve the measurement of assertiveness relating to adolescents’ safe sexual behaviors.

**FIGURE 15.1**  Example of a portion of a content validation form.

The questionnaire usually asks for detailed comments about items judged to be unclear, not relevant, or not appropriate, such as how wording might be improved, or why the item is deemed to be not relevant. Another dimension that could be included for each item in a first phase evaluation concerns an overall recommendation—for example: retain the item exactly as worded, make major revisions to the item, make minor revisions to the item, and drop the item entirely.

In addition to evaluating each item, the initial expert panel should be asked to consider whether the items taken as a whole adequately cover the construct domain. The items on a scale constitute the operational definition of the construct, so it is important to assess whether the operational definition taps each dimension adequately. Experts should be asked for specific guidance on items or subdomains that should be added. For scales constructed within an IRT framework, the experts should also be asked whether the items as a whole span a continuum of difficulty (i.e., whether the underlying hierarchy is adequately captured).

If there is agreement among the experts, the next step is straightforward: Their opinion is used to guide decisions about retaining, revising, deleting, or adding items. When there is disagreement, however, it may require further investigation. Perhaps the experts did not understand the task, perhaps the conceptual definitions were ambiguous, and so on.

The typical formula for evaluating agreement among experts on individual items is the number agreeing, divided by the number of experts. When the dimension being rated is relevance, the standard method for computing a content validity index *at the item level* (I-CVI) is the number giving a rating of either 3 or 4 on the 4-point relevance scale, divided by the number of raters. For example, if five experts rated an item as 3 and one rated the item as 2, the I-CVI would be .83. Because of the risk of chance agreement when ratings are dichotomous—relevant versus not relevant—we recommend that I-CVIs should be .78 or higher (Polit, et al., 2007). This means that there must be 100% agreement among raters when there are 4 or fewer experts. When there are 5 to 8 experts, one rating of “not relevant” can be tolerated, and when

there are 9 or more experts, even more can disagree on relevance.

Items with lower-than-desired I-CVIs need careful scrutiny. It may be necessary to recontact the experts to better understand genuine differences of opinion or to strive for greater consensus. If there are legitimate disagreements among the experts on individual items (or if there is agreement about lack of relevance), the items should be revised or dropped.

### Content Validation of the Scale

In the second round of content validation, a smaller group of experts (3 to 5) can be used to evaluate the relevance of the revised set of items and to compute the scale content validity (S-CVI). Although it is possible to use a new group of experts, we recommend using a subset from the first panel because then information from the first round can be used to select the most qualified judges. With information from round 1, for example, you can perhaps identify experts who did not understand the task, who had a tendency to give high (or low) ratings, who were not as familiar with the construct as you thought, or who otherwise seemed biased. In other words, data from the first round can be analyzed with a view toward evaluating the performance of the experts, not just the items. This analysis might also require discussion with some of the experts to fully understand the reason for incongruent or anomalous ratings.


In terms of selecting experts based on their ratings in the first round, here are some suggestions. First, it may be imprudent to select experts who rated every item as “highly relevant” (or “not relevant”). Second, it would not be wise to invite back an expert who gave high ratings to items that were judged by most others to not be relevant, or vice versa. Third, the proportion of items judged relevant should be computed for all judges. For example, if an expert rated 8 out of 10 items as relevant, the proportion for that judge would be .80. The pattern across experts can be examined for “outliers.” If the average proportion across raters is, for example, .80, you might consider not inviting back for a second round experts whose average proportion was either very low (e.g., .50) or very high (e.g., 1.0). Qualitative feedback from an expert in round 1, in

the form of useful comments, might indicate both content capability and a commitment to the project. Finally, items known not to be relevant can be included in the first round to identify judges who rate irrelevant items as relevant and thus may not really be experts after all.

After ratings of relevance are obtained for a revised set of items, the S-CVI can be computed. There is more than one way to compute an S-CVI, as noted in Chapter 14 (Polit & Beck, 2006). We recommend the approach that averages across I-CVIs. On a 10-item scale, for example, if the I-CVIs for 5 items were .80 and the I-CVIs for the remaining 5 items were 1.00, then the S-CVI/Ave would be .90. An S-CVI/Ave of .90 or higher is desirable.

In summary, we recommend that for a scale to be judged as having excellent content validity, it would be composed of items that had I-CVIs of .78 or higher and an S-CVI (using the averaging approach) of .90 or higher. This requires strong items, outstanding experts, and clear instructions to the experts regarding the underlying constructs and the rating task.

---

 **TIP:** When you describe content validation in a report, be specific about your criteria for accepting items (i.e., the cutoff value for your I-CVIs) and the scale (the S-CVI). The report should indicate the range of obtained I-CVI values and the method used to compute the S-CVI.

---

## ADMINISTRATION TO A DEVELOPMENT SAMPLE

At this point, you will have whittled down and refined your items based on your own and others' careful scrutiny. The next step in scale development is to undertake a quantitative assessment of the items, which requires that they be administered to a fairly large development sample. As with content validation, this may involve a two-part process, with preliminary assessment occurring in the first phase and subsequent efforts to evaluate the scale's psychometric adequacy in the second.

Testing a new instrument is a full study in and of itself, and care must be taken to design the study to yield useful evidence about the scale's worth. Important steps include the development of a sampling plan and data collection strategy.

### Developing a Sampling Plan

The sample for testing the scale should be representative of the population for whom the scale has been devised, and should be large enough to support complex analyses. If it is not possible to administer the items to a random sample (as is typical), it is advantageous to recruit a sample from multiple sites—preferably in different areas—to enhance representativeness and to assess geographic variation in interpreting items. Other strategies to enhance representativeness should be sought, as well—for example, making sure that the sample includes older and younger respondents, men and women, people with varying educational and ethnic backgrounds, and so on, if these characteristics are relevant. You should also consider taking steps to ensure that the sample includes the right subsets of people for a “known groups” analysis.

How large is a “large” sample? There is neither consensus among experts nor hard-and-fast rules. Some suggest that 300 is an adequate number to support a factor analysis (Nunnally & Bernstein, 1994), while others offer guidance in terms of a ratio of items to respondents. Recommendations range from 3 or 4 people per item to 40 or 50 per item, with 10 per item being the number most often recommended. That means that if you have 20 items, your sample should probably be at least 200. Having a sufficiently large sample is essential to ensure stability in the covariation among the items.

### Developing a Data Collection Plan

Decisions have to be made concerning how to administer the instrument (e.g., by mailed or distributed questionnaires, over the Internet), and what to include in the instrument. In deciding on a mode of administration, you should choose an approach

that best approximates how the scale typically would be administered after it is finalized. Thought should also be given to administration setting. For example, if the scale is designed as a screening tool for hospitalized patients, then hospitals should be the setting for collecting the development data.

The instrument should include the scale items and basic demographic information. Thought should also be given to including other measures on the instrument—which would be essential if you do not plan to undertake a separate study to evaluate the scale's validity.

Various types of validation measures are possible to evaluate the facets of construct and criterion-related validity discussed in Chapter 14. For example, you might include a measure of constructs similar to, but distinct from, the target construct to evaluate discriminant validity. Measures of other constructs hypothesized to be correlated with the target construct should be included. If the data confirm a relationship predicted by theory or prior research, this would lend evidence to the new scale's validity. Finally, it may be useful to include measures to assess response biases, especially social desirability. Item correlations with a measure of social desirability could suggest potentially biased items. (More complex approaches to evaluating and addressing the effects of social desirability and “faking bad” biases are discussed in Streiner and Norman, Chapter 6). Brief social desirability scales have been developed (e.g., Reynolds, 1982; Strahan & Gerbasi, 1972).



**TIP:** In deciding on what other measures to include in the study, keep in mind that respondents' willingness to cooperate may decline as the instrument package gets longer.

## Preparing for Data Collection

As in all data collection efforts, care should be taken to make the instrument attractive, professional looking, and easy to understand. Friends, colleagues, mentors, or family members should be asked to evaluate the appearance of the instrument before it is reproduced.

Instructions for completing the instrument should be clear, and a readability assessment of the instructions is useful. There should be no ambiguity about what is expected of respondents. Guidance in understanding the end points of response options should be provided if points along the continuum are not explicitly labeled. The instructions should encourage candor. Sometimes, social desirability can be minimized by stating that there are no right or wrong answers. Anonymity also reduces social desirability bias, and is recommended—unless the scale needs to be administered twice to estimate test–retest reliability. Pett and colleagues (2003) offer useful suggestions for laying out an instrument and for developing instructions to respondents.

One other consideration is how to sequence the items in the instrument. At issue is something that is called a *proximity effect*, the tendency to be influenced in responding to an item by the response given to the previous item. This effect would tend to artificially inflate estimates of internal consistency. One approach to deal with this is the random ordering of items. An alternative, for scales designed to measure several related dimensions, is to systematically alternate items that are expected to be scored into different subscales.

**Example of item ordering:** Lange and Yellen (2009), who refined a Spanish version of a scale to measure satisfaction with nursing care, deliberately placed two positively worded items at the beginning of the instrument, because Lange's previous work suggested that negatively worded items at the beginning confused people. After the first two items, negative items were positioned throughout the scale at random.

## ANALYSIS OF SCALE DEVELOPMENT DATA

The analysis of data from multi-item scales is a topic about which entire books have been written. We provide only an overview here. We assume that readers of this section have basic familiarity with statistics. Those who need a refresher should consult Chapters 16 through 18.

## Basic Item Analysis

The performance of each item on the preliminary scale needs to be evaluated empirically. Within classical measurement theory, what is desired is an item that has a high correlation with the true score of the underlying construct. We cannot assess this directly, but if each item is a measure of that latent variable, then the items should correlate with one another.

The degree of **inter-item correlation** can be assessed by inspecting the correlation matrix of all the items. If there are items with substantial negative inter-item correlations, some should perhaps be reverse-scored (e.g.,  $\text{NEWITEM} = 8 - \text{OLDITEM}$ , for 7-point scales). Unless intentional, however, negative correlations are likely to reflect problems and may signal the desirability of removing some items. For items on the same subscale, inter-item correlations between .30 and .70 are often recommended (e.g., Ferketich, 1991), with correlations lower than .30 suggesting little congruence with the underlying construct and ones higher than .70 suggesting over-redundancy. However, the evaluation depends on the number of items in the scale. An average inter-item correlation of .57 is needed to achieve a coefficient alpha of .80 on a 3-item scale, but an average of only .29 is needed for a 10-item scale (DeVellis, 2003).

A next step is to compute preliminary total scale scores (or subscale scores) and then to calculate correlations between individual items and total scores on the scales they are intended to represent. If item scores do not correlate well with scale scores, it is probably measuring something else and will lower the reliability of the scale. There are two types of **item-scale correlations**, one in which the total score includes the item under consideration (*uncorrected*), and another in which the individual item is removed in calculating the total scale score. The latter (*corrected*) approach is preferable because the inclusion of the item on the scale inflates the correlation coefficients, and the inflation factor increases as the number of items on the scale decreases. The standard advice is to eliminate items whose item-scale correlation is less than .30 but some recommend a criterion as high as .50.

DeVellis (2003) also recommends looking at basic descriptive information for each item, as a double check. Items should have good variability—without it, they will not correlate with the total scale and will not fare well in a reliability analysis. Means for the items that are close to the center of the range of possible scores are also desirable (e.g., a mean near 4 on a 7-point scale). Items with means near one extreme or the other tend not to discriminate well among respondents.

Other item analysis techniques have been developed. Some scale developers compute item *p* levels or *difficulty levels*, which are indicators of how “difficult” each item is. For example, if 60 people agreed with an item and 40 disagreed with it, it could be said that the *p* level for the item was .60 because 60% found it “easy” to agree. Items in the mid-range of difficulty are most desirable. Another index that is too complex to explain here is the *discrimination index*, which examines the discriminative ability of each item. As mentioned earlier, item response theory has given rise to a number of excellent diagnostic tools for examining the performance of individual items. These and other item analysis techniques are described elsewhere (e.g., Gable & Wolfe, 1993; Nunnally & Bernstein, 1994; Waltz et al., 2010).

**Example of item analysis:** Heo and colleagues (2005) undertook several item analytic procedures with data from a sample of 638 patients in their evaluation of the Minnesota Living with Heart Failure Questionnaire. They computed item-total correlations, inter-item correlations, item *p* levels, and a discrimination index. As a result of these and other analyses, they recommended that 5 items be deleted.

## Exploratory Factor Analysis

A set of items is not necessarily a scale—the items form a scale only if they have a common underlying construct. **Factor analysis** disentangles complex interrelationships among items and identifies items that “go together” as unified concepts. This section deals with a type of factor analysis known as **exploratory factor analysis (EFA)**, which essentially



assumes no *a priori* hypotheses about dimensionality of a set of items. Another type—confirmatory factor analysis—uses more complex modeling and estimation procedures, as described later.

Suppose we developed 50 Likert-type items measuring women's attitudes toward menopause. We could form a scale by adding together scores from several individual items, but which items should be combined? Would it be reasonable to combine all 50 items? Probably not, because the 50 items are not all tapping the same thing—there are various *dimensions* to women's attitude toward menopause. One dimension may relate to aging and another to loss of reproductive ability. Other items may involve sexuality, and yet others may concern avoidance of monthly menstruation. These multiple dimensions to women's attitudes toward menopause should be captured on separate subscales. Women's attitude on one dimension may be independent of their attitude on another. Dimensions of a construct are usually identified during the conceptualization phase and when the items are being evaluated by experts. Preconceptions about dimensions, however, do not always “pan out” when tested against actual responses. Factor analysis offers an objective, empirical method of clarifying the underlying dimensionality of a large set of measures. Underlying dimensions thus identified are called **factors**, which are weighted combinations of items in the analysis.



**TIP:** Before undertaking an EFA, you should evaluate the *factorability* of your set of items. Procedures for a factorability assessment are described in Polit (2010).

### Factor Extraction

EFA involves two phases. The first phase (**factor extraction**) condenses items into a smaller number of factors and is used to identify the number of underlying dimensions. The goal is to extract clusters of highly interrelated items from a correlation matrix. There are various methods of performing the first step, each of which uses different criteria for assigning weights to items. A widely used factor extraction method is **principal components analysis**

(PCA) and another is **principal-axis factor analysis**.

The pros and cons of alternative approaches to factor extraction have been nicely summarized by Pett and colleagues (2003). Our discussion focuses mostly on PCA, although the two methods often (but not always) lead to the same conclusion about dimensionality.

Factor extraction yields an *unrotated factor matrix*, which contains coefficients or *weights* for all original items on each extracted factor. Each extracted factor is a weighted linear combination of all the original items. For example, with three items, a factor would be item 1 (times a weight) + item 2 (times a weight) + item 3 (times a weight). In the PCA method, weights for the first factor are computed such that the average squared weight is maximized, permitting a maximum amount of variance to be extracted by the first factor. The second factor, or linear weighted combination, is formed so that the highest possible amount of variance is extracted from what *remains* after the first factor has been taken into account. The factors thus represent independent sources of variation in the data matrix.


Factoring should continue until no further meaningful variance is left, so a criterion must be applied to decide when to stop extraction and move on to the next phase. There are several possible criteria, which makes factor analysis a semisubjective process. Several criteria can be described by illustrating information from a factor analysis. Table 15.1 presents fictitious values for eigenvalues, percentages of variance accounted for, and cumulative percentages of variance accounted for, for 10 factors. **Eigenvalues** are equal to the sum of the squared item weights for the factor. Many researchers establish as their cutoff point for factor extraction eigenvalues greater than 1.00. In our example, the first five factors meet this criterion. Some believe that the eigenvalue rule is too generous—that is, extracts too many factors (DeVellis, 2003). Another cutoff benchmark, called the *scree test*, is based on a principle of discontinuity: A sharp drop in the percentage of explained variance indicates the appropriate termination point. In Table 15.1, we might argue that there is considerable discontinuity

TABLE 15.1 Summary of Factor Extraction Results

FACTOR	EIGENVALUE	PERCENTAGE OF VARIANCE EXPLAINED	CUMULATIVE PERCENTAGE OF VARIANCE EXPLAINED
1	12.32	29.2	29.2
2	8.57	23.3	52.5
3	6.91	15.6	68.1
4	2.02	8.4	76.5
5	1.09	6.2	82.7
6	.98	5.8	88.5
7	.80	4.5	93.0
8	.62	3.1	96.1
9	.47	2.2	98.3
10	.25	1.7	100.0

between the third and fourth factors—that is, that three factors should be extracted. Another guideline concerns the amount of variance explained by the factors. Some advocate that the number of factors extracted should account for at least 60% of the total variance and that for any factor to be meaningful it must account for at least 5% of the variance. In our table, the first three factors account for 68.1% of the total variance; 6 factors contribute 5% or more to the total variance.

So, should we extract 3, 5, or 6 factors? One approach is to see whether there is any convergence among these guidelines. In our example, two of them (the scree test and total variance test) suggest three factors. Another approach is to determine whether any of the rules yields a number consistent with our original conceptualization about dimensionality. In our example, if we had designed the items to represent three theoretically meaningful subscales, we might consider three factors to be the right number because there is sufficient empirical support for that conclusion. Indeed, some have argued that restricting the factor solution to a prespecified number of factors that is consistent with the original conceptualization can yield important information regarding how much variance is accounted for by the factors.

 **TIP:** Polit (2010) provides a “walk-through” demonstration of how decisions are made in undertaking an exploratory factor analysis.

### Factor Rotation

The second phase of factor analysis—**factor rotation**—is performed on factors that have met extraction criteria, to make the factors more interpretable. The concept of rotation can be best explained graphically. Figure 15.2 shows two coordinate systems, marked by axes A1 and A2 and B1 and B2. The primary axes (A1 and A2) represent factors I and II, respectively, as defined *before* rotation. Points 1 through 6 represent six items in this two-dimensional space. The weights for each item can be determined in reference to these axes. For instance, before rotation, item 1 has a weight of .80 on factor I and .85 on factor II, and item 6 has a weight of  $-.45$  on factor I and .90 on factor II. Unrotated axes account for a maximum amount of variance but may not provide a structure with conceptual meaning. Interpretability is enhanced by rotating the axes so that clusters of items are distinctly associated with a factor. In the figure, B1 and B2 represent rotated factors. After rotation, items 1, 2, and 3 have large weights on factor I and

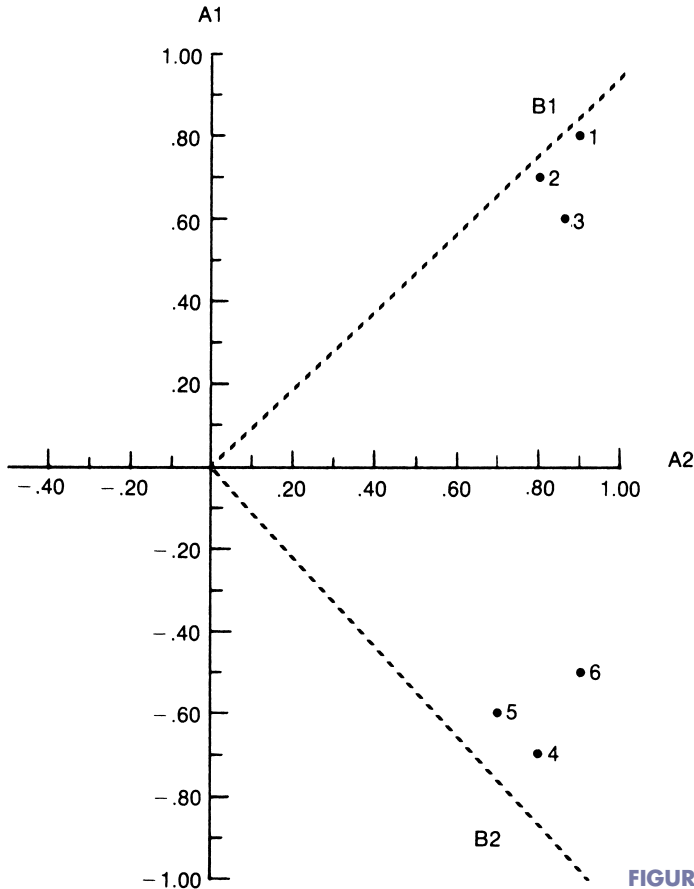


FIGURE 15.2 Illustration of factor rotation.

small weights on factor II, and the reverse is true for items 4, 5, and 6.

Researchers choose from two types of rotation. Figure 15.2 illustrates **orthogonal rotation**, in which factors are kept at right angles to one another. Orthogonal rotations maintain the independence of factors—that is, orthogonal factors are uncorrelated with one another. **Oblique rotations** permit rotated axes to depart from a 90-degree angle. In our figure, an oblique rotation would have put axis B1 between items 2 and 3 and axis B2 between items 5 and 6. This placement strengthens the clustering of items around an associated factor, but results in correlated factors. Some writers argue that orthogonal rotation leads to greater theoretical clarity; others claim that it is unrealistic. Advocates of oblique rotation point

out that if the concepts *are* correlated, then the analysis should reflect this fact. In developing a scale with multiple dimensions, we likely would expect the dimensions to be correlated, so oblique rotation might well be more theoretically meaningful. This can be assessed empirically: If an oblique rotation is specified, the correlation between factors is calculated. If the correlations are low (e.g., less than .15 or .20), an orthogonal rotation may be preferred because it yields a simpler model.

Researchers work with a **rotated factor matrix** in interpreting the factor analysis. As an example, the matrix in Table 15.2 shows information from a factor analysis of the School-Age Temperament Inventory (SATI) for 12 of the scale's 38 items (McClowry, et al., 2003). The entries under each

**TABLE 15.2** Factor Loadings: School-Age Temperament Inventory

ITEM	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
1. Does not complete homework <sup>a</sup>	.04	<b>.82</b>	-.01	.07
2. Is shy with adults he (she) doesn't know	.02	.00	<b>.78</b>	.00
3. Runs when entering or leaving	.16	.03	.00	<b>.79</b>
4. Is bashful when meeting new children	.09	.01	<b>.80</b>	-.01
5. Stays with homework until finished	.04	<b>.84</b>	.00	.06
6. Yells or snaps at others when angry	<b>.79<sup>b</sup></b>	.05	.04	.12
7. Runs or jumps when going down stairs	.18	.11	.05	<b>.74</b>
8. Is moody when corrected for misbehavior	<b>.75</b>	.10	.13	.02
9. Runs to where he (she) wants to go	.09	.07	-.10	<b>.77</b>
10. Responds intensely to disapproval	<b>.78</b>	.11	.06	.12
11. Has difficulty completing assignments <sup>a</sup>	.08	<b>.78</b>	.01	.04
12. Seems uncomfortable at someone's house	.11	.06	<b>.75</b>	.06

<sup>a</sup>Item was reverse-coded before factor analysis.

<sup>b</sup>Bolded entries represent high loadings on a factor and are used to name and interpret the factor.

Adapted from Table 1 of McCloskey, S. G., Halverson, C. F., & Sanson, A. (2003). A re-examination of the validity and reliability of the School-Age Temperament Inventory. *Nursing Research*, 52(3), p. 180.

factor are the weights, or **factor loadings**. For orthogonally rotated factors, factor loadings can range from -1.00 to +1.00 and can be interpreted like correlation coefficients—they express the correlation between items and factors. In this example, item 1 is highly correlated with Factor 2, .82. By examining factor loadings, we can find which items “belong” to a factor. For example, items 6, 8, and 10 have sizable loadings on factor 1. Loadings with an absolute value of .40 or higher often are used as cutoff values, but somewhat smaller values may be acceptable if it makes theoretical sense to do so. The underlying dimensionality of the items can then be interpreted. By inspecting the content of items 6, 8, and 10, we can search for a common theme that makes the items go together. The developers of the SATI called this first factor *Negative Reactivity*. Items 1, 5, and 11 have high loadings on Factor 2, which they named *Task Persistence*. Factor 3 and 4 are called *Approach/Withdrawal* and *Activity*, respectively. The naming of factors is a process of identifying underlying constructs—and

this naming often would have occurred during the conceptualization phase.

The results of the factor analysis can be used not only to identify the dimensionality of the construct, but also to make decisions about item retention and deletion. If items have low loadings on all factors, they may be good candidates for deletion (or revision, if you can detect wording problems that may have caused different respondents to infer different meaning from the item). Items with fairly high loadings on multiple factors may also be candidates for deletion. Items with marginal loadings (e.g., .34) but that had good content validity probably should be retained for the reliability analysis.

### Example of exploratory factor analysis:

Heaman and Gupton (2009) developed a scale called the Perception of Pregnancy Risk Questionnaire. The 9-item scale was tested with 199 women in the third trimester of pregnancy. Exploratory factor analysis resulted in a 2-factor solution: Risk for Baby (5 items with loadings ranging from .40 to .99) and Risk for Self (4 items with loadings from .51 to .92).


## Reliability Analysis

After a final set of items is selected based on the item analysis and factor analysis, a reliability analysis should be undertaken to calculate coefficient alpha. Alpha, it may be recalled, provides an estimate of the proportion of variance in the scale scores that is attributable to the true score and thus is a key indicator of the scale's quality.

Most computer programs for doing reliability analysis provide extensive information, including many item analysis diagnostics we described earlier. Especially important at this point in scale development is information about the value of coefficient alpha for the scale—and for a hypothetical scale with each individual item removed. If the overall alpha is extremely high, it may be prudent to eliminate redundancy by deleting items that do not make a sizeable contribution to alpha. (Sometimes removal of a faulty item actually *increases* alpha.) A modest reduction in reliability is sometimes worth the benefit of lowering respondent burden. Scale developers must consider the best trade-off between brevity and reliability.

One thing that should be kept in mind is that reliabilities tend to capitalize on chance factors in a sample of respondents and will often be lower in a new sample. Thus, you should aim for reliabilities a bit higher in the development sample than ones you would consider minimally acceptable so that if the alphas deteriorate they will still be adequate. This is especially true if the development sample is small.

---

 **TIP:** If you have the good fortune to have a very large sample, you should consider dividing the sample in half, running the factor analysis and reliability analysis with one subsample, and then rerunning them with the second as a cross-validation of factor structure and scale reliabilities.

---

## FINAL STEPS: SCALE REFINEMENT AND VALIDATION

In some scale development efforts, the bulk of work is over at this point. For example, if you developed a scale as part of a larger substantive project because

you were unable to identify a good measure of a key construct, you may be ready to pursue your substantive analyses. If, however, you are developing a scale for others to use, a few more steps remain.

## Revising the Scale

The analyses undertaken in the development study often suggest the need to revise or add items. For example, if subscale alpha coefficients are lower than .80 or so, consideration should be given to adding items for subsequent testing. In thinking about new items, a good strategy is to examine items that had high factor loadings. Such items presumably correlate most strongly with the latent variable and so may offer powerful clues for additional items.

There may be other reasons for adding new items. For example, if a confirmatory factor analysis is envisioned as part of a scale validation effort, there should be at least 4 items for each factor (subscale) because of technical problems with dimensions having three or fewer items.

Finally, you should carefully examine the content of the items remaining in your scale. Sometimes alphas are inflated by items that have similar wording, so it is wise to make decisions about retaining or removing items based not only on their contribution to alpha, but also on content validity considerations. It may prove worthwhile to re-examine the I-CVIs of each item in making final decisions.

## Scoring and Transforming the Scale

Scoring the scale is often easy with Likert-type items: Item scores are typically just added together (with reverse scoring of items, if appropriate) to form subscale scores, and subscale scores are sometimes added together to form total scale scores. Scoring in this manner should, however, be a conscious decision.

When individual items are simply added together, the implicit assumption is that all of the items are equally important indicators of the latent variable. If there are theoretical or empirical reasons for suspecting otherwise, a system of weighting the items (so that more important items are given more weight



in the total score) might be considered. For example, a scale to assess a person's risk of a disease or condition (e.g., risk of cardiovascular disease) might benefit from weighting some items (e.g., high blood pressure) more heavily than others. Weighting is sometimes accomplished empirically, for example, by using factor loadings from a PCA to weight items. Weighting is discussed more extensively in Streiner and Norman (2008), and Pett and colleagues (2003) provide detailed information about factor scores.

A related consideration is whether the scores should be *transformed*. If, for example, subscales have different numbers of items, the means will almost surely vary even if the average intensity is similar across dimensions—making it difficult to make comparisons across dimensions. For this reason, some scale developers deliberately try to construct scales that have an equal number of items per subscale. Another approach is to transform scores, most typically through the use of *standard scores* or *z scores* (see Streiner & Norman, 2008).

## Conducting a Validation Study

Scale developers ideally should take steps to gather new data about the worth of their instrument in a validation study. Those who are not able to undertake a second study should strive to undertake many of the activities described in this section with data from the original development sample. Designing a validation study entails much of the same issues (and advice) as designing a development study, in terms of sample composition, sample size, data collection strategies, and so on. Thus, we focus here on analyses undertaken in a validation study. Internal consistency reliability should be recomputed in the validation sample.

## Confirmatory Factor Analysis

**Confirmatory factor analysis** (CFA) is playing an increasingly important role in validation studies. CFA is preferable to EFA as an approach to construct validity because CFA is a hypothesis testing approach—testing the hypothesis that the items belong to specific factors, rather than having the dimensionality of a set of items emerge empirically, as in EFA.

CFA is a subset of an advanced class of statistical techniques known as **structural equation modeling** (SEM). CFA differs from EFA in a number of respects, many of which are quite technical. One concerns the estimation procedure. As we explain in Chapter 18, many statistical procedures used by nurse researchers employ *least-squares estimation*. In SEM, the most frequently used estimation procedure is *maximum likelihood estimation*. (Maximum likelihood estimators are ones that estimate the parameters most likely to have generated the observed measurements.) Least-squares procedures have several stringent assumptions that are generally untenable—for example, the assumption that variables are measured without error. SEM approaches can accommodate measurement error and avoid other restrictions as well.

CFA involves the testing of a **measurement model**, which stipulates the hypothesized relationships among underlying latent variables and the *manifest variables*—that is, the items. The measurement model is essentially a factor analytic model that seeks to confirm a hypothesized factor structure. Loadings on the factors (the *latent variables*) provide a method for evaluating relationships between observed variables (the items) and unobserved variables (the factors or dimensions of a construct).

We illustrate with a simplified example involving a scale designed to measure two aspects of fatigue: physical fatigue and mental fatigue. In the example shown in Figure 15.3, both types of fatigue are captured by five items each: items I1 to I5 for physical fatigue and items I6 to I10 for mental fatigue. According to the model, respondents' item responses are *caused by* their physical and mental fatigue (and thus the straight arrows indicating hypothesized causal paths) and are also affected by error ( $e_1$  through  $e_{10}$ ). Moreover, it is hypothesized that the error terms are correlated, as indicated by the curved lines connecting the errors. Correlated measurement errors on the items might arise as a result of the person's desire to "look good" or to acquiesce—factors that would systematically affect all item scores. The figure also shows that the two latent fatigue variables are hypothesized to be correlated.

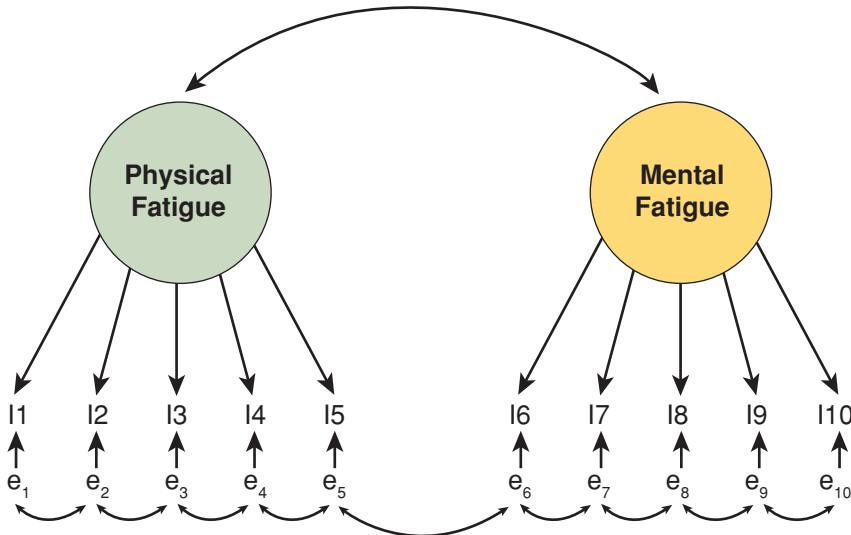


FIGURE 15.3 Example of a measurement model.

The hypothesized measurement model would be tested against actual data. The analysis would yield loadings of observed variables on the latent variables, the correlation between the two latent variables, and correlations among the error terms. The analysis would also indicate whether the overall model fit is good, based on a **goodness-of-fit statistic**. If the hypothesized model is *not* a good fit to the data, the measurement model could be respecified and retested.

CFA is a complex topic, and we have described only basic characteristics. Further reading on the topic is imperative for those wishing to pursue it (e.g., Brown, 2006; Harrington, 2008).

#### Example of confirmatory factor analysis:


Kalisch and colleagues (2010) undertook a psychometric assessment of the Nursing Teamwork Survey with a sample of over 1,700 nurses in acute care facilities. Exploratory factor analysis with a randomly selected half of the sample revealed five factors. Confirmatory factor analysis with the other half of the sample confirmed the factor structure.

#### Other Validation Activities

A validation effort would not be complete without undertaking additional activities designed to pro-

vide evidence of the scale's validity, such as ones described in Chapter 14. The assessment of criterion or construct validity primarily relies on correlational evidence. In criterion-related validity, scores on the new scale are correlated with an external criterion. In construct validity, scores on the scale can, for example, be correlated with measures of constructs hypothesized to be related to the target construct, or supplementary measures of the same construct (convergent validity), or measures of a closely related but distinguishable construct (divergent validity). Contrast validity using a known-groups approach requires selecting people with membership in groups expected to be different, on average, on the scale. It is desirable to produce as much validity evidence as possible.

If a CFA is not possible (perhaps because of lack of training in using it), it is nevertheless advisable to undertake a “confirmatory” factor analysis using the more traditional methods, such as PCA, with the validation sample. Comparisons between the original and new factor analyses can be made with respect to factor structure, loadings, variance explained, eigenvalues, and so on. In the new analysis, the number of factors to be extracted and rotated can be prespecified, since this is now the working hypothesis about the underlying dimensionality of the construct.

 **TIP:** Scale development and validation activities should be reported in the nursing literature so that others can benefit. An editorial in the journal *Research in Nursing & Health* provides guidance regarding information to include in an instrument development paper (Froman & Schmitt, 2003), and further guidance is offered by DeVon and colleagues (2007).

## Establishing Cutoff Points

Scales produce scores along a continuum, but there are constructs for which it is important to dichotomize scale scores. A familiar example is classroom and licensing examinations: There must be a score (cutoff point) that distinguishes those who pass and those who fail. Diagnostic and screening scales need to provide information about whether there is “case-ness” or not.

Various methods—both empirical and subjective—have been developed for establishing cutoff points on scales (Streiner & Norman, 2008). As described in Chapter 14, the method that has the most credibility is the construction of receiver operating characteristic (ROC) curves and its associated indicators. Data for undertaking such an analysis typically come from a validation study. Scale developers who intend to develop ROC curves need to select highly reliable criteria for dividing people into groups (e.g., those with and those without the condition being screened), and the criteria must be independent of participants’ responses on the scale.


## Establishing Norms

In some cases, it might be desirable to *standardize* a new scale and establish norms. This typically occurs if the expectation is that (a) the scale will be widely used, and used by people who will rely on solid comparative information to help them evaluate scores, and (b) average scale scores vary markedly by members of well-defined subpopulations. Norms are most commonly established for key demographic characteristics, such as age and gender.

Sampling is the most critical aspect of a standardization effort. The sample used to establish norms should be geographically dispersed (within the desired scope) and representative of the population for

whom the scale is intended. In most cases, this means using probability sampling. A large standardization sample is required so that subgroup values are stable.

After the scale is administered to a standardization sample, various descriptive statistics are computed. Norms are often expressed in terms of percentiles. For example, an adult male with a score of 72 on the scale might be at the 80th percentile, but a female with the same score might be at the 85th percentile. Guidelines for norming instruments have been developed by Waltz and her colleagues (2010) and Nunnally and Bernstein (1994).

 **TIP:** If you expect the scale to be used by others, you should develop a manual for its use. The manual should include the items; the underlying conceptual rationale; the process used to develop, refine, and evaluate items; instructions for using the scale, including scoring and interpretation; information about norms and cutoff points, if relevant; and information about the scale’s psychometric properties. Guidelines for preparing manuals are published in *Standards for Educational and Psychological Testing* (AERA, APA, & NCME Joint Committee, 1999). Scale developers should consider registering a copyright, even if they do not plan to publish the scale commercially.

## TRANSLATING SCALES INTO OTHER LANGUAGES

Scales are increasingly being used with people from various linguistic and cultural backgrounds. Developing equivalent scales in other languages requires nearly as much care and effort as developing an original scale. We provide a brief overview and offer suggestions for further reading on this important topic.

### Centered Versus Decentered Translations

Translation is often approached in a “centered” way, in which the scale is translated into another language, with no effect on the wording of the original instrument. In a *centered* (or asymmetric) *translation*, loyalty to the original scale items is maintained. Such translations typically occur after-the-fact, that is, after the original scale has been validated and used, and has been identified as a candidate for translation because it has desirable features.

By contrast, a *decentered translation* involves the possibility of modifications to items on the original scale. Such decentered (or symmetric) translations often reflect the goal of replacing culturally exclusive language with more universally understood language. This approach is often adopted when the developer knows in advance that the scale will be used in two languages, so translation activities are built into the scale development process.



**TIP:** When a translation is anticipated upfront, scale developers should consider the following as they are crafting items: (1) avoid metaphors, idioms, and colloquialisms; (2) use specific words rather than ones open to interpretation, such as “daily” rather than “frequently”; (3) avoid pronouns—repeat nouns if necessary to avoid ambiguity; (4) write in the present tense and avoid the subjunctive mode, such as “should”; and (5) use words with a Latin root if the target language is a Romance language such as Spanish or French (Hilton & Skrutkowski, 2002; Lange, 2002).

**Example of a decentered translation:** Coffman (2008) translated the Diabetes Self-Efficacy Scale into Spanish using a decentered translation approach.

## Conceptual Equivalence

The goal of a translation is to achieve equivalence between an original version of a scale and a translated version. *Equivalence*, however, is itself a complex concept. Over a dozen different types of equivalencies have been suggested, although only a few are given consideration in a typical translation (Beck, et al., 2003; Streiner & Norman, 2008).

A particularly important consideration early in a translation concerns *conceptual equivalence*. Do people in the two cultures view the construct in the same way? As an example, consider *obesity*. A person who is obese in some western cultures might not be considered obese elsewhere. A related question is, does the construct even have *meaning* in the other culture? For example, does the construct of *pleasure* or *enjoyment* have meaning in devastatingly poor societies in which daily survival is a struggle? (Note that conceptual equivalence is an important issue even among subcultures that speak the same language.) Thus, one of the first tasks in a translation effort is to ascertain whether

the meaning of the construct as defined in the scale reflects the meaning of the construct within the “target” culture. Experts knowledgeable about the culture in question are often consulted in this early step.

## Semantic Equivalence: Back Translations


*Semantic equivalence* is the extent to which each item’s *meaning* is the same in the target culture after translation as it was in the original. Literal translations are rarely satisfactory. The translation needs to preserve the underlying meaning of the original wording rather than the exact wording.

The most respected translation process for achieving semantic equivalence involves **back-translation** (Brislin, 1970), in which an instrument is translated from an original *source language* into a *target language*, and then translated back into the source language by translators who are unfamiliar with the original wording. The process typically involves several important steps.

## Selecting and Preparing Translators


The first step is to select translators. At a minimum, two translators are needed, but four or more working as a team is usually desirable. For example, for translations from English to Spanish, it is important to include native Spanish speakers from various regions because of regional linguistic and cultural variations (e.g., Mexico, Puerto Rico, Central America, etc.). Translations are typically done *into* the native tongue of the translator. So, for example, for an English-to-French translation, the items would first be translated by a native French speaker and the back translation would be translated by a native English speaker. Being bilingual is not a sufficient qualification for doing a translation; ideally, translators would have some professional training and experience and have first-hand familiarity with both cultures and be capable of understanding the conceptual underpinnings of the construct.

The scale developer needs to carefully explain the construct and the intent of each item to the translators. Translators should also be given guidance about expectations for their performance—for example, they need to be told what reading level to aim for, whether colloquialisms are discouraged, the importance of semantic equivalence, and so on.

 **TIP:** It is sometimes productive to have two independent translations, a procedure that was used by Wang and colleagues (2006) in their translation of the High School Questionnaire: Profile of Experiences into Chinese. One translation was done by a person who taught English in Taiwan, and the second version was done by a graduate student majoring in English.

### Undertaking an Iterative Process

The translation/back-translation process is often an iterative one, requiring multiple rounds of translation, review, and group discussion to arrive at consensus. It begins with the translation of items, followed by the back translation by translators blinded to the original wording. Then a comparison is made between the wording of original items and their back-translated counterparts to detect any possible alterations resulting from the translation. The theory is that if the original and back-translated versions are identical, the translated item is equivalent in meaning.

 **TIP:** Some back translators infer the original item wording (even when the translation is poor), rather than actually translating from the target language back to the source language, so it is advisable to instruct back translators to treat the target-language version as the original.

More often than not, the original and back-translated versions are *not* identical. If there are serious differences, it may be fruitful to have a group discussion among the translators, and then begin the process anew with a second group of translators after making changes to either the qualifications of the translators, item wording in the source language, or instructions to the translators (or some combination of these).

**Example of a Back-Translated Item:** Beck and colleagues (2003) provided the following example from the development of the Spanish version of the Postpartum Depression Screening Scale:

Original item: I was afraid that I would never be my normal self again

Translated item: temia no volver a ser otra vez la misma de antes

Back-translated item: I was afraid I was not going to be the same person as before

When the original and back-translated items are reasonably close, it is time to involve a committee to review what has transpired and to arrive at a consensus about the translated version of the scale items. The committee may be the team of translators but may also be three or more bilingual persons who have not participated in the translation process. Committee members are given complete information on each item—the original version, translated versions, and back-translated versions. They may also be given supplementary information, such as the desired reading level and the *actual* reading level of the various versions. Committee work may require several hours of discussion before consensus is reached.

The committee may conclude that a back-translated item is a better match to the translated item than the original wording. In such a case, if a decentering approach has been used, the wording of the original item can be changed to reflect a more universally understood construction.

### Testing the Translated Version

Translated scales need to be tested in a manner analogous to testing the original scale. Pretesting with a small sample from the target culture is important, and cognitive questioning is especially valuable with a translated instrument.

A good way to further evaluate semantic equivalence is to pretest both versions with a sample of bilingual people. Two forms of the instrument should be prepared (source language first on one, target language first on the other), with forms distributed randomly to the pretest sample. Responses on the two versions then can be compared, at both the scale and item level.

Finally, the translated scale should be submitted to a full psychometric evaluation with a large sample of respondents. These efforts not only provide evidence of the soundness of the translated scale, but also support inferences about equivalence. For example, if the internal consistency of the translated scale is substantially lower than the original, there is something wrong with the translation. Confirmatory factor analysis is




another strategy that is useful in facilitating conclusions about both the conceptual and semantic equivalence of the scales. Other construct validation procedures ideally would also be used with a sample from the new culture (e.g., known groups).

**Example of a scale translation study:** Pinar and colleagues (2009) undertook a translation of the Health-Promoting Lifestyle Profile II (HPLPII). The scale was translated from English into Turkish by three translators and back-translated by three independent translators. An expert panel of 10 health professionals reviewed the process. The instrument was pretested with 30 monolingual Turkish speakers. Then cultural equivalence was assessed by administering both the English and Turkish versions to 109 bilingual people. Psychometric evaluation of the translated scale's reliability and validity was undertaken with a sample of 920 people. Validation efforts (including both EFA and CFA) indicated good construct validity of the translated scale. Test-retest and internal consistency reliability were high.

## CRITIQUING SCALE DEVELOPMENT STUDIES

Articles on scale development appear regularly in many nursing journals. If you are planning to use a scale in a substantive study, you should carefully review the methods used to construct and validate the scale—and to translate it, if a translated version is under consideration. You should also evaluate whether the evidence regarding the scale's psychometric adequacy is sufficiently sound to merit its use. Remember that you run the risk of undermining the statistical conclusion validity of your study (i.e., of having insufficient power for testing your hypotheses) if you use a scale with weak reliability. And you can run the risk of poor construct validity in your study if your measures are not strong proxies for key constructs.

Box 15.1  provides guidelines for evaluating a research report on the development and validation of a scale.

### BOX 15.1 Guidelines for Critiquing Scale Development and Validation Reports



1. Does the report offer a clear definition of the construct? Does it provide sufficient context for the study through a summary of the literature and discussion of relevant theory? Is the population for whom the scale intended adequately described?
2. Does the report indicate how items were generated? Do the procedures seem sound? Is information provided about the reading level for the scale items?
3. Does the report describe content validation efforts, and was the description thorough? Is there evidence of good content validity?
4. Were appropriate efforts made to refine the scale (e.g., through pretests, item analysis)?
5. Was the development or validation sample of participants appropriate in terms of representativeness and size?
6. Was factor analysis used to examine or validate the scale's dimensionality? If yes, does the report offer evidence to support the factor structure and the naming of factors?
7. Were appropriate methods used to assess the scale's reliability? Were reliability estimates sufficiently high?
8. Were appropriate methods used to assess the scale's criterion or construct validity? Is the evidence about the scale's validity persuasive? What other validation methods would have strengthened inferences about the scale's worthiness?
9. Does the report provide information for scoring the scale and interpreting scale scores—for example, means and standard deviations, cutoff scores, norms?
10. If the study involves a translation, were appropriate procedures (e.g., back translation, a committee approach, validation efforts) used to ensure scale equivalency?

## RESEARCH EXAMPLE

**Studies:** Postpartum Depression Screening Scale: Development and psychometric testing (Beck & Gable, 2000); Further validation of the Postpartum Depression Screening Scale (Beck & Gable, 2001); Postpartum Depression Screening Scale: Spanish version (Beck & Gable, 2003).

**Background:** Beck studied postpartum depression (PPD) in a series of qualitative studies, using both a phenomenological approach (1992, 1996) and a grounded theory approach (1993). Based on her in-depth understanding of PPD, she began in the late 1990s to develop a scale that could be used to screen for PPD, the Postpartum Depression Screening Scale (PDSS).

**Statement of Purpose:** Beck and an expert psychometrician undertook methodologic studies to develop, refine, and validate a scale to screen women for postpartum depression, and to translate the scale into Spanish.

**Scale Development:** The PDSS is a Likert scale designed to tap seven dimensions, such as sleeping/eating disturbances and mental confusion. A 56-item pilot form of the PDSS was initially developed with 8 items per dimension, using a 5-point response scale. Beck's program of research on PPD and her knowledge of the literature were the basis for specifying the domain. Themes from Beck's qualitative research were used to develop 7 dimensions, and to craft the items to operationalize those dimensions. The reading level of the final PDSS was assessed to be at the third-grade level and the Flesch reading ease score was 92.7.

**Content Validity:** Content validity was enhanced by using direct quotes from the qualitative studies as items on the scale (e.g., "I felt like I was losing my mind"). The pilot form was subjected to two content validation procedures with a panel of five content experts. Feedback from these procedures led to some item revisions.

**Construct Validity:** The PDSS was administered to a sample of 525 new mothers in six states (Beck & Gable, 2000). Preliminary item analyses resulted in the deletion of several items, based on item-total correlations. The PDSS was finalized as a 35-item scale with seven subscales, each with 5 items. This version of the PDSS was subjected to confirmatory factor analyses, which involved a validation of Beck's hypotheses about how individual items mapped onto

underlying constructs, such as cognitive impairment. Item response theory was also used, and provided supporting evidence of the scale's construct validity. In a subsequent study, Beck and Gable (2001) administered the PDSS and two other depression scales to 150 new mothers and tested hypotheses about how scores on the PDSS would correlate with scores on other scales. The results indicated good convergent validity.

**Criterion-Related Validity:** In the second study, Beck and Gable correlated scores on the PDSS with an expert clinician's diagnosis of PPD for each woman. The validity coefficient was .70, which was higher than the correlations between the diagnosis and scores on other depression scales, indicating its superiority as a screening instrument.

**Internal Consistency Reliability:** In both studies, Beck and Gable evaluated the internal consistency reliability of the PDSS and its subscales. Subscale reliability was high, ranging from .83 to .94 in the first study and from .80 to .91 in the second study. Figure 15.4 shows a reliability analysis printout (from the Statistical Package for the Social Sciences, or SPSS, Version 17.0) for the five items on the Mental Confusion subscale from the first study. In Panel A, we see that the reliability for the 5-item subscale is high, .912. The first column of Panel B (Item Statistics) identifies subscale items by number: Item 11, Item 18, and so on. Item 11, for example, is the item "I felt like I was losing my mind." The item means and standard deviations for the 522 cases suggest a good amount of variability on each item. Panel C presents intercorrelations among the 5 items. The correlations are fairly high, ranging from .601 for item 25 with 53, to .814 for item 11 with 25. Panel D (Summary Item Statistics) presents various descriptive statistics about the items. In Panel E, the fourth column ("Corrected Item-Total Correlation") presents correlation coefficients for the relationship between women's score on an item and their score on the total subscale, after removing the item from the scale. Item 11 has a corrected item-total correlation of .799, which is very high; all five items have excellent correlations with the total subscale score. The final column shows what the internal consistency would be if an item were deleted. If Item 11 were removed from the subscale and only four items remained, the reliability coefficient would be .888—less than the reliability for all 5 items (.912). Deleting any of the items on the subscale would reduce its internal consistency, but only by a rather small amount.

**A Reliability Statistics**

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.912	.912	5

**B Item Statistics**

	Mean	Std. Deviation	N
Item11	2.36	1.424	522
Item 18	2.21	1.270	522
Item 25	2.21	1.374	522
Item 39	2.40	1.351	522
Item 53	2.28	1.349	522

**C Inter-Item Correlation Matrix**

	Item11	Item 18	Item 25	Item 39	Item 53
Item11	1.000	.654	.814	.646	.649
Item 18	.654	1.000	.603	.659	.751
Item 25	.814	.603	1.000	.652	.601
Item 39	.646	.659	.652	1.000	.724
Item 53	.649	.751	.601	.724	1.000

**D Summary Item Statistics**

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	2.292	2.205	2.399	.194	1.088	.008	5
Item Variances	1.835	1.612	2.029	.416	1.258	.023	5
Inter-Item Correlations	.675	.601	.814	.213	1.354	.006	5

**E Item-Total Statistics**

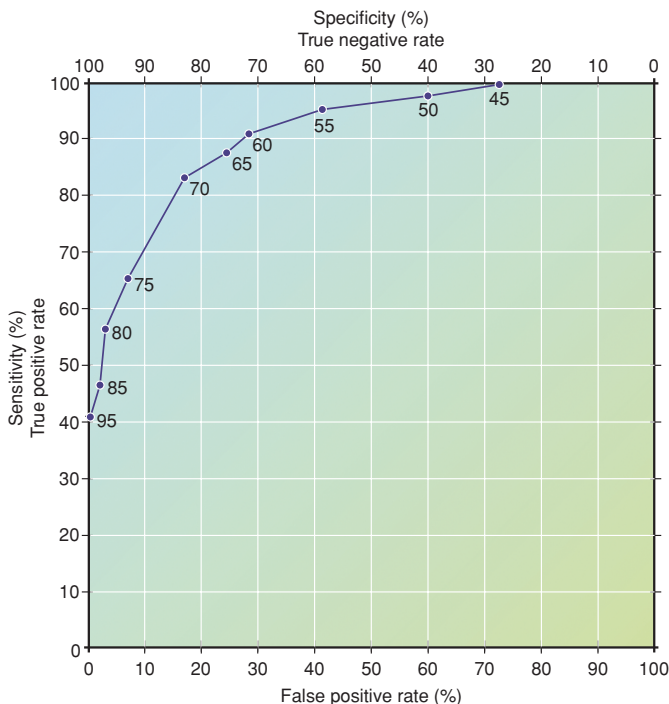
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Item11	9.09	21.371	.799	.715	.888
Item 18	9.24	23.006	.770	.623	.895
Item 25	9.25	22.097	.769	.691	.894
Item 39	9.06	22.290	.869	.610	.894
Item 53	9.18	22.176	.781	.666	.891

<sup>1</sup> In 2010, SPSS, Inc. was acquired by IBM and the software (starting with version 18.0) is now called PASW Statistics (or IBM SPSS).

**FIGURE 15.4** SPSS reliability analysis for the Mental Confusion subscale of the Postpartum Depression Screening Scale.

**Sensitivity and Specificity:** In the second validation study, ROC curves were constructed to examine the sensitivity and specificity of the PDSS at different cutoff points, using the expert diagnosis to establish PPD caseness. In the validation study, 46 of the 150 mothers had a diagnosis of major or minor depression. To illustrate the trade-offs the researchers made, the ROC curve (Figure 15.5) revealed that with a cutoff

score of 95 on the PDSS to screen in PPD cases, the sensitivity would be only .41, meaning that only 41% of the women actually diagnosed with PPD would be identified. A score of 95 has a specificity of 1.00, meaning that all cases *without* an actual PPD diagnosis would be accurately screened out. At the other extreme, a cutoff score of 45 would have 1.00 sensitivity but only .28 specificity (i.e., 72% false positive),



**FIGURE 15.5** Receiver operating characteristic (ROC) curve for Postpartum Depression Screening Scale (PDSS): Major or minor postpartum depression. Area = 0.91 (SD = 0.03). Used with permission from Beck, C. T., & Gable, R. K. (2001). Further validation of the Postpartum Depression Screening Scale. *Nursing Research*, 50, p. 161.

an unacceptable rate of overdiagnosis. Beck and Gable recommended a cutoff score of 60, which would accurately screen in 91% of true PPD cases, and would mistakenly screen in 28% who do not have PPD. Beck and Gable found that using this cutoff point would have correctly classified 85% of their sample. In their ROC analysis, the area under the curve was excellent, .91.

**Spanish Translation:** Beck collaborated with translation experts to develop a Spanish version of the PDSS. Eight bilingual translators from four backgrounds (Mexican, Puerto Rican, Cuban, and South American) translated and back-translated the items. The translators met as a committee to review each others' wordings and to arrive at a consensus. The English and Spanish versions were then administered, in random order, to a bilingual sample. Scores on the two versions correlated highly (e.g., .98 on the "Sleeping/Eating Disturbances" subscale). The alpha reliability was .95 for the total Spanish scale, and ranged from .76 to .90 for subscales. Confirmatory factor analysis yielded information that was judged to indicate an adequate fit with the hypothesized measurement model, and screening performance was found to be good (Beck & Gable, 2005).

## SUMMARY POINTS

- Scale development begins with a sound conceptualization of the construct (the **latent variable**) to be measured, including its dimensionality.
- After deciding on the type of scale to construct, items must be generated; common sources for items include existing instruments, the research literature, concept analyses, qualitative studies, focus groups, and clinical observations.
- In classical measurement theory, a **domain sampling model** is assumed; the basic notion is to sample a homogeneous set of items from a hypothetical universe of items.
- In generating items, a number of decisions must be made, including how many items to generate (typically a large number initially), what to use as the continuum for the response options, how many response options there should be, whether to include positive and negative item stems, how

- intensely worded the items should be, and what to do about references to time.
- Items should be inspected for clarity, length, inappropriate use of jargon, and good wording; the scale's **readability** should also be assessed.
  - External review of the preliminary pool of items should also be undertaken, including review by members of the target population (e.g., via a small pretest that could include **cognitive questioning**).
  - Content validity should be built into the scale through careful efforts to conceptualize the construct, and through content validation by a panel of experts—including the calculation of a quantitative index such as the CVI to summarize the experts' judgments of the relevance of scale items.
  - Once content validity has been established at a satisfactory level, the scale must be administered to a development sample—typically 300 or more respondents who are representative of the target population.
  - Data collected from the development sample are then analyzed using a number of techniques, including **item analysis** (e.g., a scrutiny of **inter-item correlations** and **item–scale correlations**); **exploratory factor analysis (EFA)**, and reliability analysis.
  - EFA is used to reduce a large set of variables into a smaller set of underlying dimensions, called **factors**. Mathematically, each factor is a linear combination of variables in a data matrix.
  - The first phase of EFA (**factor extraction**) identifies clusters of items that are strongly intercorrelated and is used to define the number of underlying dimensions in the items empirically; a widely used factor extraction method is **principal components analysis (PCA)**, but another important alternative is **principal axis factor analysis**.
  - The second phase of factor analysis involves **factor rotation**, which enhances the interpretability of the factors by aligning items more distinctly with a particular factor. Rotation can be either **orthogonal** (which maintains the independence of the factors) or **oblique** (which allows correlated factors). **Factor loadings** of the items on the rotated factor matrix are used to interpret and name the factors.
  - After the scale is finalized based on the preliminary analyses, a second study is often undertaken to validate the scale, using a variety of validation techniques; one widely used approach is **confirmatory factor analysis (CFA)**.
  - CFA involves tests of a **measurement model**, which stipulates the hypothesized relationship between latent variables and *manifest variables*. CFA is a subset of sophisticated statistical techniques called **structural equation modeling**.
  - Well-constructed scales with good psychometric properties are increasingly likely to be translated for use in other cultures. Translations are often **centered** on the original language, but a **decentered approach**, which would allow modifications to the wording of items in the original scale, may be preferred when it is anticipated during the development phase that the scale will be used in two languages.
  - Both *conceptual equivalence* and *semantic equivalence* are critical to the success of a translated effort. The “gold standard” for semantic equivalence involves **back-translation**, in which the scale is first translated from the *source language* into the *target language*, and then translated back to the source language by translators blind to the original wording. The next step typically involves a committee that convenes with the goal of arriving at a consensus translation.
  - The translated version is then tested in a manner similar to the original scale. Evidence for semantic equivalence and psychometric soundness comes from pretests of both original and translated scale with a sample of bilingual people, and comparison of reliabilities, factor structures, and other validity estimates between the two scales.

## STUDY ACTIVITIES

Chapter 15 of the *Resource Manual for Nursing Research: Generating and Assessing Evidence for Nursing Practice*, 9th edition, offers exercises and study suggestions for reinforcing concepts



presented in this chapter. In addition, the following study questions can be addressed:

1. Read a recent scale development paper and see how many of the steps discussed in this chapter were followed. Do omitted steps (if any) jeopardize the evidence about the scale's quality?
2. Use the critiquing guidelines in Box 15.1 to evaluate scale development procedures in the studies by Beck and Gable, referring to the original studies if possible.

## STUDIES CITED IN CHAPTER 15

- Beck, C. T. (1992). The lived experience of postpartum depression: A phenomenological study. *Nursing Research*, 41, 166–170.
- Beck, C. T. (1993). Teetering on the edge: A substantive theory of postpartum depression. *Nursing Research*, 42, 42–48.
- Beck, C. T. (1996). Postpartum depressed mothers interacting with their children. *Nursing Research*, 45, 98–104.
- Beck, C. T., & Gable, R. K. (2000). Postpartum Depression Screening Scale: Development and psychometric testing. *Nursing Research*, 49, 272–282.
- Beck, C. T., & Gable, R. K. (2001). Further validation of the Postpartum Depression Screening Scale. *Nursing Research*, 50, 155–164.
- Beck, C. T., & Gable, R. K. (2003). Postpartum Depression Screening Scale: Spanish version. *Nursing Research*, 52, 296–306.
- Beck, C. T., & Gable, R. K. (2005). Screening performance of the Postpartum Depression Screening Scale—Spanish version. *Journal of Transcultural Nursing*, 16, 331–338.
- Bu, X., & Wu, Y. (2008). Development and psychometric evaluation of the instrument: Attitude toward patient advocacy. *Research in Nursing & Health*, 31, 63–75.
- Coffman, M. (2008). Translation of a Diabetes Self-Efficacy Instrument: Assuring content and semantic equivalence. *The Journal of Theory Construction & Testing*, 12, 58–62.
- Ellenbecker, C. H., Byleckie, J., & Samia, L. (2008). Further psychometric testing of the Home Healthcare Nurse Job Satisfaction Scale. *Research in Nursing & Health*, 31, 152–164.
- Gómez, J., Hidalgo, M., & Tomás-Sábado, J. (2007). Using polytomous item response models to assess death anxiety. *Nursing Research*, 56, 89–96.
- Hamilton, J., Stewart, B., Crandell, J., & Lynn, M. (2009). Development of the Ways of Helping Questionnaire: A measure of preferred coping strategies for older African American cancer survivors. *Research in Nursing & Health*, 32, 243–259.
- Heaman, M., & Gupton, A. (2009). Psychometric testing of the Perception of Pregnancy Risk Questionnaire. *Research in Nursing & Health*, 32, 493–503.
- Heo, S., Moser, D. K., Riegel, B., Hall, L. A., & Christman, N. (2005). Testing the psychometric properties of the Minnesota Living with Heart Failure Questionnaire. *Nursing Research*, 54, 265–272.
- Jones, R., & Gulick, E. (2009). Reliability and validity of the Sexual Pressure Scale for Women—Revised. *Research in Nursing & Health*, 32, 71–85.
- Kalisch, B., Lee, H., & Salas, E. (2010). The development and testing of the nursing teamwork survey. *Nursing Research*, 59, 42–50.
- Lange, J. W., & Yellen, E. (2009). Measuring satisfaction with nursing care among hospitalized patients: Refinement of a Spanish version. *Research in Nursing & Health*, 32, 31–37.
- Lin, C. C., Anderson, R., Chang, C., Hagerty, B., & Loveland-Cherry, C. (2008). Development and testing of the Diabetes Self-Management Instrument. *Research in Nursing & Health*, 31, 370–380.
- McClowry, S. G., Halverson, C. F., & Sanson, A. (2003). A re-examination of the validity and reliability of the School-Age Temperament Inventory. *Nursing Research*, 52, 176–182.
- Pinar, R., Celik, R., & Bahcecik, N. (2009). Reliability and construct validity of the Health-Promoting Lifestyle Profile II in an adult Turkish population. *Nursing Research*, 58, 184–193.
- Schilling, L., Dixon, J., Knafl, K., Lynn, M., Murphy, K., Dumser, S., & Grey, M. (2009). A new self-report measure of self-management of type I diabetes for adolescents. *Nursing Research*, 58, 228–236.

***Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.***