CHAPTER HIGHLIGHTS

• Measurement: Ideal versus real
• Measurement scale types
• Identifying measures
  o The literature
  o Databases of tests: PsycTESTS and HaPI
  o Books of measures
  o Department resources and professors
  o Catalogues of measures and fees charged
• Qualities of measures: Reliability and validity
  o Reliability: Test–retest; parallel forms
  o Internal consistency: Split-half; Cronbach’s alpha
  o Calculating Cronbach’s alpha
  o Validity: Content, face, criterion-related, construct
• Other characteristics of measures
  o Length and difficulty
  o Instructions for scoring
  o Social desirability concerns; measures of social desirability
  o Qualifications for use
• Developing your own measure
• Why develop your own measure?
• Steps in creating your own measure
• Writing items (recommendations from Schwarz, 1999)
• Demographic items and social sensitivity
  • Gender, race, ethnicity, income
• Item formatting (number of anchors; odd, even; stem)
• Scale types and flexibility in answering research questions
• Question order
• Online survey software
  • Comparison of features across software programs
  • SurveyMonkey versus Qualtrics
  • Program features
  • Downloading online surveys into SPSS
  • Survey appearance and labeling

OVERVIEW

Finding appropriate scales to use in research can be a challenge. There are challenges both in identifying good scales in the literature and in physically obtaining them. An added pressure is that all measures (in their entirety) must be included in the human subjects institutional review board (IRB) proposal (see Chapter 4), which must be submitted and approved before data collection can begin. For students in a one-semester research methods class, this proposal is typically submitted by midsemester; thus, finding measures for research takes on added pressure.

This chapter covers both identifying good scales and physically obtaining those scales. The chapter explains the different types of scale measurement, the advantages of using measures other people have developed, and the criteria for evaluating the quality of a measure. These criteria include the measure’s validity (does it measure what it says it measures) and its reliability (does it produce the same results over multiple administrations; do all of the items assess the construct in question). Related to reliability, you will learn how to calculate the internal consistency (e.g., Cronbach’s alpha) for a scale. Cronbach’s alpha is a specific measure of internal consistency and reflects the degree to which each item in a scale correlates with each other item on the scale.

Validity: extent to which a measure assesses what it is claimed to measure.
Reliability: in statistics, refers to the ability of a measure to produce reproducible outcomes.
Cronbach’s alpha: statistical indicator of the internal consistency of a measure.
Values typically range from 0 to 1 (a negative value is possible but uncommon). The greater the internal consistency of a measure, the greater the extent to which each item is measuring the same construct. A construct is an idea or theory whose properties are inferred from some kind of measurement and are not directly observable. Examples of constructs include intelligence, anxiety, self-esteem, and leadership. When you are choosing a measure, the internal consistency is one of the most important characteristics.

The criteria for using a measure may also include the length of the instrument and the difficulty of the items. Cost, copyright, and the requisite level of training to use a measure are other concerns. The chapter will also illustrate the drawbacks to developing your own measures and the kinds of questions (typically demographic or background) that are reasonable to generate yourself. You will not only learn how to locate and obtain measures but also the importance of identifying whether there are subscales and obtaining complete scoring instructions. The chapter will discuss the kinds of scoring issues that sometimes arise (e.g., items that are reverse scored). A resource guide at the end of this book (Appendix D) provides a list of measures commonly used in social science and behavioral neuroscience research.

THE CONCEPT OF MEASUREMENT:
IDEAL VERSUS REAL

As we discussed in Chapter 2, when examining the Method section of articles, it is helpful to keep track of the measures used in a particular area of study. Then, as you begin your search for measures, you will have some familiarity with the landscape. Early on in the research process for a semester-length course, you will have to determine your variable(s) of interest, that is, what you are trying to measure (e.g., anxiety and perceived control), and what scales exist for these constructs.

To explain the causes of behavior, researchers need some way to both describe and measure that behavior. In the social and behavioral sciences, we often try to describe and quantify abstract concepts, such as anxiety or sense of community, with scales. True, we may do something far more direct, for example, measure the reaction time to respond to a stimulus on a computer screen. Even then reaction time may be a proxy for some more abstract concept such as speed of perceptual processing.

Any measure you use will fall short in some way. It may not fully capture the essence of the concept you want to measure, or it may include material unrelated to the concept. Consider a concept like managerial competence. Your goal is to come as close as possible to measuring what you and others think managerial competence is
(e.g., ability to work with people, delegate tasks, or solve problems), without including extraneous information in your measurement (e.g., attractiveness). You approach this task by identifying an existing measure of that idea, a measure that you find in the literature or one that you create yourself (more on that later). This measure of managerial competence is your actual instrument; you give this measure to your participants. The goal is to have maximum overlap between your theoretical concept, here managerial competence, and the actual measure you use, recognizing that the overlap will never be complete.

THE PURPOSE OF MEASURES

By now, you have identified a research question and have developed some preliminary hypotheses. How will you evaluate these hypotheses? What kind of data will you use? You want data that reflect the constructs (e.g., sense of community or managerial competence) central to your research question. As an example, perhaps you are interested in the relationship between the rate and complexity of vocabulary in managers’ speech and judgments of managers’ competence. You could figure out a way to measure the rate (e.g., number of words/minute) and complexity of vocabulary (e.g., an easy way would be to use the readability statistics in the word processing program). But how are you going to measure the judged competence?

Constructs become more difficult to measure the more abstract they are. When constructs are straightforward, such as rate of speech, the operational definition (as defined in Chapter 3) is easy to specify. When you define the variable by the operations or processes used to measure or quantify it, you have an operational definition. If we were measuring safe driving behavior on a college campus, for example, we might count the number of times drivers come to a complete stop at a particular stop sign. We have quantified the behavior (safe driving) by counting full stops at this designated stop sign. Thus, we have “put into operation” our measures that define safe driving. Some operational definitions seem relatively easy to specify, as this example suggests. Measuring the manager’s judged competence is more challenging.

REVISIT AND RESPOND

• What is an operational definition? Give an example. Why is it hard to capture in full the essence of an abstract theoretical concept with a measure? Why is it harder when you measure managerial competence than when you measure rate of speech?
MEASUREMENT SCALE TYPES

In selecting a measure to use, one consideration is the kind of scale type it is. Many researchers emphasize the role of quantifiable (i.e., measured) data, although in Chapter 6, we talk about qualitative approaches where statistical analyses are not the focus. There are four generally accepted types of measurement scales: nominal, ordinal, interval, and ratio (Stevens, 1946). Using a particular scale type (e.g., nominal) can restrict the kinds of research questions you can answer and the degree to which you can detect differences in outcomes of interest; thus, knowing about the properties of each scale type is important. Figure 5.1 presents a summary of the characteristics of the four scale types, described in more detail in the next sections.

Nominal Scale Measurement

Nominal measurement scales are scales in which the values reflect no inherent order. These scales are also called categorical data, to reflect the fact that the data are assigned to categories. An example of a nominal scale question would be, “Do you own a car?” to which the respondent would reply “yes” or “no.” We can count the number of participants in a given category (i.e., how many people said “yes” vs. how many said “no”), but there is no inherent order to the responses (“yes” is not inherently better or higher than “no”). Each answer simply represents a different category on the dimension of car ownership. Categorical data are often seen in research in which open-ended questions are asked. An open-ended question allows the respondent to answer in any way he or she wishes, that is, an open-ended response. In a closed-ended question, the respondent selects his or her response from a set of provided options, that is, a closed-ended response. Open-ended questions (e.g., “What do you like about your institution?”) produce responses that can be put in categories. In response to this question about one’s institution, participants may give answers about professors, courses, social life, food, housing, sports, career advising.

<table>
<thead>
<tr>
<th>Figure 5.1 Characteristics of Four Types of Measurement Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Can be counted</strong></td>
</tr>
<tr>
<td><strong>Can be ordered</strong></td>
</tr>
<tr>
<td><strong>Known distance between each value</strong></td>
</tr>
<tr>
<td><strong>Contains a true zero point</strong></td>
</tr>
</tbody>
</table>

Copyright ©2018 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.
location, library resources, or a host of other possibilities. Notice that one category is not more important or “higher” than another. These responses could be put into separate categories, and a frequency count could be made for each category. Some categories could also be combined; for example, a category could be created for “academics” into which we might put professors, courses, career advising, and library resources. There might be another category for quality of student life, which might include the other aspects (social life, food, housing, sports, and location). Again, there is no inherent ordering of these categories; they are equally valued.

In terms of statistics, categorical data are often analyzed in terms of frequencies (i.e., counts of the number of responses in each category) and their corresponding percentages. Beyond frequency analyses, **chi-square** analyses can be used with categorical data. Chi-square analyses involve dimensions that are nominal (categorical). Often analyses involve a $2 \times 2$ chi-square (meaning two categories for each dimension), but chi-square analyses are not restricted to just two categories on each dimension. For example, you might ask whether first year students differ from fourth year students in the number of mentions of social life, sports, and courses as sources of satisfaction (see Table 5.1). In this table, one dimension had three categories and the other dimension had two categories ($3 \times 2$). Imagine that you had 500 first year students and 500 fourth year students provide answers to that original open-ended question about what people liked about their institution.

The responses do not need to add up to 500 because participants could mention as many items as they wished and not everyone would mention these particular items. Chi-square analyses are about proportionality, not about the absolute numbers. As you look at these numbers, you can see that there seem to be different proportionalities, especially for courses (fourth year students mention this more often than do first year students) and sports (again, fourth year students mention this more often than do first year students). First year students list social life more often than do fourth year students, but the proportional difference is not as great as in the other categories. If you do the chi-square analysis on these data, there is a significant chi-square $\chi^2(2, N = 1,764) = 282.92$, $p < .001$, reflecting the fact that the proportions across categories differ. There are online

<table>
<thead>
<tr>
<th>TABLE 5.1</th>
<th>Comparison of First- and Fourth-Year College Students’ Satisfaction With Social Life, Sports, and Courses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Social Life</td>
</tr>
<tr>
<td>First Year</td>
<td>436</td>
</tr>
<tr>
<td>Fourth Year</td>
<td>336</td>
</tr>
</tbody>
</table>
calculators to do calculations such as this (e.g., http://www.socscistatistics.com/tests/chisquare2/default2.aspx) without needing to enter the data into a software statistical package such as SPSS Statistics®.

Nominal data are used in what is known as nonparametric statistics. A major difference between parametric and nonparametric statistics is that nonparametric statistics are viewed as less powerful than are parametric statistics, in that larger sample sizes are needed to achieve the same degree of power that you have with a parametric test (Howell, 2013).

**Ordinal Scale Measurement**

In ordinal data, there is an “ordering,” but there is no knowledge of the distance of one choice in that order from another. What if you were given a list of types of transportation to take in the city: taxi, car, Uber®, bicycle, train, subway, and bus. You were asked to rank order your preference for these transportation modes (1–7) with 1 being first choice to seventh being last choice. Once you had completed that ranking, we have your preferences, but we have no idea how closely preferred these choices are. You might really like car, Uber, and taxi (which you ranked 1, 2, and 3), but you might have very low preference for choices 4–7. We have more information than we do with nominal data (the ordering of your choices) but not all that much more. Among the common statistics used with ranked data are Spearman’s rho and Kendall’s tau (Howell, 2013). In Spearman’s rho, you are calculating the correlation between sets of ranks; in Kendall’s tau, the focus is on the inversions in the ranks (e.g., one participant ranks three items: 1—car, 2—Uber, 3—taxi, and another ranks them 1—car, 2—taxi, and 3—Uber). There is an inversion in the second and third ranks for the second participant relative to the first. In both measures (rho and tau), the essence is still a degree of agreement between the participants.

**Interval Scale Measurement**

Interval scales are the most widely used scale type in the social and behavioral sciences. Interval scale data are also sometimes called continuous data. In interval scales, the anchors are assumed to be equally spaced, such that in a 5-point scale, the distance between any two anchors is the same (e.g., 1 to 2 is the same distance as 4 to 5). Interval scales are used in parametric statistics, such as analysis of variance. As mentioned earlier, parametric statistics have some advantages over nonparametric statistics because of their greater power.

One commonly used type of interval scale in the social and behavioral sciences is the Likert scale. There’s some disagreement about the pronunciation, but it should be pronounced Lick-urt—like licking an ice cream cone (http://core.ecu.edu/psyc/wuenschk/StatHelp/Likert.htm: this link is entertaining reading in that experts talk about what makes something a Likert scale). The Likert scale is a scale in which there are typically five response options or anchors that represent degrees of agreement-disagreement to a
statement: strongly agree, agree, neither agree/disagree, disagree, and strongly disagree. Each point on a rating scale, not just the extremes, is called an anchor. Most researchers treat the Likert scale as interval, but it should be noted that some treat it as ordinal.

Ratio Scale Measurement

A ratio scale has one element missing in an interval scale: a true 0 point. In a ratio scale, 0 represents the absence of a quantity (such as having 0 correct answers on a memory test; other examples include measurements of weight and height; the Kelvin scale for temperature is also a ratio scale). In the social and behavioral sciences, we typically see ratio scales used in performance tasks, such as recognition, recall, or problem-solving tasks. In these instances, performance may be represented as the number of correct answers, which would have the possibility of 0 (none correct).

REVISIT AND RESPOND

- Describe the four different scale types (nominal, ordinal, interval, ratio). Explain the frequent use of interval scale data in the social and behavioral sciences.

SENSITIVITY OF A SCALE AND ANCHOR VALUES

When people talk about scale sensitivity, they usually refer to the idea that the scale is sufficiently fine-grained to detect differences that exist. Figure 5.2 presents an example.

Consider again our earlier example of asking participants to indicate what they like about their institution. If we listed a series of options (courses, sports, food, location) and asked people to use a rating scale with three anchors (1 = not at all; 2 = somewhat; 3 = a lot), we might find little variability in the responses. With just three anchors, there is little room to express a range of opinion. Alternatively, with seven anchors (the second example), a greater range of opinion could be expressed. We would therefore refer to the second option as having greater sensitivity than the first option. When you are selecting measures to use in research, the sensitivity of the measure is one criterion to consider.
As a general principle, measures that have 5–7 anchors are preferred to those with a smaller or larger number of anchors. With an even number of items (e.g., 4), you will have no neutral point, and having 3 items takes us back to the problem of lack of sensitivity. With 8 anchors, again there is no neutral point, and with 9 anchors, respondents may not be able to make meaningful distinctions with so many choices (Thomas, 2004).

**Psychological Meaning of Scale Anchors**

An interval scale may incorporate a 0 point, but it is a relative, not an absolute, value. The use and placement of an anchor of 0 in an interval scale also carries with it a psychological meaning that should be considered. In Figure 5.3 illustrating two different 7-point rating scales, the first example has no 0 and the second includes 0 as its lowest scale value.

These two rating scales are arithmetically identical, but would users necessarily interpret them the same way?

**THE PROCESS OF IDENTIFYING MEASURES: THE LITERATURE**

With the knowledge of the different scale types (nominal, ordinal, interval, ratio), you are ready to search for a measure. The general search strategies here can be in any domain of social and behavioral science. When you began reading the empirical literature in your topic area, as a way to generate a specific research question, you probably looked at more than 20 articles in that area. As discussed in Chapter 2, those articles are important resources because it makes sense to concentrate first on existing measures, that is, measures other researchers have used. When we talk about measures, we are typically referring to dependent variables (see Chapter 3). If we were to use the word *stimuli*, we would typically be referring to independent variables (see Chapter 3).

Not all disciplines in the social and behavioral sciences use the same words for the section that contains information about the measures. In the discipline of psychology, this section of a journal is called the Method (note that it is singular; as introduced in Chapter 2), and it has three subsections: Participants, Materials (sometimes called Apparatus or Measures or Instruments), and Procedure. The Participants are the “who”
The Research Experience

of your study; the Materials are the “with what” of your study; the Procedure is the “how” of your study. In other social and behavioral sciences, the terminology is similar, although not identical.

**The Measures/Materials/Instruments Section**

The Measures/Instruments section is the first place to look to locate a measure for your study. In that section, the author is supposed to provide the name of the measure used, a citation to that source (that is, where the author found that measure), the number of items, a sample item (including the anchors), and the internal consistency or reliability of that measure (or other aspects of what are called the *psychometric properties* of the instrument). Psychometric properties are quantifiable aspects of a measure that indicate its statistical quality. Other information (such as how to score the scale) may or may not be provided. Pay particular attention to measures that are commonly used because that might give you some idea of the endorsement of that scale by researchers. If a scale was only used in one article and that article is not recent (i.e., not within the last 10 years), you might be wise to keep looking.

Given the page restrictions for journal articles, you are unlikely to find all of the items for a measure (unless the measure is very short) in the Measures/Instruments section. You may not even find all of the basic information mentioned in the previous paragraph (e.g., sample items). Authors do not always provide the Cronbach’s alpha for a given measure or another indication of the scale’s reliability. Without that information, it is hard to determine the quality of the measure.

What is your next step? When describing the measure, the author provided a citation for that measure in the Measures/Instruments section. Your next step is to locate that citation in the References section of the article, and then obtain THAT article (using the tree backward techniques discussed in Chapter 2). You keep repeating that step (that is, backtracking) until one of two things happens at the point when no further sources are listed to examine: You have the full scale or you don’t. Through this process of backtracking, you will typically find the article where the scale was introduced and validated. The items for the scale are usually in that article. If not, there may be an author note in the article to contact the corresponding author to obtain the scale. In that case, e-mail the corresponding author to obtain the items. If that person has retired or is otherwise unlocatable, a good strategy is to write an author who has recently used the scale and ask whether that person can send you the items.

**DATABASES OF TESTS (PsycTESTS AND HaPl)**

Over the last few years, online databases of measures have become available. The most well known of these is **PsycTESTS®**. A second useful source is the **Health and**
Psychosocial Instruments (HaPI®) database. PsycTESTS, which appeared in 2011, is a database provided through the American Psychological Association (APA) that now contains almost 30,000 records, with coverage beginning in 1910 (to the present). If you look on line at the APA site for this database (http://www.apa.org/pubs/databases/psyctests/), it gives you the current count of records available; this count is updated monthly. Typically the actual test or at least sample test items are available, as well as information about the psychometric properties of the test (if those have been published). The database focuses on tests not commercially available, which is an important consideration for researchers (i.e., cost).

This database is very useful. It is possible to use a keyword to search for a test that measures a particular construct. As an example, if we use the keyword “self esteem,” it yields 860 results (May 26, 2016). You can sort the results by relevance, oldest, and newest. The default presentation is by relevance. If you took a look at the first measure listed and clicked on the full text symbol, you would see the screenshot in Figure 5.4:

Note that the first page gives you the name of the authors, the number of items (6), and the response format of the items (a Likert scale). You also see the citation for the source of this scale and a section labeled Permissions. Pay attention to the permissions to see the requirements for use of this scale. In this instance, you may use it for noncommercial research and educational purposes without contacting the authors. The next page gives you the specific items. Note also that we have seen no mention of the psychometric properties of the scale. But if we go back to the initial listing for this record after we have clicked on the hyperlink for this item (“Self Esteem Measure”), we come to a page called the detailed record. On that page is a description, which in this case includes the following at the end of the paragraph:

The Self-Esteem Measure consists of six items (e.g., “You have a lot of good qualities”) analogous to some of the items used in Rosenberg’s (1989) scale. Each item is rated on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). Cronbach’s alpha for this measure was .845. (PsycTESTS Database Record (c) 2015 APA, all rights reserved)

The description gives you the Cronbach’s alpha. There is no statement about reverse scoring. If there were reverse-scored items, this is where you would find out. Usually in cases where there is reverse scoring, some items need to be reverse-scored and others do not. A well-known example is Morris Rosenberg’s Self-Esteem Scale (1965); of the 10 items, 5 need to be reverse scored. What this means is that some items are phrased in a way that endorses the construct, in this case, possessing self-esteem, (e.g., high self-esteem; “On the whole I am satisfied with myself”), whereas other statements are phrased in a way that opposes the construct (i.e., low self-esteem; “At times I think I am no good at all”). If the...
anchors range from 1 to 4, with a higher number reflecting higher self-esteem, you would need to reverse the numbering (i.e., 4 to 1) for the reverse-scored items. Then, all items are “going in the same direction”; that is, that for all items a high score means the same thing (i.e., endorsement of high self-esteem).

What is the downside to using PsycTESTS? Consider what just happened in Figure 5.4 (Williams & McCarthy, 2014). That scale came first, by relevance, from our keyword search. The original Rosenberg Self-Esteem Scale (RSES; 1965), not modified or
translated, is the seventieth entry on this list. PsycTESTS is definitely convenient, and if you can’t find a relevant scale in the Measures section of the empirical articles you consulted, PsycTESTS is useful. PsycTESTS is also helpful in seeing whether all items for a given test you want are provided, if they were not available in the literature. The drawback to consulting PsycTESTS before consulting the literature is that the most widely used and reliable scale to measure a given construct may not be listed first. Start with the literature.

HaPI (Health and Psychsocial Instruments) is produced by the Behavioral Measurement Database Services (BMDS). A website describing HaPI [https://www.ebscohost.com/academic/health-and-psychosocial-instruments-hapi] states that “over 80 unique instruments” are offered within almost 175,000 records (see also http://bmdshapi.com). The records are from journals in both the health and psychosocial sciences. Let’s do the same keyword search in HaPI as we did in PsycTESTS: self esteem. If we use that keyword in our search, the result is 4,301 records (May 26, 2016). But there is redundancy in those records. The number of unique instruments offered in the database is far different than the total number of records. You can sort by relevance, author, source, date newest, and date oldest. Relevance is the default. The first record that appears is the Rosenberg Self-Esteem Scale (German Version) by Rosenberg [translated by Brauhardt, Rudolph, and Hilbert (2014)]. There is no information about the number of items, item format, or psychometric properties of the scale, although there is a reference to the original Rosenberg (1965) book that introduced and validated the scale. HaPI is useful if you want to see the articles in which a particular scale has appeared, as well as to locate the original source of the scale, but it is not directly useful in providing detailed information about the scale.

BOOKS OF MEASURES

Other useful bibliographic resources (again, after you have consulted the literature) are books of measures. There are books that provide measures and books that critique measures. Offerings include Encyclopedia of Psychological Assessment (2 volumes; Fernandez-Ballesteros, 2003), Measures for Clinical Practice and Research: A Sourcebook (5th ed.; Corcoran & Fischer, 2013), Tests Critiques Compendium: Reviews of Major Tests From the Test Critiques Series (Keyser & Sweetland, 1987), and Directory of Unpublished Experimental Mental Measures (Goldman & Mitchell; 9 volumes, 2008 is the most recent). The sourcebook by Kevin Corcoran and Joel Fischer is particularly useful, especially for clinical research. Volume 1 covers couples, families, and children. Volume 2 focuses on adults. In these volumes, the measures are provided as well as the following information for each scale: purpose, authors, description, norms, scoring, reliability, validity, primary reference, and availability. Regarding availability, the usual statement is that the measure may be copied from the volume; if not, the author’s contact information (or other avenues for obtaining the article) is provided.
DEPARTMENT RESOURCES AND PROFESSORS

Another source of measures is nearby: your department and its professors. See what measures and equipment your professors have on-hand and/or recommend. Your department may have a file of frequently used scales; these may be available to students with the appropriate training. Typically clinical measures, e.g., the Wexler Adult Intelligence Scale (WAIS) or projective techniques such as the Rorschach, require specific advanced training (e.g., master's or doctoral level; see also later section in this chapter on Qualifications for Use). In addition, the appendices of honors, master's, and Ph.D. theses in your department will contain complete measures. These are excellent resources.

CATALOGUES OF MEASURES AND FEES CHARGED

Almost without exception, measures offered by test publishing companies cost money. Examples of such companies are Psychological Assessment Resources (PAR), Inc. (http://www4.parinc.com), which offers assessments in such areas as achievement, intelligence, neuropsychology, personality, and career development. Consulting Psychologist’s Press (https://www.cpp.com/products/index.aspx) covers areas in career and organizational development. Among the most well-known of its offerings is the Myers-Briggs® Type Indicator, which is often used in career counseling offices to provide an assessment of personality and help people identify their areas of interest and “types.” EdITS (http://www.edits.net) offers services in career counseling and assessment. Although many of the instruments offered through EdITS focus on career counseling, there are some mood assessments, e.g., the Multiple Affect Adjective Checklist-Revised (MAACL-R®) that are useful for a range of research topics. Western Psychological Services (http://www.wpspublish.com/app/) offers assessment instruments in clinical, industrial/organizational, and neuropsychology, among other areas. The Psychological Corporation (http://www.pearsonclinical.com) focuses on clinical assessment (e.g., the MMPI®). Finally, Mind Garden (http://www.mindgarden.com) has a range of instruments covering leadership, employee burnout, gender roles, and environments. Of the companies offering tests for a fee, Mind Garden might have instruments most likely to be used by student researchers. Among their offerings are the University Residence Environment Scale®, which assesses the social climate of residential groups of students; the Bem Sex Role Inventory®, which is widely used to assess gender role; and the State-Trait Anxiety Inventory for Adults®, which is widely used to assess both state (at the moment) and trait (generalized) anxiety.
What these companies have in common is that they charge fees, although some provide discounts to researchers. Although some tests offered by these companies are available only through these companies (many of these tests are clinical instruments that require advanced training in clinical psychology), there are scales available through other routes without charge. As an example, if we take the Bem Sex Role Inventory (BSRI) and check PsycTESTS, we find the test there (all the items) and the Permissions statement says, “May use for Research/Teaching,” and no fee is listed. On Mind Garden, the manual for the BSRI is $50; a Remote Online Survey License is $2/participant, with a minimum purchase of 50 (thus, $100). Before assuming that you have to pay for an instrument, a recommended course of action is to check the literature first and then PsycTESTS.

Finally, a word is in order about the availability of tests that you find “online.” Doing an online search for a measure (e.g., searching for “BSRI”) may lead you to a copy of the test. Should you use this measure? There are at least two considerations. First, there is an ethical issue. If the test is one for which particular permissions and/or fees are required, then using a test you found online without paying the fees and having the qualifications for use may be considered unethical. Second, without knowing the items on the original test and their anchors, you can’t assume that the test you found is complete and correct.

**REVISIT AND RESPOND**

- Where should you start your search for a measure? How do PsycTESTS and HaPI differ? What do you learn from reading about the permissions for a scale? What kinds of tests typically have qualifications for use? If a fee is charged for a scale, what other avenues might you pursue before paying the fee?

**QUALITIES OF MEASURES: RELIABILITY AND VALIDITY**

Among the most important considerations for selecting a scale are its reliability and validity. There are various types of reliability, but the core issue is the consistency of the measure. Consistency can be measured over time (test–retest reliability) and can also refer to the degree to which each of the items of the scale is measuring the same construct (called internal consistency, as introduced in Chapter 2, often assessed as Cronbach’s alpha). Validity refers to whether the measure assesses what it claims to assess (e.g., does a measure of stress actually assess what people commonly understand stress to be?). We will concentrate on reliability because it is easier for researchers to evaluate this aspect of a measure, although considerations of validity are important as well. We will talk further
about one form of reliability, inter-rater reliability, when we talk about behavioral observation and content analyses in Chapter 6. Here we will concentrate on the reliability of psychological tests or measures.

**Test–Retest Reliability**

Test–retest reliability is a measure of the degree to which a test administered at one time ($T_1$) correlates with that measure at a second time ($T_2$). Typically, the length of time between administrations is fairly long, which will reduce the likelihood that participants will remember the questions (and their answers). The typical statistical measure of test–retest reliability is Pearson’s $r$. An example would be students who took the same test (e.g., a self-esteem scale) at two points in time.

**Parallel Forms Reliability (also called Alternate Forms Reliability)**

Many people take the SAT® twice to improve their scores. If so, they take a parallel form of the test the second time. Test developers such as the Educational Testing Service are aware that taking the same test twice could artificially inflate students’ scores, merely as a function of familiarity with the items. For that reason, test developers continuously create new items (which are included as part of an existing test but do not contribute to a test taker’s score). By including such trial items, test developers are assessing the degree to which performance on the old and new items correlate. Test developers need to be sure that the new items are of equivalent difficulty to create a truly parallel form of the test. Therefore, a researcher who is concerned that multiple administrations of a test might affect a participant’s scores [think back to Campbell and Stanley’s (1963) “Testing” threat to internal validity;, see Chapter 3] might look for an instrument with parallel forms reliability (also called alternate forms reliability).

**Measures of Internal Consistency: Split-Half Reliability and Cronbach’s Alpha**

Internal consistency is an important kind of reliability for researchers. Internal consistency reflects the degree to which each item in a scale is measuring the same construct (i.e., the degree to which each item correlates with every other item on the scale). There are two common measures of internal consistency, Cronbach’s alpha and split-half reliability. Cronbach’s alpha is more widely used, but it is worth explaining split-half reliability.

The name “split-half” suggests what happens to calculate this kind of reliability. A test is randomly divided into two halves, and participants take both halves, with the items intermingled, at the same time. Then, the two halves can be separated and a correlation
can be conducted between the scores on the halves. This approach to assessing the internal consistency of a measure is often used if you want to change the length of a test (lengthen or shorten it). In the Measures/Instruments section of a paper, you may see this kind of reliability referred to as Spearman–Brown’s split-half reliability coefficient.

The most widely reported measure of internal consistency is Cronbach’s alpha (α). You want the alpha for a measure to be high (.8 or greater; Nunnally, 1978) because the kind of variability you want should come from differences produced by exposure to different conditions, not from variability introduced in scale scores when “poor” test items don’t correlate well with other items.

Researchers often accept as a “given” the cutoffs or standards for selecting measures that other researchers attribute to the original source (Lance, Butts, & Michels, 2006). In the case of alpha, the “given” has been .70 or higher, which researchers frequently cite from Jum Nunnally (1978). Charles Lance et al. (2006) revisited Nunnally’s (1978) comments and showed that Nunnally said something quite different. First, Nunnally said that there are different standards for different research situations and that the .70 value is only appropriate for “early stages of research on predictor tests or hypothesized measures of a construct” (p. 245). If you are using an established scale for basic research, the .8 level may be acceptable, but Nunnally went on to say that in applied settings where important decisions are being made on the basis of someone’s test score, even a value of .9 would be “the minimum that should be tolerated, and a reliability of .95 should be considered a desirable standard” (p. 246). The value of .70 is a long way from .95!

In the Measures/Instruments section of an empirical paper, look for some statement about the internal consistency of a measure and abide by the standards that Nunnally (1978) suggested. You may not always be able to find an appropriate measure that reaches .8 or higher, but that should be a goal. Some commonly used measures have a high Cronbach’s alpha. For example, RSES has a reported alpha of .88 (PsycTESTS) and the Frost Indecisiveness Scale, which measures compulsive indecision, has a reported alpha of .9 (PsycTESTS).

THE IMPORTANCE OF COMPUTING YOUR OWN CRONBACH’S ALPHA

As a researcher, the Cronbach’s alpha value is one of the important criteria you will use in selecting a measure, but you should also conduct a Cronbach’s alpha on your own data and report that in your Results section. You cannot assume that the alpha reported in the literature will be the same for your sample; it may even be higher than the level
reported in the literature. It may also be lower, which could be cause for concern. When you calculate the Cronbach's alpha, for example, using SPSS, you will be able to see what the scale value would be if each item on the scale were deleted (see Figure 5.8), and you may be able to improve the Cronbach's alpha.

In SPSS, the steps for calculating a Cronbach’s alpha are:

1. Under Analyze from the Menu bar, select “Scale” and choose “Reliability Analysis.” An “action box” for Items will open.

2. Move over all of the items you want for the reliability analysis from the variable list to this box (see Figure 5.5).

3. Click open the “Statistics” option and click “Item,” “Scale,” and “Scale if item deleted” under “Descriptives.” Then click “Continue” (see Figure 5.6). In the main analysis box, the default model is Alpha so you don’t need to change that. You are ready to click “OK.”

In the example here, 10 items from the 39-item Padua Inventory (a scale to measure obsessive-compulsiveness) were used to calculate an alpha from a dataset from Ann
FIGURE 5.6  Screenshot of Reliability Analysis: Statistics Options in SPSS

Devlin (2008). We see that the alpha based on these 10 items (Padua 30–Padua 39) is .822 (see Figure 5.7). We also see in Figure 5.8 (labeled “Item-Total Statistics”) that the alpha would IMPROVE if we deleted Padua 30. The alpha would be .849 without Padua 30 (see Figure 5.8). What you would do in this instance is go back in to SPSS, take out Padua 30 from your reliability analysis, and rerun the reliability analysis. You would then see whether any other items could be deleted to improve the alpha (in this case, the answer was no).

FIGURE 5.7  Reliability Statistics

---

1The alpha for all 39 items in this dataset is .937; for all 39 items, the alpha cannot be improved by deleting any items. In fact, deleting any item would lower that alpha.
156 The Research Experience

Content validity: validity of a measure that focuses on its representativeness from the domain of interest (e.g., a spelling test for fifth graders composed of words selected from books read by fifth graders).

Face validity: type of validity in which the measures subjectively appear to assess what you claim (e.g., a measure of leadership that asks about decisiveness).

Criterion-related validity: degree to which test scores predict the behavior of interest; two types (predictive and concurrent).

Convergent validity: demonstration of agreement between measures hypothesized to be theoretically related; contrasted with divergent validity.

Discriminant validity: situation in which measures hypothesized to be theoretically unrelated are, in fact, unrelated.

FIGURE 5.8 Item-Total Statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean if Item Deleted</th>
<th>Variance if Item Deleted</th>
<th>Corrected Item-Total Correlation</th>
<th>Cronbach’s Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>padua30</td>
<td>1.8016</td>
<td>11.663</td>
<td>.164</td>
<td>.849</td>
</tr>
<tr>
<td>padua31</td>
<td>1.8925</td>
<td>10.641</td>
<td>.424</td>
<td>.817</td>
</tr>
<tr>
<td>padua32</td>
<td>2.0165</td>
<td>10.314</td>
<td>.668</td>
<td>.790</td>
</tr>
<tr>
<td>padua33</td>
<td>2.0578</td>
<td>11.740</td>
<td>.400</td>
<td>.817</td>
</tr>
<tr>
<td>padua34</td>
<td>1.8098</td>
<td>9.608</td>
<td>.728</td>
<td>.780</td>
</tr>
<tr>
<td>padua35</td>
<td>1.9834</td>
<td>10.115</td>
<td>.699</td>
<td>.786</td>
</tr>
<tr>
<td>padua36</td>
<td>1.7189</td>
<td>8.814</td>
<td>.669</td>
<td>.789</td>
</tr>
<tr>
<td>padua37</td>
<td>2.0330</td>
<td>11.222</td>
<td>.564</td>
<td>.805</td>
</tr>
<tr>
<td>padua38</td>
<td>1.9669</td>
<td>10.792</td>
<td>.584</td>
<td>.800</td>
</tr>
<tr>
<td>padua39</td>
<td>2.1322</td>
<td>12.264</td>
<td>.631</td>
<td>.818</td>
</tr>
</tbody>
</table>

REVISIT AND RESPOND

- Define internal consistency in your own words. Is there one set Cronbach’s alpha value for all research purposes? If not, what circumstances require the highest alpha? List the steps you would take to run a reliability analysis on SPSS.

QUALITIES OF MEASURES: VALIDITY

The validity of a scale is its ability to measure what it claims it measures. There are different kinds of validity; some involve numerical calculations, whereas others involve the presentation of a reasoned argument. Here we will cover four categories of validity: content, face, criterion-related (predictive and concurrent), and construct, which includes convergent and discriminant validity.

Content Validity

Students are well acquainted with the essence of content validity even if they have never used that term. Imagine that you are taking a course in personality theory and are having a test on Monday. As you think about studying for the exam, you want the professor to include material that you actually covered in the course (from the text, from lectures), not obscure material in the reading that you didn’t cover. Content validity is the idea that
the information (here on your test) adequately samples the domain of interest (material covered thus far in your personality course). You can't numerically show that a test has content validity; rather, you present an argument that the test was constructed in a way that it adequately sampled (i.e., included items from) the domain of interest.

Try This Now

If you were going to construct a test of fifth-grade spelling words with content validity, how would you go about constructing the test?

Face Validity

Face validity refers to the idea that a test “looks like” what it claims to assess. The “looks like” part can refer to the content of the written items, or it may literally refer to how the visual items look (or how the auditory items sound). If you told participants they were going to take a test to assess their ability to identify well-known classical music, participants would expect to hear classical music. If you didn’t play classical music, the test would lack face validity.

Why is this important for you as a researcher? If you tell participants they are going to participate in a research project about “X” (e.g., perceived suitability for working in various industries, including technology), and then they fill out measures that have to do with “Y” (e.g., sexism; using the Modern Sexism Scale, because you are really interested in gender discrimination in the tech industry), those measures lack face validity for the problem you identified. The measures don’t seem to have anything to do with what you said the study was about. In this situation, that mismatch may increase the demand characteristics in the study (refer to Chapter 3).

Criterion-Related Validity: Predictive and Concurrent

Criterion-related validity answers the question of whether a test predicts an outcome (criterion) of interest. This kind of validity involves a numerical calculation, a correlation, between the test score and the outcome (criterion) score. The difference between the predictive form and the concurrent form is the timing. In predictive validity, you are claiming that a measure predicts (forecasts) some outcome of interest that will occur in the future. Thus, the test occurs at one point in time and the outcome at a second point in time. In concurrent validity, the correlation between the test and the outcome occurs at the same point in time. An example will illustrate the difference. Most students have taken the SAT or ACT® tests, which are often required to apply to college.
These SAT or ACT scores are supposed to predict first year grade-point average (GPA), and there is even some evidence that the scores correlate well with overall GPA at the end of the fourth year of college (Schmitt et al., 2009). In other words, these tests are claimed to have predictive validity in that students take the tests in high school, yet the tests predict the GPA at the end of the first year of college. How do researchers know these tests have predictive validity? Because they have conducted studies in which there is variability in the test scores (here, SAT or ACT). That is, the researchers have the data for people who have scores that range from the low to the high end of performance, and then they correlate these scores with the first year GPAs of those students. Students with a wide range of such test scores are admitted to college (they have other qualities that influence admission), and this variability provides the opportunity to conduct such predictive validity studies.

In the case of concurrent validity, the test is taken and the outcome (e.g., performance) is measured at the same point in time.

**Try This Now**

Before reading further, can you think of some drawbacks to concurrent validity, in contrast to predictive validity?

Concurrent validity studies are often used in employment situations where you want to see if a new screening test for applicants predicts job performance. You would thus give the test you were trying out to current employees (incumbents) to see whether the test matched (i.e., correlated with) their current success as employees. A problem with this approach is that there is probably a restriction of range with these current employees. The difference between the scores of the high and low performers may not be that large—the score range is restricted (otherwise you probably would have already fired them!). In addition, you would have to assume that the time these incumbents had already spent on the job would not affect performance on your employment screening test. Thus, you would want a test that was not affected by on-the-job experience. Such a test would be better if it measured some reasonably stable indicator, like intelligence. You can see there are quite a few challenges in using measures that are validated concurrently; they are typically viewed as less convincing than are measures that are validated through the predictive approach.

An intriguing example of predictive validity made the national news in 1996 when it was mentioned on *The Tonight Show* and other media outlets. In this instance, an applicant for a position in the New London, Connecticut, Police Department was turned
away because he scored “too high” on one of the screening measures, the Wunderlic Personnel test. You can imagine all of the banter around this outcome, such as “Man too smart to be a cop.” The police department limited interviews to those who scored from 20 to 27 on the test, and the applicant had a score of 33. Why would the guidelines suggest that someone with a lower score would be a better addition to the department than someone with a higher score? The answer comes from the predictive validity studies conducted by the company that produces the test. These studies showed that applicants with the higher scores for police jobs often quit early on after becoming bored with the work. Training these applicants is expensive; thus, municipalities have a rationale for sticking to the predictive cutoffs. The rejected applicant sued but lost his appeal against the city, first in a lower court and then in the 2nd U.S. Circuit Court of Appeals. The court stated that the city’s policy was a rational way to select candidates (but also perhaps unwise!) (Larrañeta, 2000).

**Construct Validity**

**Construct validity** deals with whether the measure of interest actually assesses what it claims to assess. In a sense this is the essence of validity. How do you demonstrate construct validity? The approach involves creating a logical argument followed by a series of studies. You need to develop what Lee Cronbach and Paul Meehl (1955) called a *nomological network*. Such a network shows a series of lawful relationships, that is, the relationships among observable variables that your theory predicts. You need to situate the construct within a theoretical framework, specify a series of empirical studies that would demonstrate the relationships identified in the theoretical framework, and then conduct these studies. You might go about this by showing that the construct correlates with other measures with which you expect it to be similar to (this is called *convergent validity*) and that it does not correlate highly with measures with which it is hypothesized to be dissimilar (this is called *discriminant validity*). Cronbach and Meehl outlined the process involving this nomological network in an important paper titled “Construct Validity in Psychological Tests” (1955). Cronbach and Meehl didn’t offer a diagram of a nomological network in their 1955 paper, but Figure 5.9 shows a hypothetical nomological network. In this example, we want to develop a new (short) test of anxiety. The network shows that we expect particular relationships (positive correlations) between scores on our new test and existing measures of anxiety like the Taylor Manifest Anxiety Scale and the State-Trait Anxiety Inventory (STAI). Furthermore, we hypothesize correlations between scores on our scale and measures of electrodermal activity and heart rate variability (which can detect anxiety). We would not expect our new anxiety measure to correlate highly with a measure of leadership or intelligence, given that anxiety is distinct from these constructs.
REVISIT AND RESPOND

- Give an example of how you would establish content validity. What does it mean to say a measure has face validity? Why is a measure that has demonstrated predictive validity more powerful than a measure with concurrent validity? In your own words, explain what a nomological network is.

LENGTH AND DIFFICULTY OF MEASURES

When selecting measures, the internal consistency will be among the most important criteria you consider. Practical concerns, including the length of the measure and the difficulty of the items, are important as well. When we talked about threats to internal validity (see Chapter 3), one factor on the list from Donald Campbell and Julian Stanley (1963) was Maturation, which could include fatigue. The number of scales
you give participants, and the number of items on each scale, should be considered in selecting measures. Your participants are volunteers. How many measures do you plan to use? What is the length of each? Widely used measures vary considerably in length, from the 10-item Rosenberg Self-Esteem Scale (1965) to the NEO-PI-3® (McCrae, Costa, & Martin, 2005), a widely used measure of personality with 240 items and a validity item. The NEO-PI-3 takes about 40 minutes to administer and is typically used by professionals in counseling and education, but it is also used in research.

There is no hard-and-fast rule for the number of items you can expect participants to complete, but if filling out your materials is taking upward of 45 minutes, you might consider eliminating one or more of your measures. How do you know whether your battery of measures takes that long? Pilot testing (see Chapter 3). A set of materials that is lengthy may threaten the internal validity of your study (see Chapter 3).

A related concern is the reading level of the items. Some measures may require a level of reading proficiency that your participants do not possess, although this drawback is unlikely if your participants are college students. Even so, you should be attuned to the possibility that some words will not be understood. For example, one item on the Profile of Mood States is “peeved,” a word that not all students may know. Furthermore, you may have participants for whom English is not a first language or individuals with limited education. In the case of the NEO-PI-3, mentioned earlier, the authors reported that some adolescents had difficulty with several items (the test was originally designed for adults). In fact, Robert McCrae et al. (2005) reported that 30 items were replaced because the items were not understood by some adolescent respondents (at least 2% had such difficulty). An example of an old item and the one replacing it are “I am easy-going and lackadaisical” (old item) and “I'm not very ambitious” (replacement) (McCrae et al., 2005, p. 270). The authors reported that the new readability level was grade level 5.3 on the Flesch–Kincaid scale. The Flesch–Kincaid reading ease and grade-level scales tell you how difficult it is to understand a passage in English. You can test out this function for your own writing using Microsoft Word® software (this is available under Tools/Spelling and Grammar/Options/Grammar/Show readability statistics).

INSTRUCTIONS FOR SCORING

After you obtain a measure, you need to know how to score the measure. In some measures, as many as a third of the items may not be included as part of the score (e.g., Bem Sex Role Inventory; Bem, 1974). Many measures include items that are reverse scored (e.g., RSES, 1965; see discussion earlier in the chapter). Reverse-scored items are discussed again in the section on Data Entry in Chapter 10.

Flesch–Kincaid scale: provides statistics for the ease of reading of a given passage; available through word processing programs.
 NAMES OF MEASURES AND SOCIAL DESIRABILITY CONCERNS

The name of a measure is generally meant to communicate its content. For example, the RSES (1965) is a measure of self-esteem.

Try This Now

What problems can you see with telling participants that they are going to take a test that measures self-esteem even if you don’t use the full name of the scale?

Scale names or labels may communicate information about the measures that set up unwanted expectancies, that is, demand characteristics. A demand characteristic involves some component of the research process that unintentionally influences the participants’ responses (see Chapter 3). You may not want participants to know precisely what you are assessing, a kind of passive deception (see Chapter 4), because it may change the nature of their responses. Often researchers are concerned that respondents will answer in a manner that is more socially desirable (as defined in Chapter 1), which occurs when participants modify their responses to present themselves more favorably. Changing the name of the measure on the survey may eliminate this kind of demand characteristic; another approach is to provide a label without any content (e.g., Scale 1), or no label at all. A measure of managers’ competence might become “Managers and Their Work.”

Measures of Social Desirability

Most people want to be perceived favorably. We may modify our answers to present ourselves more favorably (e.g., smarter or less biased) than in truth we are. Researchers who administer measures on personal qualities and social issues where social desirability may play a role (e.g., sexual behavior or racism) typically include a measure of social desirability. Such measures allow researchers to assess the degree to which respondents typically answer in a favorable manner. Then, the total score on this measure is used as a control, called a covariate, in the analyses. In the case of a social desirability measure, what you are essentially saying (statistically) is that once you have controlled for people’s socially desirable responses you can see the “true” relationship between the IV and DV. Using a covariate helps us reduce error variance. Common measures of social desirability

Covariate: variable that can affect the relationship between the dependent variable and the independent variables being assessed.
are the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984, 1991) and the Marlowe–Crowne Social Desirability Scale (Crowne & Marlowe, 1960).

Another direction researchers are taking to assess socially desirable responding is using reaction time tests, most notably, the Implicit Association Test (IAT; Greenwald et al., 2003), which was mentioned in Chapter 1 in the context of dealing with biases when investigating socially sensitive issues. The underlying premise of the IAT is that responding takes time; if some associations come more easily (rapidly) to mind, such as the association of “young people” with “good” rather than of “old people” with “good,” the response to the first pairing will be faster. A website has been set up called Project Implicit (https://implicit.harvard.edu/implicit/research/) for researchers interested in using the IAT. You can contact the research team for help if you are interested in using the IAT in your work. Researchers have used the IAT to tap into implicit attitudes regarding a wide range of social concerns.

**QUALIFICATIONS FOR USE**

As already indicated, some measures require specific training and advanced degrees for use. Most of these measures are for clinical and neuropsychological testing, and most of these are only commercially available. If you look at the NEO-PI-3 in the PAR catalogue (http://www4.parinc.com), it lists a qualification level, S or B. Those with qualification level S can purchase products listed as A (no qualifications required) and as S (“a degree, certificate, or license to practice in a health care profession or occupation . . . ; plus appropriate training and experience in the ethical administration, scoring, and interpretation of clinical behavioral assessment instruments”). Those with qualification level B can purchase products at levels A, S, and B. For level B, the bar is a four-year degree, a relevant major and additional coursework in measurement theory or related topics, or a license or certificate in the use of psychological tests from an appropriate agency (http://www4.parinc.com/Supp/Qualifications.aspx).

If you want to read more about the ethical issues surrounding test use and professional standards, consider the work of Lorraine Eyde (e.g., Eyde, Robertson, & Krug, 2010).

**REVISIT AND RESPOND**

- What is socially desirable responding? Name two approaches to dealing with this threat to internal validity.
DEVELOPING YOUR OWN INSTRUMENT

The advantages of using existing measures far outweigh the disadvantages. Primary among the positive aspects is that you can select a scale with known reliability and validity. It takes considerable skill and time to develop a scale and to demonstrate its reliability and validity. Given this level of effort, why would researchers develop their own measures? (a) No adequate measure exists to assess a particular construct or to improve significantly on an existing measure (e.g., creating a shorter version or validating the measure in another language); (b) the existing measure costs money; and (c) professional qualifications are required. Here we will concentrate on the fundamental reason for creating a new measure—that no current measure adequately addresses the area of interest.

Developing a New Measure

On occasion, students develop research questions for which they can find no preexisting measure. New forms of social media are developing so rapidly that scales may not exist to assess their effect. For example, a recent student project dealt with Tinder, a program to facilitate social interaction between users who express mutual interest. You “swipe right” if you are endorsing someone whose picture and information appear. At the time the student developed the project, there were few online dating measures specifically targeting the Tinder app, and the student developed three items to supplement a question asking the degree to which someone was likely to “swipe right.”

Advice for Writing Items Yourself

Researchers typically write their own demographic or background items assessing participants’ characteristics such as age, race, and gender. Among the best guidelines available about writing items is Norbert Schwarz’s (1999) article titled “Self-Reports: How the Questions Shape the Answers.” Through clear examples, Schwarz showed how question format and response choices affect participants’ answers. The complete article by Schwarz is recommended reading. What follows are comments on some of the major points.

How your questions are structured communicates as much information as the content itself, in Schwarz’s (1999) view. What respondents understand and you intend may differ unless you are careful. Schwarz explained that respondents are doing their best to interpret what researchers have in mind, given their understanding of the rules of communication, and further that respondents assume that everything that is stated is in some way related to the research goals: “Unfortunately, as researchers we are often not fully aware of the information that our questionnaires—or our experimental procedures . . .—provide, and hence miss the extent to which the questions we ask determine the answers we receive” (p. 103).
A major point is that quite different responses may emerge in an open- versus a closed-ended question approach. In addition, response scales guide participants. In an example of whether people consider themselves successful in life, Schwarz (1999) used a scale from “not at all successful” to “extremely successful.” Response anchors go from 0 to 10 in one version and from –5 to +5 in a second version. This numerical distance is the same, in absolute terms. In Version 1, fewer than 15% of the respondents used the anchors in the 0–5 range, whereas in Version 2, the percentage in the –5 to 0 range was 34%. Participants see these scales differently, said Schwarz. The first scale (0 to 10) is considered to be unipolar (varying degrees of one attribute). In contrast, the second scale is viewed as bipolar, with the two ends representing opposites. The psychological meaning of the scales differs: In Version 1, the range from 0 to 5 may indicate not having success to show, that is, nothing above a threshold of 0; in Version 2, the lower 6 points of the scale (–5 to 0) may reflect actual failures, that is, falling below the threshold of 0. This example clearly illustrates the role that scale range, represented by the anchors, can play in shaping our responses (also refer back to Figure 5.2 in this chapter).

Frequency options are another structural aspect to consider. Schwarz (1999) noted that in understanding a scale, people use their knowledge about the world (i.e., their schemas) such as how often events occur (frequently vs. rarely). This point should remind you of the discussion of the availability heuristic and representativeness in Chapter 1. The middle value in a scale is usually interpreted as typical or usual (the average frequency), whereas the anchors at the ends are considered extremes of the available range. If respondents’ actual behavior is above that represented by an end anchor (e.g., texting 10 times/day), they may shift their response to align it with the “norm” (as represented in the scale range). This behavior, that is, seeking to conform, is also an aspect of social desirability, discussed earlier in the chapter.

Using an open-ended approach for the response format will help make sure the response options don’t have unnecessary weight, in Schwarz’s (1999) view. An example would be:

“How many times a day do you text?” __________/times per day.

Specifying the units of measurement (i.e., __________/times per day) for the question is important, he noted, to limit vague answers such as “a lot.” The worst choices for frequency alternatives, he stated, involve the quantifiers “sometimes,” “often,” “frequently,” and similar open-to-interpretation options. The meaning of those words depends on the respondent’s subjective standard. This variability in interpretation creates a problem for the researcher.
Try This Now

In the question “How many times a day do you text?” what assumption about the user’s habits is assumed? How could you improve on that question?

Writing Demographic Items and Social Sensitivity

Researchers who are administering a survey typically obtain some background information about their sample, as a way to both characterize the sample and rule out alternative explanations (see Chapter 3). Usually this category of information is called demographics and includes such characteristics as the age, gender, race, educational level, relationship status, and income level of the participant (as defined in Chapter 2). Often, options are listed and instructions to the participant are to “Circle the appropriate response for each item.”

With regard to gender, a common method is to print the variable name and the choices as follows:

Gender: M F

Try This Now

What problems do you see in giving participants these response options for gender?

Many possibilities for gender extend beyond male and female. Given these possibilities, a recommendation is to let people self-report their gender. Remember that the research experience is an interpersonal one. By providing the opportunity for people to self-report their background characteristics, you show sensitivity to the fact that people define themselves in many different ways. For example, one gender-related term that some individuals now use is “gender fluid.” To accommodate all possibilities, when asking about gender the item would look like this:

Gender: _______________

In general, obtain the most specific information possible (from the open-ended format), which can later be collapsed into categories if desired. The recommendation is to use the
open-ended (i.e., fill in the blank) format for your demographic items. For example, age would look like:

Age: ____________

and not as follows, where the participant circles a category:

16–20
21–25
26–30

Use of age spans may obscure group differences within a category. The advantage of the open-ended approach is its specificity; you throw away potentially useful information when you start with categories. The disadvantage is the additional effort required if you want to collapse categories, but if you are using online survey software (see later section in this chapter) that is downloadable into a program like SPSS, collapsing data into categories involves just a few steps.

**Asking About Race and Ethnicity**

For the reasons just discussed about sensitivity and capturing specific information, using open-ended questions to ask about race and ethnicity (they are different) is also recommended. Many researchers see these terms as socially constructed. For some race is defined more in terms of biology, whereas ethnicity is more a reflection of culture, often tied to a geographical region. You may wonder why it is important to ask about these characteristics if your study does not focus on these variables. Asking about these characteristics is important in terms of being able to describe your sample and whether it is representative of the population from which it is drawn. Moreover, in indicating the extent to which research findings are generalizable, it is important to indicate the backgrounds of your participants. As a side note (discussed more in Chapter 11 in the section on use of language), the *Publication Manual of the American Psychological Association* (APA, 2010b) has a very good section on reducing bias (by topic), including gender, sexual orientation, racial and ethnic identity, disabilities, and age.

How will you know whether your sample is representative in terms of characteristics like gender or race? If you are using college students as your participants, you may be able to obtain the percentages for these demographic categories from your registrar, admissions office, office of institutional research, or the institution’s website.
Asking About Income

Family income is an item that presents many challenges because participants may not know their household or family income. Moreover, asking people about money is considered an invasive question (i.e., “none-of-your-business”). Don Dillman (Dillman, Smyth, & Christian, 2014), one of the recognized authorities on survey construction, has noted that income is one question that people object to answering. In the experience of many researchers, there are more nonresponses to an item that asks about income than there are to other demographic items.

Some researchers give a list of categories as a way to help situate the range of possibilities for respondents, but this approach has several problems. First, doing so may miss important information within categories. Second, it may make people feel uncomfortable because they fit into one extreme (either the low or the high end of the income scale). Third, and related to the first problem of missing information within categories, you may have a ceiling effect, as introduced in Chapter 3. Some respondents may have substantially higher incomes than your scale categories reflect, which may lead to a restricted range at the upper end of the scale. When everyone is at the low end of the income distribution (or fails the test or performs poorly), this is called a floor effect (also introduced in Chapter 3). When income clusters at either end, the restricted range limits the likelihood you will have significant results related to that demographic variable.

Creating Your Own Questions: Item Format

In writing demographic items, the recommendation was to use an open-ended response format whenever possible. If you are developing the measure of a construct, you will likely use response anchors (i.e., decide on the number of anchors for each item and on what the anchors mean). In a scale of precautionary measures (Devlin, 2000), participants were asked to “Rate your likelihood to take these [precautionary] steps on a scale ranging from 1 (‘not at all likely to do that’) to 5 (‘definitely likely to do that’)” for each of the 19 behaviors in the precautionary measures scale (e.g., “Avoid areas where few people seem to be”). In line with Schwarz (1999), there are formatting issues to consider, including the number of anchors and the stem itself. The stem is the statement or question or prompt to which the respondent replies.

As discussed earlier in this chapter in the section on scale sensitivity and anchor values (see Figure 5.2), 5–7 anchors are recommended for surveys. Having too few points (e.g., 3) reduces the variability needed for group differences to emerge. With too many points (e.g., 12), responses become less reliable because people are overloaded with the number of choices.
Odd or Even?

Try This Now

Does it matter whether you give respondents an even (six) or an odd (seven) number of anchors?

An odd number of choices (e.g., 7) gives people the opportunity to “be in the middle,” that is, to express a neutral position. In the case of six anchors, they must be on one side of the spectrum or the other; if you want their choice to be categorized as positive or negative, for example, or to express clear agreement or disagreement, select an even number of anchors.

The Stem

The stem presents the problem statement or content of the question; it frames the material for respondents. Consider the following options:

(a) How satisfied do you feel with your institution?

| Extremely dissatisfied | 1 | 2 | 3 | 4 | 5 | Extremely satisfied |

(b) How do you feel about your institution?

| Extremely dissatisfied | 1 | 2 | 3 | 4 | 5 | Extremely satisfied |

Try This Now

Before reading further, what is the difference between (a) and (b)? What effect do you think each option has on the respondent?

The first example “leads the witness.” Because the stem includes “how satisfied do you feel,” it suggests that the respondent must possess some degree of satisfaction about the current institution. The second stem does not lead; instead, the anchors provide the context. The first example contains a demand characteristic (discussed in Chapter 3 and earlier in this chapter related to social desirability); reading the stem the participant may
think what is necessary is a report of satisfaction [think back to Schwarz’s (1999) point
that respondents are doing their best to interpret what researchers have in mind]. Also
note that the stem is bolded and the response options are not. We will talk about the
appearance of survey items later in the chapter.

**Pilot Testing**

Pilot testing questions may increase the internal validity of your research (see
Chapter 3). Test them on a small sample of people similar to your participants. Con-
sider their feedback, and revise your items accordingly. Fix problems of misunder-
standing, errors pointed out, and any other issues that arise. Because you understand
a word in a particular way doesn’t mean others will. Include operational definitions of
concepts if you want people to respond with the same semantic understanding (think
back to the beginning of the chapter where we talked about the difference between the
ideal concept and the actual measure).

---

**REVISIT AND RESPOND**

* How do the questions shape the answers, according to Schwarz (1999)? Explain
  why Schwarz recommended using open-ended response formats. How can a
  researcher demonstrate social sensitivity in the way questions are asked? Why
  would you choose to use an even number of response anchors? How can the stem
  of a question “lead the witness”? 

**SCALE TYPES AND FLEXIBILITY IN ANSWERING RESEARCH QUESTIONS**

Earlier in this chapter we discussed the four basic scale types (nominal, ordinal, interval,
ratio) and their strengths and weaknesses. You may remember that the interval scale is
the most widely used in the social and behavioral sciences. It combines the advantages of
parametric statistics with the ability to capture greater sensitivity in the DV than is true
for nominal or ordinal scales. When you are writing demographic questions, consider
the strengths and weaknesses of these scale types, especially the difference between what
you can learn from asking a “yes–no” question versus one that permits a more sensitive
measurement. We talked about this sensitivity of a scale and the anchor values in earlier
in this chapter (refer to Figure 5.2).

If you are asking a factual question that can only be answered with a “yes” or “no”
(e.g., are you enrolled in four courses this semester?), then the nominal scale approach
makes sense. Nominal scales are the foundation of **content analysis** (taking open-ended responses and creating categories to capture themes; see Chapter 6) because you are essentially recording whether a participant mentioned a particular category or not (e.g., mentioned courses or not in response to a question about areas of satisfaction with the institution). As soon as you move to questions involving attitudes and perceptions, nominal scale items are less useful because they provide very little information and their range is restricted. Think about asking questions in a way that will involve an interval scale, where a range of responses can be expressed (i.e., the degree of satisfaction with courses at the institution). As we saw earlier in the chapter when talking about scale sensitivity and also about item format, a scale range of 5–7 anchors is generally recommended.

### THE ORDER OF QUESTIONS IN A SURVEY

One of the most widely consulted books about survey research is Dillman’s book (2000) on mail and Internet surveys. Now in its 4th edition (Dillman et al., 2014), the order of approaches listed in the book’s current title reflects the changing landscape of doing survey research: *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Any researcher whose work is primarily survey-based should read this book. For Dillman, a questionnaire is like a conversation; topics that are similar should be grouped together rather than forcing the respondent to revisit a topic area already covered. He recommends starting the questionnaire with questions about the research topic [i.e., what the participant was told in recruitment materials and in the informed consent document (see Chapter 4) was the focus of the research]. Don’t start with the demographic items. In practice, demographic items should be the last items before a manipulation check.

Another issue is sensitive items, for example, dealing with experience in therapy or sexual assault. Dillman’s advice (Dillman et al., 2014) is to place these items far into the survey and gradually build up to these difficult issues. Such issues are more likely to be answered once the participant has already invested 5–10 minutes responding to other items.

### REVISIT AND RESPOND

- List a recommendation from Dillman (Dillman et al., 2014) for the order of items in a survey. Where should demographic items be located in a survey?
ONLINE SURVEY SOFTWARE TOOLS

Online survey software tools such as Survey Monkey® (introduced in Chapter 4) and Qualtrics® (introduced in Chapter 1) enable you to present your survey online for your respondents. This approach has several advantages to both researchers and participants. First, once you create your survey online, it can be sent to anyone, anywhere, with a survey link the program generates for you. The online approach offers a great deal of flexibility (for the researcher and the respondents). In Chapter 9, we will talk in more detail about using Amazon MTurk® (introduced in Chapter 2), which provides an online route to obtain participants through crowdsourcing. Crowdsourcing is obtaining something (e.g., ideas) from people online, and one form of crowdsourcing is seeking services. In this case, respondents (“workers”) are paid (by the researcher, the “requester”) to take the survey. The survey you create through SurveyMonkey or Qualtrics is typically the link that is sent to the “worker” through the Amazon MTurk site.

Another advantage to online surveys is accuracy. Because respondents are typically responding to set option choices (see Figure 5.10), the approach reduces the likelihood of data entry errors (because some online platforms allow you to download the data directly into statistical software). Third, the approach promotes sustainability because you cut down on the use of paper. Fourth, because some of these programs (e.g., Qualtrics) enable you to download the responses directly into SPSS, this saves time for the researcher (and again cuts down on data entry errors).

What are the disadvantages? First, these programs cost money. The account subscription is usually paid by the institution, but someone has to pay. Second, there is a learning curve to using these programs. Many instructors and students feel that these programs add another significant component to the material that must be covered in a research methods course. Instructors don’t always have the time (or the expertise) to cover these survey tools; students may be faced with learning the intricacies of the programs by themselves. Third, when you rely on this approach, some respondents may not have access to computers and the Internet and cannot access your survey. Fourth, the flexibility these programs offer can be a threat to internal validity. If respondents can take the survey (by accessing the link) anytime and anywhere, the circumstances under which this occurs will not be uniform across participants. This variability in procedure is a threat to internal validity. The researcher has no idea whether the respondent was watching television while completing the survey, for example.

To deal with this problem, some departments require students in the participant pool to come to a physical location (where the researcher is present) to access the survey. Usually the researcher has the e-mail address of each respondent, with the link included in the body of the e-mail (in an e-mail draft), and sends that e-mail as soon as the participant appears at the research site. This approach also cuts down on the likelihood that
respondents will start the survey and stop after a few items; in this case, the demand characteristic created by the presence of the researcher is not such a bad thing!

FEATURES OF ONLINE SURVEY SOFTWARE PROGRAMS

An online article by Michaela Mora from 2013 (http://www.relevantinsights.com/free-online-survey-tools) compares some of the features of several software survey tools: SurveyMonkey, SurveyGizmo®, and QuestionPro®. Figure 5.11 has been created for you summarizing some of the critical information from Mora’s article.
What you should notice is that the price is “free,” but that is because these are “limited” versions of the software. If you want more features, you have to pay. Particularly noteworthy is that the “free” version of SurveyMonkey limits you to 10 questions and does not offer data exporting. You need the “Gold” plan (or higher) for that, which costs $300 annually (June 12, 2016; https://www.surveymonkey.com/pricing/?utm_source=header).

If we compare the two most widely known survey software programs (SurveyMonkey and Qualtrics), we learn the following (see Figures 5.12 and 5.13):

For SurveyMonkey, if you want randomization (the ability to randomize questions, pages, or conditions across participants), which would be an important tool for research, you need the $300 annual “Gold” package.

Many researchers view Qualtrics as more powerful than SurveyMonkey. For Qualtrics, the “free account” is really a trial version because once you have 100 responses, you need to upgrade to a paid account (see Figure 5.13). For some student research projects, 100 responses may be more than sufficient.
Qualtrics seems to be a more dominant presence in academic research ([http://www.informationweek.com/qualtrics-dominates-academic-survey-research/d/d-id/1110904](http://www.informationweek.com/qualtrics-dominates-academic-survey-research/d/d-id/1110904)), but SurveyMonkey is also a viable choice for student research. There are

**FIGURE 5.12 Overview of SurveyMonkey**

<table>
<thead>
<tr>
<th>SurveyMonkey</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Free account</strong></td>
</tr>
<tr>
<td>• 10 questions</td>
</tr>
<tr>
<td>• 100 responses</td>
</tr>
<tr>
<td>• Standard e-mail support</td>
</tr>
<tr>
<td><strong>Next level</strong></td>
</tr>
<tr>
<td>• $26/month</td>
</tr>
<tr>
<td>• Unlimited questions</td>
</tr>
<tr>
<td>• 1,000 responses</td>
</tr>
<tr>
<td>• Data exporting</td>
</tr>
</tbody>
</table>


**FIGURE 5.13 Overview of Qualtrics**

<table>
<thead>
<tr>
<th>Qualtrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Free trial account</td>
</tr>
<tr>
<td>• 100 responses, one active survey at a time</td>
</tr>
<tr>
<td>• No question limit, 8 question types</td>
</tr>
<tr>
<td>• Access to all logic functions and randomization; online reporting</td>
</tr>
<tr>
<td>• Full access pricing for academic users typically $500/yr.; 1,000 responses and unlimited surveys; unlimited phone and e-mail support</td>
</tr>
</tbody>
</table>
some more advanced functions that Qualtrics but not SurveyMonkey has, such as the ability to set viewing time for a page, but SurveyMonkey has been expanding its functional capabilities, and the differences between the programs may not be as pronounced as they were.

PROGRAM FEATURES

These programs have many features, such as the preset question types and response options, where you can change both the kind of anchors and the number of anchors for a given question (see Figures 5.14 and 5.15, from a Qualtrics file). You can also include a progress bar, which shows participants how much of the survey they have completed.

FIGURE 5.14  Screenshot of Preset Question Types and Response Options in Qualtrics

Source: © 2016 Qualtrics.

Copyright ©2018 by SAGE Publications, Inc. This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.
Again, learning the programs takes time. Figure 5.16 shows part of a Qualtrics survey in which the randomizer function is used. This function allows the researcher to randomize the order of presentation of different conditions (contained in separate blocks) across participants. Note that there is an initial Block for Consent, the next Block is for Instructions, then the Condition (here named “nice photo” or “ugly photo”) would be shown (note, only the researcher sees that label), followed by a Block with the DVs (74 questions), and ending with a Block that contains the Debriefing.

Copyright ©2018 by SAGE Publications, Inc. 
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.
Figure 5.16 is presented as an illustration of the kinds of features available, and it provides a reality check. These programs take time to learn, and you need to spend the time if you plan to be successful using these programs. Also, as a reminder, if you use technology, the links to the surveys may not always work. Prepare a backup plan.

**GOOGLE DOCS FORMS**

**Google Docs® Forms** can be used as a platform available to collect data, and it is free. Using Google Docs Forms, you can create a survey and then invite people to participate via e-mail. Their responses are saved automatically using a Google Docs spreadsheet, which in turn can be downloaded to Microsoft Excel® on your desktop. There are many tutorials on the Web that illustrate the steps involved in using this approach. A comparison of Google Docs Forms with SurveyMonkey (prepared by SurveyMonkey, which obviously emphasizes its advantages) can be found at [https://www.surveymonkey.com/mp/surveymonkey-better-than-google-forms/](https://www.surveymonkey.com/mp/surveymonkey-better-than-google-forms/).
DOWNLOADING ONLINE SURVEYS INTO SPSS

From the researcher’s standpoint, one advantage to using software survey tools is the ability to download the files directly into a statistics program, such as SPSS, or into a spreadsheet, such as Excel. Once the file is downloaded into SPSS, there is usually some additional work to do, particularly to label variables if you did not give them names in your online survey. In addition, although you will be able to identify which participant was in which condition, you still need to create new variables to represent those conditions in SPSS. The example in Figure 5.17 uses Qualtrics.

The far right-hand column in Figure 5.17 shows which participant had which condition, but if you want to run a $2 \times 2$ analysis (one factor had two levels of photos; the other factor had two levels of text), you need two new columns, one that represents the level of photo the participant saw and the other that represents the level of text the participant saw. It is simple to create those new columns in SPSS; the point is that your work is not necessarily done when you download those responses into SPSS (see also Chapter 10 on Managing Data and Results). Furthermore, you need to have some

![Figure 5.17 Example of a Qualtrics Survey Downloaded Into SPSS](source: © 2016 Qualtrics.)

<table>
<thead>
<tr>
<th>Q91_1</th>
<th>Q91_2</th>
<th>Q91_3</th>
<th>Q91_4</th>
<th>Q91_5</th>
<th>Q91_6</th>
<th>Q90</th>
<th>Q91</th>
<th>Q93</th>
<th>Q94</th>
<th>Q95</th>
<th>DO_BR_FL_13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>Connecticut</td>
<td>5</td>
<td>6</td>
<td>Good photo/good text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>Pennsylvania</td>
<td>4</td>
<td>8</td>
<td>Bad photo/bad text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>massachusetts</td>
<td>5</td>
<td>4</td>
<td>Bad photo/bad text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Good photo/good text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MA</td>
<td>4</td>
<td>6</td>
<td>Good photo/bad text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Good photo/good text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NY</td>
<td>8</td>
<td>1</td>
<td>Bad photo/good text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>California</td>
<td>6</td>
<td>8</td>
<td>Good photo/good text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bad photo/bad text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>New York</td>
<td>1</td>
<td>7</td>
<td>Bad photo/bad text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bad photo/bad text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>MA</td>
<td>4</td>
<td>1</td>
<td>Bad photo/bad text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bad photo/bad text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>New York</td>
<td>6</td>
<td>8</td>
<td>Bad photo/bad text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bad photo/bad text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>New Hampshire</td>
<td>3</td>
<td>5</td>
<td>Good photo/good text</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bad photo/good text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bad photo/good text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>New Jersey</td>
<td>6</td>
<td>7</td>
<td>Good photo/good text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>Nw York</td>
<td>9</td>
<td>9</td>
<td>Good photo/good text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>New York</td>
<td>2</td>
<td>1</td>
<td>Bad photo/good text</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: © 2016 Qualtrics.
idea what you are viewing. You can see that the researcher did not label each variable in the Qualtrics survey. Thus, there are a lot of Qs with some numbers after them (e.g., Q93). Without a specific label, you would need to go back to the Qualtrics survey to see what question Q93 is. A better approach would be to take the time to label all of those variables in the original survey. Again, the labeling is seen by the researcher, not by the participant.

On balance, the advantages of being able to download data directly into SPSS probably outweigh the disadvantages, but one more comment is worth considering. Yes, online surveys are increasingly used, but not every participant in every research project will have access to a computer. Even if you are using an online survey, it is good practice to have a paper version available.

SURVEY APPEARANCE

How your survey looks may affect the quality of the data respondents provide. In particular, it is important to think about bolding, numbering, and spacing.

Earlier in the chapter we looked at the difference between question stems and anchors. One of those questions is represented here. You notice that the stem is bolded but the anchors are not.

14. How do you feel about your institution?

| Extremely dissatisfied | 1 | 2 | 3 | 4 | 5 | Extremely satisfied |

Try This Now

Before reading further, what is the purpose of bolding the stem?

Bolding the stem draws attention and separates it from the anchors. Furthermore, in the illustration here, we have added a question number (14). Each question should be numbered, again to draw attention to it. Another issue is the spacing; you should leave enough space between questions to reduce the possibility that respondents will skip over them.

Online survey software spaces the items for you, both vertically on the page and the presentation of the anchors themselves (which can be horizontal or vertical). It is also possible to adjust that spacing to meet your needs. One function of Qualtrics allows
researchers to add a prompt that appears if a respondent has skipped one or more items on the page. We cannot force people to answer items (research is voluntary; see Chapter 4 dealing with ethical issues), but we can give them the opportunity to answer an item they may have inadvertently skipped.

Online survey software programs typically have a section where you can make such adjustments to the survey. In Qualtrics, this section is called “Look and Feel.” In addition to the adjustments already mentioned, you can determine how many questions per page you want, highlight questions or parts of questions, and add a visual separator between questions. Qualtrics also allows you to have your survey delivered via a smartphone as well as a computer (and will alert you when your survey may not be suitable for smartphone delivery, such as surveys with large graphics).

The care you take in preparing your survey will reflect positively on you and may encourage respondents to take the research project more seriously. This care may also contribute to the internal validity of the study, in that participants may complete more of your questions.

**Try This Now**

One feature of Qualtrics is the ability to randomize the presentation of questions on a page and to randomize the distribution of conditions across participants. If you were going to use a paper survey, how would you accomplish those two kinds of randomizations?

**REVISIT AND RESPOND**

- List two advantages of using online survey software such as Qualtrics and SurveyMonkey. List one disadvantage. What is the advantage of having participants fill out a survey link in your presence rather than simply sending them the link to fill out on their own time? What is the point of “bolding” the stem of a question and labeling each question? Why is it important to label all of your variables in the online survey software? Does the participant see these variable labels? Once you download your data from the online survey software into a statistical program such as SPSS, what else might you need to do if you have different conditions?
Summary
This chapter has given you the tools you need to identify and select existing measures to use in your research. You know the criteria for selection including internal consistency, length, and difficulty. You also know the logical places to look for the measures, including the Measures/Instruments section of empirical journal articles, your department resources, the author of a given scale, PsycTESTS, and test publishers. If the need arises, you can also even create your own scale and calculate its Cronbach’s alpha. Furthermore, you know Schwarz’s (1999) recommendations for the construction of scale items and response choice formats and can write sound demographic items yourself.

In addition, you should be familiar with some of the most widely used survey software, including Qualtrics and SurveyMonkey, and be able to identify their advantages and disadvantages. Finally, you should recognize that how your survey looks to respondents may very well affect the quality of the data they provide.

In the event that you did not consider them earlier, here are the REVISIT and RESPOND prompts that appeared in this chapter.

- What is an operational definition? Give an example. Why is it hard to capture in full the essence of an abstract theoretical concept with a measure? Why is it harder when you measure managerial competence than when you measure rate of speech?
- Describe the four different scale types (nominal, ordinal, interval, ratio). Explain the frequent use of interval scale data in the social and behavioral sciences.
- Where should you start your search for a measure? How do PsycTESTS and HaPI differ? What do you learn from reading about the permissions for a scale? What kinds of tests typically have qualifications for use? If a fee is charged for a scale, what other avenues might you pursue before paying the fee?
- Define internal consistency in your own words. Is there one set Cronbach’s alpha value for all research purposes? If not, what circumstances require the highest alpha? List the steps you would take to run a reliability analysis on SPSS.
- Give an example of how you would establish content validity. What does it mean to say a measure has face validity? Why is a measure that has demonstrated predictive validity more powerful than a measure with concurrent validity? In your own words, explain what a nomological network is.
- What is socially desirable responding? Name two approaches to dealing with this threat to internal validity.
- How do the questions shape the answers, according to Schwarz (1999)? Explain why Schwarz recommended using open-ended response formats. How can a
researcher demonstrate social sensitivity in the way questions are asked? Why would you choose to use an even number of response anchors? How can the stem of a question “lead the witness”?

- List a recommendation from Dillman (Dillman et al., 2014) for the order of items in a survey. Where should demographic items be located in a survey?
- List two advantages of using online survey software such as Qualtrics and SurveyMonkey. List one disadvantage. What is the advantage of having participants fill out a survey link in your presence rather than simply sending them the link to fill out on their own time? Why is the point of “bolding” the stem of a question and labeling each question? Why is it important to label all of your variables in the online survey software? Does the participant see these variable labels? Once you download your data from the online survey software into a statistical program such as SPSS, what else might you need to do if you have different conditions?

**BUILD YOUR SKILLS**

1. Draft your own demographic questions to assess major, geographical upbringing, and political orientation.

2. Sign up for a free trial account of SurveyMonkey or Qualtrics or use Google Doc Forms to create a brief survey to test out the various question types and features.

*SAGE edge*

gagepub.com/devlin

Sharpen your skills with SAGE edge! SAGE edge for students provides a personalized approach to help you accomplish your coursework goals in an easy-to-use learning environment. You’ll find action plans, mobile-friendly eFlashcards, and quizzes, as well as videos, web resources, and links to SAGE journal articles to support and expand on the concepts presented in this chapter.