14

Theory and Measurement

Types of Measurement Strategies

I have been struck again and again by how important measurement is to improving the human condition.

—William Gates (2013)

The concepts of reliability, validity, and measure generalizability are core to measurement theories. Chapter 13 described theory construction perspectives that can help you identify factors that impact the reliability and validity of measures and also help you select extant measures for your research. However, it did not consider measure and scale construction per se, that is, theory that drives the formal construction of measures. The present chapter does so. As with Chapter 13, our intent is not to provide a checklist of good measurement practices, although you will learn about such practices as you progress through the chapter. Rather, we stress that to measure a construct well, you need to adopt a theory construction mindset that specifies determinants of key measurement properties and that takes into account, at minimum, the population you are studying, the structure and content of the measures you use or construct, and the assessment context that you will ultimately use.

Our focus is on self-reports, observer reports, and "objective" measures that collectively form the backbone of research in the social sciences. Although our discussion is dominated by quantitative research orientations, much of what we cover is relevant to qualitative research; it certainly bears on mixed-methods research that combines qualitative and quantitative methodologies. As in Chapter 13, we emphasize idea generation rather than formal tests of those ideas. Although you could pursue such tests and potentially publish their results, our goal is to encourage and elaborate a theory construction mindset. We omit the complex topic of multi-item scale construction using approaches such as item response theory, Thurstone scaling, and Guttman scaling. These topics are beyond the scope of this chapter (they are discussed on our companion website at *www.theory-construction.com*). In addition, we do not attempt to provide comprehensive theories of self-reports, observer reports, and "objective" measures, given the inevitable

complexity of such theories. Again, our goal instead is to create a mindset about the use of theory construction for measurement and to provide guidelines and examples to help you think along such lines.

CONSTRUCTING A THEORY OF SELF-REPORTS

Self-reports are widely used in the social and health sciences. Three processes underlie self-reports for organizing one's theory construction efforts. First, people must understand the question posed to them as intended by the researcher; that is, there must be *comprehension*. Second, people must form or retrieve from memory judgments and opinions in response to the posed question to form an answer in their minds. This *judgment* process typically involves cognitive and affective mechanisms playing themselves out in working memory. Third, once the judgment/opinion is formed, people must communicate it to the researcher. Sometimes people provide their responses in an open-ended format and other times on a rating scale. The act of communicating one's answer to the investigator is called *response translation*. Measurement theorizing about comprehension, judgment, and response translation can shape how you ask questions and how you structure response options. We have found that organizing our thinking around these processes is a useful theory construction frame. In the next sections, we specify factors that impact each process separately.

Constructing Theories of Comprehension

Your first step is to construct a theory of comprehension relative to your measures. Comprehension of a question obviously is impacted by a person's vocabulary. If a question is posed in a language you do not know, you will have little sense of what the question means. At the same time, the human mind has a remarkable capacity to bring past experience to bear to make sense of what is technically gibberish as the brain fills in gaps and reorganizes "nonsense to sense" in text processing. As evidence, read the following text:

> Aoccdrnig to rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a total mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe. Amzanig huh?

Such "fill-in" processing is a two-edged sword, sometimes working in favor of good measurement and other times not.

To effectively construct questions that are understandable, we ultimately need a theory of comprehension and text processing. What factors impact comprehension of a question, text, or a preamble? How can you address these factors to maximize comprehension? Components of your theory might generalize across different substantive domains, but you also undoubtedly will need to introduce nuances specific to your

research area. Our discussion emphasizes general factors that impact comprehension to help stimulate idea generation on your part. To organize our framework, we draw on the core measurement facets from Chapter 13, namely, characteristics of the population, characteristics of the questions, and characteristics of the assessment context that can affect comprehension.

Population Characteristics

Characteristics of your population will impact your respondents' ability to comprehend the questions you pose. As such, you need to theorize about these characteristics. What guidance can extant social science and measurement theories provide as you think about this matter? Let's consider some relevant factors.

Literacy. One variable that has been found to affect comprehension is *literacy*. Literacy is defined as the ability to read and write, but the concept is sometimes nuanced further via the concepts of *functional literacy* and *functional illiteracy*. Functional illiteracy refers to having trouble in everyday life because of reduced reading and writing skills. Distinctions often are made between specific types of literacy, such as health literacy, financial literacy, e-digital literacy, and numeracy. In terms of general literacy, research in the United States suggests that a person's reading level is usually three to five grade levels below the highest grade the person has completed (Agency for Healthcare Research and Quality, 2015). If your target population is juniors and seniors in high school, for example, you probably can assume about a seventh-grade reading level and write your questions accordingly. This heuristic does not always apply, so it should be used only as a rough guideline. For example, in Detroit, a workforce studying illiteracy found that about half of those who were functionally illiterate had high school degrees (Detroit Regional Workforce Fund, 2011). Several brief assessment tools for literacy can be used in pilot testing to gain a better understanding of the literacy of your population, if need be (e.g., Davis et al., 1993). Even if questions are to be read to respondents, literacy may be relevant. For example, it has been shown that learning to read reinforces other key abilities, such as verbal and visual memory, phonological awareness, and visuospatial and visuomotor skills (Ardila et al., 2010), all of which impact information processing.

Computerized readability formulas, such as the Flesch–Kincaid readability index available in Microsoft Word, can be applied to a question or text to gain a sense of its reading level. These indices are limited because they focus on the length of words and sentences rather than comprehension per se; shorter words and shorter sentences are assumed to be more readable. As a general guideline, it has been suggested that material written at the fourth- to sixth-grade level based on these indices is easy to read; seventh- to ninth-grade materials are of average difficulty; and material at or above a tenth-grade level is deemed difficult (Agency for Healthcare Research and Quality, 2015).

As noted, there are many subtypes of literacy that can impact the way you approach question design. The number of content-oriented literacy types is considerable (e.g., food literacy, health literacy, financial literacy, social literacy). In some ways, these dif-

ferent forms of literacy reflect how knowledgeable people are in different topic areas. Thinking about such matters relative to your research topic may help you determine terms you can safely use in questions.

One type of literacy that generalizes across many content domains is that of *numeracy* (Reyna, Nelson, Han, & Dieckmann, 2009). Numeracy is the ability to understand and use numbers, including the ability to perform simple arithmetic operations, compare numerical magnitudes, and to understand fractions, proportions, percentages, and probabilities. Numeracy is relevant, for example, if your questions include percentages (e.g., agreement or disagreement with the statement "This new treatment controls cancer in about 40% of cases like yours") or if your response metric uses percentages ("If women use birth control pills for 6 months, what percentage of them do you think will accidentally become pregnant?"). In a study of patients considering a new treatment with a 40% cure rate, Weinfurt and colleagues (2005) found that 72% correctly interpreted the percentage; 16% of patients interpreted the statement to mean either the doctor was 40% confident the treatment would work or that the treatment would reduce disease severity by 40%; and 12% indicated they did not understand the statement at all. If your population has low numeracy, it may be best to avoid questions with these formats (Reyna et al., 2009).

Working Memory Capacity and Cognitive Skills. Another potentially relevant consideration is working memory capacity. Psychologists distinguish three types of memory: short-term memory, working memory, and long-term memory. *Short-term memory* refers to the short-term storage of information without actively processing that information. *Working memory* refers to processes to interpret, elaborate, and act on a subset of the information that is in short-term memory. The cognitive processes used by working memory include attention, encoding, integration, processing, and retrieval of information from long-term memory, all of which are used to interpret the information (which is part of the comprehension process), and, ultimately, to form judgments. *Long-term memory* is our permanent storehouse of memories in the brain, which also can be accessed by working memory.

The processes used in working memory are central to comprehension and reasoning when answering a question that has entered short-term memory. Short-term memory can store information for about 10 seconds unless working memory processes or "manipulates" it (Miller, 2003). Unless acted upon by working memory, information in short-term memory decays rapidly. It is difficult to specify the capacity of short-term/working memory, but Miller's (1956) classic work suggests that for adults it is 7 bits of information, ± 2. More recent work indicates that the limit may be lower, whereas others have argued that it is impossible to know because of phenomena like information chunking (e.g., where the numbers 1, 2, 3 are chunked into 123; Miller, 2003). Still, it is safe to say that the amount of information people can keep in their conscious mind is time sensitive and rather limited and that Miller's 7 ± 2 is a rough guideline for appreciating the limits of conscious information processing. Consideration of working memory limitations in question design is particularly relevant for research with children and the elderly, because comprehension is hindered by limited working memory capacities.

With such populations, questions must be that much shorter, simpler, and more concrete.

In addition to working memory, other cognitive skills relevant to comprehension include processing speed (how quickly one digests new information); attention (the ability to sustain focus); verbal reasoning (the ability to understand linguistic information); abstract thinking (the ability to think abstractly); verbal memory (how efficiently one encodes and recalls linguistic information); and visual memory (how efficiently one encodes and recalls visual-spatial information). Are any of these skills relevant to the measurement strategies in your research? How? How would respondent limitations on one or more of these skill(s) shape the questions you design?

Language Diversity. Migration has increased in many countries, and with it has come linguistic diversification. Individuals who do not have a good speaking knowledge of the native language in which questions are written or phrased may have difficulties comprehending questions. Kleiner, Lipps, and Ferrez (2015) found that resident foreigners were more likely to have comprehension problems due to reduced language mastery *and* reduced motivation to conscientiously complete an interview, presumably because of task difficulty. In some cases, you may need to offer versions of your survey in different languages. Methods for effectively translating surveys use *forward–back translation* methodology. This involves expert translation to the new language, followed by independent retranslation of the translated survey back to the original language. The original version and the back-translated versions are then compared to each other, and the disparities found signal potential translation issues. See the World Health Organization (2018) website for elaboration of the steps in the forward–backward translation approach. Peytcheva (2008) argues that the language used in a survey can itself create a cultural lens for interpreting and responding to questions (see also Johnson, Holbrook, & Cho, 2006).

Attention and Motivation. People will be more likely to miscomprehend questions if they do not attend to them, partially attend to them, or read/listen to them superficially. As such, respondent motivation to take the assessment task seriously is another factor that can affect comprehension. Providing research participants with motivating instructions (e.g., stressing how important it is to attend to everything carefully and how their participation will contribute to science and/or society) can help. Remuneration is often used as a task motivator, although research suggests that the effect of payments on task motivation is complex (Bonner, Hastie, Sprinkle, & Young, 2000). Wenemark, Persson, Brage, Svensson, & Kristenson (2011) apply formal motivation theory to the analysis of survey responses. What factors might impact the motivation to work conscientiously on the self-report tasks in your research?

In sum, it is important for you to construct a theory of question comprehension as you plan your research. What characteristics of your study population might impact question comprehension and how you approach measurement? What is the literacy level of your population? What specific types of literacy are relevant to your questions? What cognitive skills does your population need to answer your questions, and how do you

adjust for low skills in question design? Is the working memory capacity of your population limited? Does your population include people whose native language is different from the language you intend to use? Do you need different language versions of your questions? What can you do to increase motivation to work conscientiously on your task?

Characteristics of Questions

The survey research literature offers many ad hoc but common-sense principles for improving question comprehension. For example, Alwin and Beattie (2016) discuss research supporting the KISS principle—"*K*eep *i*t *s*imple, *s*tupid"—to encourage question designers to be brief and to the point in constructing questions. Other tips include the following: (1) avoid technical terms, jargon, and slang (e.g., many respondents might refer to marijuana as "weed," but it is unwise to assume this term is universally understood); (2) avoid abbreviations (e.g., the item "I know the whereabouts of my child 24/7" might fit the way many parents talk, but it might be confusing to some parents); (3) avoid words with ambiguous meanings; (4) strive for wording that is specific and concrete as opposed to general and abstract; and (5) avoid questions with single or double negations (e.g., "I am not disapproving of legislation restricting access to abortion")

Lexical and Semantic Processing. One can think about question comprehension linguistically at two levels: the lexical and the semantic. At the *lexical level,* concern is with how people understand the individual words in the question. At the *semantic level,* concern is with the meaning of the overall question/sentence or portions of it based on those words. Sentence meaning is not a simple function of the individual words that make up the sentence. Rather, there is an interplay between lexical and semantic processes in shaping individual word interpretation and the implied meanings of a sentence. In the question "How many tobacco cigarettes did you smoke yesterday?," the focus is on the action of smoking a tobacco cigarette. People may be able to provide meaningful and accurate definitions of the words *cigarette* and *smoking* but differ in how they comprehend and interpret the act of "smoking a tobacco cigarette." Suessbrick, Schober, and Conrad (2000) found that 23% of respondents said the question referred to only cigarettes one had finished, 23% felt it included cigarettes one finished or partly smoked, and 54% defined it as taking even one puff from a cigarette. The counts that people report in response to this question can differ depending on their understanding of the implied action. In web-based or computer-assisted surveys, some researchers allow people to click on terms or collections of terms embedded in questions if they are unsure of their meaning, at which point a pop-up bubble is shown on the screen that defines the clicked terms. In our opinion, researchers ideally should frame their questions so that such clarifications are not necessary, but such prompts might nevertheless serve as useful comprehension aids.

Language Structure. Variables related to linguistic structure also can affect comprehension. Sentence complexity impacts reading fluency, which, in turn, can affect

comprehension, perhaps by undermining task motivation because people find the task difficult or too time consuming. Thompson and Shapiro (2007) emphasize the importance of the order in which major elements appear in a sentence. For example, it is easier to process the sentence "John (S) kicked (V) the ball (O)" than "The ball (O) was kicked (V) by John (S)" or "It was the ball (O) that John (S) kicked (V)." Lenzner, Kaczmirek, and Lenzner (2010) identified text features that detract from comprehension, including (1) the use of low-frequency words (people understand words that occur more frequently in their language); (2) the use of vague relative terms, like *many, often, rarely,* and *substantially*; (3) the use of vague or ambiguous noun phrases (e.g., "John showed the boy a picture of his mother," which is ambiguous with respect to whether the picture was of John's mother or the boy's mother); (4) complex syntax (in which a sentence contains both a dependent clause with words like *if, when, while, after, before, because,* coupled with an independent clause, yielding something like "I voted for Donald Trump because it would help the economy"—which is confusing for someone who voted for Donald Trump but not because he or she thought it would help the economy); (5) complex logical structures (sentences that include numerous logical operators, like *or,* which can require respondents to keep a large amount of information in mind); (6) low syntactic redundancy (using uncommon grammatical structures); and (7) bridging inferences (needing to make inferences from one part of the sentence to another). Graesser, Cai, Louwerse, and Daniel (2006) developed software that automatically evaluates questions for the presence of most of these features. We recommend its use.

Assessment Context Characteristics

Features of the assessment context are another source of question miscomprehension. Any facet of the context that detracts from people paying attention to questions (noise, interruptions, poor lighting, the presence of others, smells) can affect question comprehension. Similarly, facets of the assessment context that reduce respondent motivation to approach the task conscientiously are relevant because they lessen attention to the task at hand. Sometimes these facets can be subtle, such as when question ordering impacts task motivation. For example, the conventional wisdom is that (1) initial questions in a survey should be simple and build rapport; (2) questions at the beginning should address the topics described during informed consent; (3) questions on the same topic should be grouped together; and (4) questions on sensitive topics should occur at the end of the survey (Krosnick & Presser, 2010). For online surveys, respondents often become more conscientious when prompted by a pop-up message reminding them to read items carefully if the computer detects they are making too many item nonresponses, not differentiating their responses to questions (answering them all in the same way), or answering questions too quickly so as to suggest only casual reading (Zhang & Conrad, 2016).

In face-to-face interviewing, methodologists distinguish between *standardized interviewing* and *conversational interviewing* (Schober & Conrad, 2002). In standardized interviewing, interviewers are required to read a question exactly as worded and provide only neutral answers to respondent probes. This approach has the advantage of reducing

inappropriate interviewer influence and ensuring that everyone responds to the same "stimuli." In conversational interviewing, interviewers read questions as worded, but they then use whatever clarifications are necessary to convey the meaning of the question. This often improves responses by ensuring that all respondents interpret the question in the same way. Conversational interviewing thus tends to increase comprehension but at the risk of interviewers inappropriately putting words into the respondents' mouths. Conversational interviews generally take longer than standardized interviews and require more interviewer training (Schober & Conrad, 2002). Replication of results has become an issue in some social science disciplines (Earp & Trafimow, 2015), which also raises concerns about the use of conversational interviewing. Which approach is best for your research?

In sum, good measurement practice requires you to theory construct about factors that impact question comprehension for your research and then use that theory to guide your approach to measurement. Relevant variables in your theory will include characteristics of your target population, characteristics of the questions per se, and characteristics of the assessment context you intend to use, among others. We have elaborated general variables in each of these domains to help stimulate your thinking. These variables are summarized in Figure 14.1, which is a simplified representation

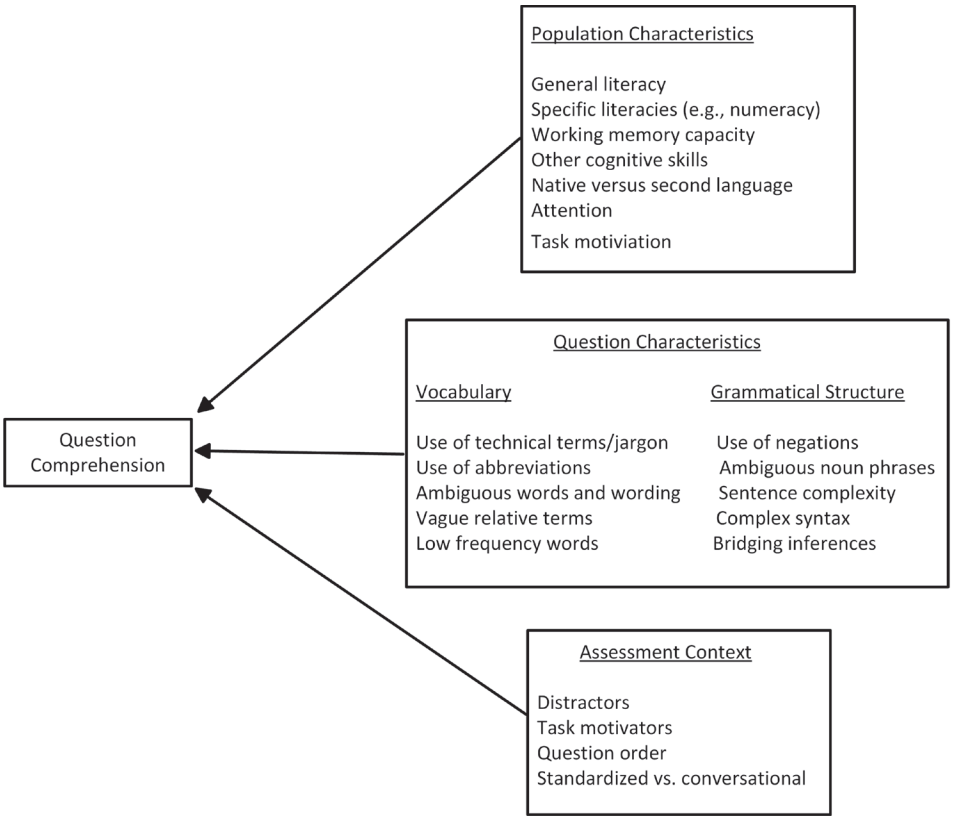


FIGURE 14.1. Factors impacting comprehension.

because it omits likely interactive/moderated relationships between factors and dependencies between the various causes.

Constructing a Theory of Judgment

Once a question has been understood, people form an answer in their minds. Although it may seem counterintuitive, understanding how people formulate judgments and answers also can impact question design. This fact reveals, again, the need for theory construction for measurement purposes, but now around the judgment process. Social scientists have studied so many types of judgments that we cannot begin to consider this topic comprehensively. Instead, we illustrate the process using a common question type in social science research, namely, that of asking people the number of times they have performed a behavior or experienced an event in the past (e.g., the number of times they have seen a doctor in the past year; the number of days they used marijuana in the past month).

When individuals are asked to recall the number of times they have performed a behavior or experienced an event over a specified time period, they might use several cognitive strategies. One strategy is to think of the relevant time period and then try to recall each occurrence of the event, counting them up as each event is recalled. Individuals could either begin with the most recent event and count backward in time, called a *think backward strategy*, or they could start at the beginning of the time period and count recalled instances that occur sequentially since the inception date, called a *think forward strategy* (Loftus & Fathi, 1985). Some studies favor the use of think backward strategies to maximize recall accuracy because the cues for recent events are more readily available in memory and can serve as cues for accurate recall of the earlier events (Loftus, Smith, Klinger, & Fiedler, 1992). Other studies favor a think forward strategy because it involves recalling events as they unfolded over time, providing a more natural structure (Loftus & Fathi, 1985). Overall, research tends to favor think backward strategies for maximizing recall accuracy, but there are exceptions (Bradburn, Rips, & Shevell, 1987). For both strategies, the memory of one event may interfere with the memory of another event. In addition, there may be order effects, with events that occur more recently having a greater impact on the final judgment. Whatever the case, the fundamental nature of this judgment process is episodic in that individuals try to recall specific episodes of the event and then count them up.

A second mental strategy is to use a rule-based judgment process. In this case, the individual invokes a stored rule or algorithm from memory that is used to generate the requested frequency without recourse to recalling specific episodes of the event. For example, asked how many days you have smoked marijuana in the past month, you may reason you do so every day and therefore report a frequency of 30.

Other cognitive strategies involve a combination of episodic and rule-based processes. When adolescents make a judgment of the frequency of marijuana use during the past 12 months, they may think episodically of the number of times they smoked marijuana during the past month and then adopt a generating rule that multiplies this result by 12. They also might cognitively invoke "correction" factors to account for months when unusual events occurred, such as being away on a family vacation.

These processes have implications for the choice of a time period over which to measure the count. Questions focused on short time intervals, such as 1 month, probably encourage episodic recall strategies, which can be reasonably accurate for individuals who engage in the behavior infrequently (Blair & Burton, 1987). However, a focus on counting individual episodes may be counterproductive for individuals who engage in the behavior frequently because of episode interference and the difficulty of keeping the episodes distinct in memory. A longer time frame (e.g., 3 months or 6 months) might yield more accurate recall because individuals who engage in the behavior infrequently will still rely on episodic cognitive strategies, whereas those who engage in the behavior more frequently will adopt a rule-based process that is less subject to episodic distortion, presuming individuals use reasonably accurate rule-generating criteria. If the focal time period becomes too long, such as recall over a period of 1 or more years, then those who engage in the behavior infrequently may have trouble recalling episodes that occurred in the distant past, and those who use rule-generating criteria may apply a rule that is appropriate for recent events (e.g., the past 3 or 4 months) but not for such a broad time base. These considerations led Jaccard, McDonald, Wan, Dittus, and Quinlan (2002) to predict that recall accuracy would be best for *moderate* time periods (3 or 6 months) as compared to short (1 month) or long (12 months) periods when young adults were asked to recall their frequency of sexual intercourse. (Of course, when the time period is extremely short, such as 1 day, recall should be accurate because individual episodes are distinct and readily recalled by all). Note that the Jaccard et al. prediction is counter to the common-sense notion that shorter recall periods, by fiat, lead to greater accuracy. Their predictions were borne out in their data.

Often, when we measure past behavior, we do so with the idea of measuring the behavioral proclivities of individuals, such as the tendency to smoke marijuana a great deal, a moderate amount, or very little/not at all. If the time frame used in the question is very short, then occurrence of the behavior may reflect idiosyncratic, situation-specific influences rather than a general behavioral proclivity on the part of the individual. For example, if the time interval is a week, a person might smoke marijuana during that week while away visiting a distant friend who has marijuana readily available and who pressures the person to smoke; however, this is a situational exception rather than a general behavioral proclivity. A longer time period gives a better "sampling" of behaviors for identifying a behavioral proclivity. If the time period is too long, however, then the underlying proclivity might change between the start of the period (when the proclivity was, say, high) and where the individual is now (when the proclivity is low). This logic also favors the use of moderate time frames, unless one is specifically interested in situational influences.

Finally, with longer time frames, individuals may have greater difficulty keeping clear in their minds the precise time period on which they should focus. This is typically addressed by using psychological "landmarks" (such as a birthday or an important or notable event) that define the beginning of the period and serve as a reference point for the individual (Shum, 1998). Even in the presence of such landmarks, however, an individual's thoughts might drift to periods outside the landmarks. Such cognitive drifting is more likely to occur when the time period is long in duration.

An interesting phenomenon that sometimes occurs for count judgments is *cognitive rounding* or *cognitive heaping* (Smitherman, Martin, Penzien, & Lipton, 2013). When making judgments about the number of days on which an event occurred over the past 3 months, for example, individuals tend to round estimates to the nearest 5 days once the frequency exceeds 10 days; above 20 to 30 days, rounding tends to occur to the nearest 10 days. This "cognitive heaping" can yield an ill-shaped distribution of scores; it can bias means and standard deviations; and it raises questions about whether a reported score of, say, 27 (someone who did not round) is truly distinct from a score of 25 (someone who likely rounded). Research has observed individual differences in cognitive rounding. For example, in a study of reports of headache frequency, Smitherman and colleagues (2013) found that women were more likely to cognitively heap than men; younger patients were more likely to cognitively heap than older patients; and depressed patients were more likely to cognitively heap than nondepressed patients. Cognitive rounding can be addressed by using shorter time frames (heaping is less likely to occur at lower frequencies, and a shorter time frame will produce lower frequencies); using daily diaries in place of longer recall periods (thereby reducing the recall time to one day but then aggregating entries across many days to better represent behavior proclivities); or making statistical adjustments during data analysis.

As a whole, this discussion clarifies some of the theoretical issues one must consider when framing questions about recall of event frequencies. What time interval should you use in your question to maximize recall accuracy? If you want to study behavioral proclivities, what time interval will allow for an adequate "sampling" of behavior but not be too protracted? Will recall accuracy be maximized by people using episodic strategies (behavior occurs infrequently) or rule-based strategies (behavior occurs frequently)? How can you, or should you, try to encourage one or the other through question phrasing or instructional sets? If respondents are likely to use episodic strategies, should you encourage a think backward or think forward strategy through instructional sets or question wording, or should you say nothing at all and leave it to the respondent to do whatever comes naturally? For longer time periods, should you use "landmarks" and, if so, what landmarks should you use? How will you handle cognitive rounding?

On a more general level, when a question is posed to a person, the verbatim question enters short-term memory and working memory then extracts the gist of the question (Reyna et al., 2009). Relevant information from long-term memory is accessed by the individual based on this gist, coupled with the processing of information in the surrounding context as respondents formulate an answer to the question. Cognitive scientists distinguish two appraisal systems that operate in any given situation (Gross, 2007; Smith & Kirby, 2000). The first system is a cognitive appraisal system, where we interpret the situation we are in, make note of who is present, think about the intentions and orientations of the people who are present, and formulate other cognitions about the context. The second system is an affective appraisal system that alerts us to the emotions, feelings, and affective reactions we are experiencing and that, in turn, may predispose us to act or interpret matters in certain ways. These cognitive and emotional appraisals happen at lightning-fast speeds, often automatically, and sometimes without awareness. These appraisals and the question posed to us form the basis of the informa-

BOX 14.1. Quantifying Love

Early in my career, I (Jaccard) interacted with traditional anthropologists who were skeptical (to put it mildly) of quantitative methods. One of them would repeatedly and defiantly say "let's see you quantify love!" It turns out that numerous theories of love have been offered by social scientists (e.g., Hatfield & Walster, 1978; Lee, 1973; Rubin, 1970; Sternberg, 1997). Sternberg (1997) posited a theory of love with three major components: passion, intimacy, and commitment. Passion is associated with physical arousal and emotional stimulation, whereas intimacy is associated with feelings of closeness and attachment; commitment is associated with a conscious decision to be together over the long run. According to Sternberg, different types and stages of love are represented by different combinations of these components. One couple might be high in passion and high in intimacy, but low in commitment; another couple might be low in passion and low in intimacy, but high in commitment. As part of his research program, Sternberg developed three scales, one for each component, consisting of 15 items per scale. Each item was rated on a 9-point disagree–agree scale. Total scores on each subscale were based on a sum of the items. Sample items for passion are "Just seeing _____ excites me"; "My relationship with _____ is very romantic"; "I find _____ to be very personally attractive"; "I especially like physical contact with _____"; Sample items for intimacy are "I have a warm relationship with _____"; "I communicate well with _____"; "I feel close to _____"; "I feel that I really understand _____." Sample items for commitment are "I know that I care about _____"; "I am committed to maintaining my relationship with _____"; "I have confidence in the stability of my relationship with _____"; "I could not let anything get in the way of my commitment to _____."

Suppose you want to describe love for couple members in the United States with a specific focus on these three dimensions. You might decide you need a random sample of 1,000 couples to have a reasonable representation of the population. A qualitative study of that magnitude would be a massive undertaking, especially if you wanted to compare subgroups on the dimensions based on gender, ethnicity, age, and social class. Granted, a summary score for each of the three components is limited. But if primary interest is with these three dimensions per se, the quantitative approach is not unreasonable for a study of that magnitude. Sternberg went to great lengths to establish the reliability and validity of these measures.

Interestingly, Sternberg later altered his measurement approach (Sternberg, Hojjat, & Barnes, 2001). He noted that people typically are exposed to large numbers of diverse stories about love, either love stories by and of themselves or love stories embedded in larger stories. The stories come from observing people in relationships, experiencing one's own relationships, watching television and movies, reading novels, and so on. Based on these stories, Sternberg reasoned, people create their own stories or narratives about what love is or should be. Potential partners may "fit" a person's personal stories to a greater or lesser degree. Sternberg felt that relationships

(continued)

might be more stable and satisfying if there was a match in the personal love stories of the couple members. Based on extensive qualitative interviews and content analyses of media, Sternberg identified 25 kinds of stories that people might have about love. These stories encompassed the three dimensions of passion, intimacy, and commitment, but they were far more nuanced and qualitatively rich. Sternberg developed an assessment tool that asked individuals to rate how much they embraced themes in each of the 25 stories. He acknowledged that story content and relevance would be culturally dependent, and he encouraged research to explore this position, leading to cultural specific assessments. In your opinion, does this approach to measuring love have merit? Can we measure love or are my anthropologist friends correct?

tion we access from long-term memory for purposes of constructing an answer. Note our use of the word *constructing*: Answer formulation is a constructive process; it is not a simple process of finding the relevant information in long-term memory and passively reporting it. People actively construct answers based on their appraisals and on the information they have in working memory. When we theorize about measurement, we find it helpful to think about the cognitive and affective appraisals respondents are likely to make and the type of information likely to be accessed from long-term memory when formulating an answer to a question.

Elements of the above were evident in our prior analysis of count recall. By thinking through each question in this kind of depth, you will be better able to construct a measurement theory of judgment to guide effective question design. Research has elaborated many factors that can potentially impact accurate recall. Relevant variables include, among others, (1) cognitive abilities; (2) age; (3) mood; (4) stress and anxiety; (5) attention; (6) the salience, vividness, and distinctiveness of the information or events to be recalled; (7) the frequency and recency of exposure to the information or events to be recalled; and (8) factors that promote confusion with other information/events. The importance and relevance of these factors for accurate recall can vary as a function of the facets of measurement in your study (e.g., the population you are studying, the assessment context, the substantive topic). Of course, often you will be interested not in maximizing the accuracy of recall but rather in the judgments people make about prior, current, or future events per se independent of accuracy. Generally, as you structure questions, a thoughtful analysis of the type of judgments and judgment processes you are activating will ultimately help you frame questions and potentially interpret answers to those questions.

Constructing a Theory of Response Translation

Once an individual formulates an answer to a question, he or she needs to convey that answer to the researcher or interviewer. In quantitative research, this task is often

accomplished using rating scales. Rating scale formats are useful if we intend to process data for large numbers of people in multivariately complex ways for purposes of theory testing or describing populations. Rating scales are foreign to many people. People must learn how to use them and forge a strategy for translating their judgments onto them. Two people may make identical cognitive judgments but give different answers on the rating scale if they interpret and use the rating scale differently. This is a potential source of measurement error and needs to be addressed. How we accomplish accurate response translation requires a theory of the response translation process. Again, theory construction and measurement go hand in hand. In this section, we focus primarily on rating scales. We consider issues of metric precision, anchoring, use of adverb qualifiers, the problem of response satisficing, practice effects, and the identification of "mischievous" responders. All focus on the fundamental process of response translation.

Metric Precision

An important distinction in measurement theory and statistics is that between a *discrete variable* and a *continuous variable*. Often, the number of values that a variable can assume is relatively small and finite, such as the number of people in one's family. Such variables have a finite number of values that can occur between any two points. For example, consider the number of people who donate blood at a blood drive during the first hour of the drive. Only one value can occur between the values of 1 person and 3 persons, namely, 2 persons. We do not think of there being 1.5 or 2.7 persons. Variables for which only a finite number of values can occur between any two points are called discrete variables. In contrast, a continuous variable can theoretically have an infinite number of values between any two points. Reaction time to a stimulus is an example of a continuous variable. Even between the values of 1 and 2 seconds, an infinite number of values can occur (1.870 seconds, 1.8703 seconds, 1.87035 seconds, and so on). Many measures we use in the social sciences are discrete in character, but they are thought to reflect an underlying continuous construct. Satisfaction with a product might be measured on a 7-point scale ranging from very dissatisfied to very satisfied, but the construct it is thought to reflect (satisfaction) is continuous in character.

When measuring agreement with a statement on an opinion survey, the underlying construct of agreement is continuous, but researchers might use a different number of discriminations to assess agreement. Some researchers might use a 2-point metric (0 = disagree, 1 = agree); others might use a 3-point metric (0 = disagree, 1 = neither, 2 = agree); and still others might uses a 5-point metric (1 = strongly disagree, 2 = moderately disagree, 3 = neither, 4 = moderately agree, 5 = strongly agree). The number of categories/discriminations of a measure refers to the *precision* of that measure, with more categories being more precise. More precise measures have the advantage of better identifying individuals who truly differ in their opinions or judgments, while less precise measures can artificially lump together people who are meaningfully different into the same measurement category. A 2-point agree–disagree scale lumps into the same category and treats as the same people who strongly disagree with a statement, people who moderately disagree with the statement, and people who only slightly disagree with

it. Such "lumping" can lead to misleading inferences. For example, consider the concept of behavioral intent. People who only slightly agree with the statement "I intend to vote in the upcoming presidential election" often behave differently in their voting behavior than those who strongly agree with this statement. As a result, many social scientists conceptualize behavioral intentions as a continuous construct and measure such intent using more precise metrics than 2-point scales (see Fishbein & Ajzen, 2010).[1]

Simulation studies that have addressed issues of scale coarseness suggest that five to seven categories often are enough for many empirical applications. For example, in a classic study focused on Pearson correlations, Bollen and Barb (1981) created on a computer a very large "population" of individuals where the true population correlation between two continuous variables was either 0.2, 0.6, 0.8, or 0.9. Bollen and Barb then created coarse measures from the continuous measures for each population by breaking the continuous measures into anywhere from 2 to 10 categories. For example, a normally distributed continuous variable that ranges from –3 to +3 can be turned into a 2-point scale by assigning anyone with a score of 0 or less a "0" and anyone with a score greater than 0 a "1." Bollen and Barb computed the correlations using these "coarse" measures and examined how close they were to the case where the correlation was computed using fully continuous metrics. They found that the true correlations were relatively well reproduced by the coarse measures as long as the coarse measures had five or more categories. For example, the reproduced correlations for five-category measures were within about 0.06 correlation units of the continuous-based correlations when the true correlations were 0.60. Bollen and Barb concluded that five categories were probably sufficient for many applications. This recommendation has been replicated in many other studies using different analytic contexts (although some research suggests seven or more categories may be best in some scenarios; see Green, Akey, Fleming, Hershberger, & Marquis, 1997; Lozano, García-Cueto, & Muñiz, 2008; Lubke & Muthén, 2004; Taylor, West, & Aiken, 2006). Thus, coarse measurement of continuous constructs is not necessarily problematic unless it is very coarse. Having said that, the requisite precision needed for a measure is dependent on the research question and context.

In the literature on question design, you will encounter conflicting statements about the relationship between precision and reliability, with some methodologists suggesting that more precise measures lead to lower reliability and other methodologists suggesting the opposite. Our recommendation is to think through the needed precision for the substantive questions you are addressing (which usually will be five or more discriminations) and then to use assessment practices that maximize reliability relative to that level of precision (see Chapter 13 for a discussion of such practices). Sometimes, the construct you study will require only a few discriminations, such as whether a person voted or whether a person purchased a product. However, if your construct is continuous, then you want a reasonably precise measure. If you work with populations where rating scales are not viable (because of literacy issues), with some ingenuity, you can still ask questions orally in ways that yield precise answers. For example, one might

[1]Sometimes, greater precision leads to lower reliability as individuals grapple with having to choose from among many discriminations. Later in this chapter we discuss strategies for circumventing this dilemma.

avoid the use of rating scales altogether and orally ask a question in two steps. At Step 1, you ask if the person "agrees" or "disagrees" with a statement or concept. If the person states "agree," you follow up by asking, "Do you strongly, moderately, or only slightly agree with it?" If the person states "disagree" at Step 1, you follow up by asking, "Do you strongly, moderately, or only slightly disagree with it?" When the two steps are combined, the result is a 6-point agree–disagree metric.

Anchoring

Cognitive judgments often are impacted by *cognitive anchors,* namely, a reference point against which judgments are made. At auctions, the opening bid is an anchor or standard against which later bids are evaluated. In negotiations, the first position stated becomes an anchor for counterpositions. Rating scales typically have endpoint descriptors that serve as anchors for how one uses the scale. One commonly used rating format is a *visual analog scale* (VAS). A VAS is a horizontal line, usually 100 millimeters in length, anchored by verbal descriptors. There are many variants, but here is a common example used in pain research to rate experienced pain associated with an event (e.g., a medical condition):

Respondents are instructed to mark a point on the line that best describes the pain they experience. A score between 0 and 100 is assigned based on the number of millimeters from the left that the mark is made. (*Note:* This example is not drawn to scale.) Pain researchers almost always use "no pain" as the left anchor, but they vary the descriptor for the right anchor. Usually, the maximal anchor is extreme to avoid ceiling effects (i.e., everyone marking the upper end of the scale), but not too extreme to the point people can't comprehend it. Research finds that as the maximal anchor becomes more extreme, pain ratings decrease (Seymour, Simpson, Charlton, & Phillips, 1985), which makes sense because with an extreme right anchor, individuals essentially are pushed away from the right extreme of the scale when making their ratings.

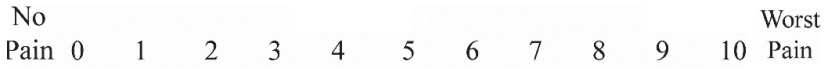
Qualitative studies have examined the painful events that people mentalize as maximum pain anchors. These tend to vary across individuals (de Williams, Davies, & Chadury, 2000). For example, adult females tend to use events associated with childbirth, whereas men use events associated with injuries (Robinson et al., 2004). Some people imagine events they think would be painful, while others recall a painful event that they experienced. These individual differences are important because two individuals can make the same cognitive judgments of pain but will translate it differently onto the rating scale if the mental representation of the maximum anchor is not the same, with more extreme representations lowering pain ratings. Indeed, the same individual may use different anchors at different time points in a longitudinal study, artificially producing change in pain ratings when no true pain change has occurred. Good psycho-

metrics provides individuals with common anchors by defining anchors and providing referents. This is true for most rating scales, not just VASs.

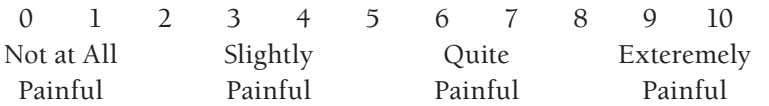
Use of Adverb Qualifiers

Sometimes a VAS will include numerical demarcations on the line as reference points, as follows:

Although this format reduces precision, it can enhance reliability and equal interval properties of the metric as people have clear demarcations to indicate their marks on, and hopefully, they apply numerical properties to the scale points (such as equal intervals between numbers). Other researchers eliminate the line altogether, as in the following:



People circle the number that best captures their judgment. The use of numbers assumes some degree of numeric literacy. To reduce such reliance, some researchers add *adverb qualifiers* to the scale at different points:



The idea is that individuals will first orient themselves to the adverb qualifier that best captures their judgment and then circle a number above that qualifier, but with the flexibility to rate somewhat lower or somewhat higher by circling a nearby number instead. Indeed, instructional sets are often given to use the scale in exactly this fashion. Numerous studies and meta-analyses have supported the addition of verbal descriptors to numerically labeled scales to increase reliability and validity (Bendig, 1953; Rodgers, Andrews, & Herzog, 1992; Saris & Gallhofer, 2007). Finally, some researchers eliminate the numbers to rid numeracy from the mix but at the cost of losing precision, as in the following:

____ Not at all painful

____ Slightly painful

____ Quite painful

____ Extremely painful

A classic scale format that also does not use numbers is called the *semantic differential* (Snider & Osgood, 1969), which appears as follows for rating a political candidate:

Candidate A

bad	____:____:____:____:____:____:____	good
weak	. . . . . .	strong
passive	____:____:____:____:____:____:____	active

The labels *extremely, quite,* and *slightly* are used for the three leftmost categories and *slightly, quite,* and *extremely* for the three rightmost categories as well to create symmetry. The middle category is labeled *neither or both*. Instructional sets usually are given that link these adverbs to the categories (see Snider & Osgood, 1969), but sometimes the adverbs are written below each category. Sometimes both labels and numbers (e.g., –3 to + 3) are associated with each category. Many different combinations are used in practice.

When choosing adverbs to use in rating scales, one should use adverbs that connote roughly equal intervals across the underlying dimension. For example, when rating the importance of each of several factors that entered into a decision, a commonly used format for each rating is:

____ Very important

____ Somewhat important

____ Not very important

____ Not at all important

Note that there seems to be unequal "psychological spacing" between these qualifiers. The difference between "not very important" and "somewhat important" seems slight compared to the difference between "somewhat important" and "very important." The choice of these particular adverbs creates an ordinal metric. There are large literatures in psychometrics that can guide the choice of adverbs to produce roughly equal intervals across the underlying dimension (Beckstead, 2014; Rohrmann, 2015). For example, using psychophysical scaling methods, Cliff (1959) found that describing something as "slightly good" is generally perceived to be about 0.50 times as "good" than the simple, unmodified "good." Adverbs can be selected based on such qualifying values to approximate equal intervals (taking into account, as well, linguistic distinctiveness and comprehension of the adverb). Rohrmann (2015) presents adverb analyses in English, German, and Chinese for dimensions of frequency (e.g., never, seldom, sometimes, often, always); intensity (not at all, a little, moderately, quite a bit, very much); probability (certainly not, unlikely, about 50:50, likely, for sure); quality (bad, inadequate, fair, good, excellent); and agreement (fully disagree, mainly disagree, neutral, mainly agree, fully agree). Beckstead (2014) summarizes research on qualifying values for frequency judgments and magnitudes. We do not recommend interpreting qualifying values in these reports

as strictly applicable to your research because studies have shown that qualifying values can vary as a function of the different measurement facets (McClelland, 1975). However, published studies such as these can serve as rough guidelines. Coupled with common sense that is sensitive to creating equal psychological differences between categories as well as proper cognitive response testing, reasonable adverb choices can be made to produce approximately equal interval metrics.

As an aside, rating scales can be unipolar (from not having a property to having much of it) or bipolar (rating an object on polar adjective opposites, such as sad–happy, dominant–submissive). Here is an example of a bipolar format:

–5	–4	–3	–2	–1	0	1	2	3	4	5
Very Sad		Moderately Sad		Slightly Sad		Slightly Happy		Moderately Happy		Very Happy

For bipolar scales, it is important that the adjectives are, in fact, polar opposites. Research suggests that formal antonyms are not always perceived as psychological antonyms (Yorke, 2001). For example, some perceive the antonym pair calm–angry as not constituting polar opposites. Early research on prototypical personalities treated masculinity and femininity as polar opposites, but this conceptualization was later rejected with the introduction of the concept of androgynous personality styles (namely, having both male and female qualities; see Lubinski, Tellegen, & Butcher, 1983). In addition, use of negative numbers in bipolar scales implies numeracy, which may be problematic. One can address this by eliminating numbers and using just the label "neither or both" for the midpoint, yielding a 7-point metric.

Satisficing

Herbert Simon was awarded the Nobel Prize in economics in 1978 for his extensive work in decision making. Earlier, in 1957, he coined the term *satisficing* (a combination of the words "satisfy" and "suffice") to refer to decision contexts where instead of carefully thinking about all available decision options, people only do what is sufficient to obtain a satisfactory result. For example, in choosing a bank, a person may open an account at the first bank that seems satisfactory rather than explore the merits of bank after bank. In the context of questionnaires, satisficing refers to the tendency to choose the first acceptable response option for a question because doing so requires the least effort.

Satisficing can take different forms. If a consumer psychologist asks people to rate different ice creams on 0 to 10 scales, a person might just rate them all 8 because he or she likes ice cream in general. Galesic and Yan (2011) used eye-tracking technology to follow people's eye movements as they read items on questionnaires. They found that about 10% of survey takers never looked at the last two response options in a 12-category list of preferred products and that they spent far more time looking at options in the first half of the list. Strategies for dealing with satisficing include (1) use of instructional sets to encourage conscientious responding, (2) keeping questions or response options short and simple, (3) splitting complex questions that have many response options into

multiple questions, and (4) including checks to diagnose satisficing. For the latter, for example, Adams and colleagues (2006) included the following item among 42 items rated on 5-point disagree–agree scales: "We use this question to discard the survey of people who are not reading the statements. Please select option 4 to this item." Individuals are more likely to satisfice if they feel time pressure to complete the task, if they find the task difficult, or if they are bored.

Practice Effects and Scale Comprehension

As noted, your population may have little experience with rating scales, in which case it may take time for them to become comfortable with the scales. To remove these warm-up effects, we find it helpful, where feasible, to include a few practice items. After the warm-up items, we also include a few items that serve as comprehension checks for rating scale use. These are items where we know what response the individual should make if he or she understands the rating scale properly (e.g., a rating on a good–bad scale of something that is obviously good). If the correct response is not given, we know we must review the rating scale instructions with the respondent. Finally, we have found that people sometimes object to frequent shifts in scale formats. Such changes may be necessary, but wherever possible, we try to use one format (e.g., a 5-point agree–disagree scale with adverb qualifiers) for most items. Doing so also limits the number of anchoring examples and practice tasks.

Mischievous Responders

Research has documented the existence of what are known as mischievous responders (Robinson-Cimpian, 2014). These are individuals who comprehend a question, make a valid judgment in their minds, but then deliberately report a false judgment, often in outlandish ways, in order to be "mischievous." Fish and Russell (2018) documented an example in the National Longitudinal Study of Adolescent Health (Add Health). They noted that an unusually high number of middle and high school youth in the study had reported they were not heterosexual. In addition, the number of students who reported being nonheterosexual but changed their answer to being heterosexual in a one-year follow-up interview was unusually high. Fish and Russell developed a mischievous index based on methods suggested by Robinson-Cimpian (2014) that used response patterns to 10 low base-rate items unrelated to sexual identity within Add Health. If people responded affirmatively to a large number of these items, it raised the possibility that they were being "mischievous." Using this index and response triangulation from multiple questions about sexual identity, Fish and Russell identified those individuals who were engaging in mischievous responding. They then used this information as a covariate in analyses of health disparities between heterosexual and nonheterosexual youth. They found evidence for only a small proportion of "mischievous" youth, but, importantly, one of the five health disparities they had documented in prior analyses became statistically nonsignificant when the mischievousness index was included as a covariate. Mischievousness can probably best be counteracted through the use of instructional sets

that discourage it and that encourage honest responding and also through including items that will detect it in the survey.

Concluding Comments on Response Translation

Many types of rating scales are used in social science research that allow people to translate their judgments and opinions into brief, quantitative formats that can be analyzed effectively. You need to theory construct about ways to facilitate that translation process for your particular research domain. We have provided general guidelines to stimulate your thinking, and these are summarized in Figure 14.2. This figure, like Figure 14.1, is a simplification because it ignores potential moderated and interactive relationships between constructs. Inevitably, some considerations will be unique to your substantive domain and your chosen scale formats.

Concluding Comments on Theory Construction for Self-Reports

In sum, when using self-reports in a research project, it is helpful to adopt a theory construction mindset as you think about the measures using three processes as a theory generation frame: comprehension, judgment, and response translation. We make it a practice to theory construct around these processes, explicitly taking into account the particular measurement facets of our study (who the population is, what the assessment context is, etc.). We engage in such theory construction not only for measures we develop but also for extant scales that purportedly have solid psychometric histories. Sometimes those "histories" do not hold up when they are subjected to rigorous analysis. More often than not, we find ourselves engaging in preliminary psychometric studies designed to provide insights into the best way to structure the measures we will use in our main study.

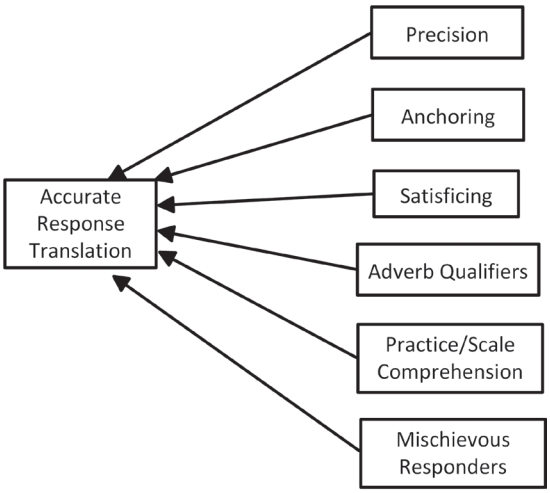


FIGURE 14.2. Factors impacting response translation.

CONSTRUCTING A THEORY OF OBSERVER REPORTS

Many studies do not rely on self-reports but instead use trained observers to record information (e.g., make behavioral counts) and/or to make judgments about people's behavior (e.g., rate a child's behavior in a playground on a rating scale reflecting how aggressive the behavior is). There are many types of observational research contexts in the social and health sciences. Sometimes observations are made in natural settings and other times in laboratory settings. Sometimes observers are hidden from view, and sometimes research participants are aware of their presence and can visually see the observer. Sometimes observers code behaviors using dichotomous formats (e.g., the behavior performed expressed positive affect or it did not), and sometimes the behaviors are rated on scales or are subject to some other form of quantification. Sometimes observers make judgments about a person's behavior, and sometimes they make inferences about psychological states (e.g., anxiety). Sometimes observations are made about people, and sometimes they are made about settings, such as the quality of a child's home environment. In our view, self-reports are a form of observer reports; it is just a matter of who the "observer" is—the person being studied (self-report) or someone else (observer report).

External observers fundamentally use the same processes that we have described for self-reports. First, observers must have a clear understanding of the behaviors/constructs they are to observe (comprehension). They then must make judgments about the person's behavior relative to those constructs (judgment). Finally, they must communicate those judgments to the researcher by translating them onto a rating scale or some other response format (response translation). As such, the theory construction frame we described for self-reports applies with equal vigor to observer reports. For example, most factors we have identified as determinants of question comprehension for self-reports also impact the comprehension of instructions to observers about what they are to observe and how they are to make their observations. If observers use ratings scales, the factors we have described for response translation for self-reports also apply to observers, namely, the need to have precise metrics; to use well-defined anchors; to use adverb qualifiers strategically selected to approximate equal interval metrics; to implement practices in order to increase observer task motivation and eliminate satisficing; and to deal with potential practice and warm-up effects.

Developing effective measurement strategies using trained observers means you need to theorize about observer comprehension, judgment, and response translation. Theorizing about these processes will inevitably invoke consideration of characteristics of (1) the observers, (2) the research participants, (3) the research context, (4) the behavioral context, and (5) the judgments to be made. These five categories form a starting point for your measurement theory construction for observer reports.

One important difference between self-reports and observer reports lies in the kinds of factors that impact the cognitive judgments made by observers. The literature in organizational psychology suggests potential biases on the part of supervisors when judging the behaviors of employees and, although made in slightly different contexts, these biases are also of potential relevance for observer reports more generally. The

biases include (1) a *halo effect,* a bias whereby the overall impression of a person ("he or she is very good") impacts evaluations of the person's performance independent of actual performance; (2) a *horn effect,* a bias that is the same as the halo effect but is focused on an overall negative impression, leading one to evaluate behaviors more negatively; (3) a *central tendency bias,* the tendency to judge everyone as average; (4) a *leniency bias,* the tendency to "go easy" on people by judging most everyone positively; (5) a *strictness bias,* the tendency to "be hard" on people by judging most everyone negatively; (6) a *contrast effect,* the tendency to see someone who stands apart from others as more different than they actually are, such as judging a person to be a very good performer if that person is surrounded by people who are poor performers; (7) an *assimilation effect,* the tendency to see someone who is slightly different from others as being more like those others than is actually the case, such as judging a person to be a poor performer if that person is surrounded by people who are poor performers; (8) a *recency bias,* the tendency to judge performance based on the most recently observed performances rather than the total performance; (9) a *primacy bias,* the tendency to judge performance based on the initially observed performances rather than the total performance; (10) *stereotyping,* the tendency to allow stereotypes of the groups a person is a member of, such as gender and ethnicity, to impact judgments of that person; and (11) an *expectancy effect,* the tendency to base judgments on what the observer expects to happen. Which of these biases might be relevant to your research that uses observers? What other biases may be relevant for your research project that are not listed here? What is your theory of judgment bias? When you are training observers, how might you address such biases and eliminate their impact to improve measurement validity? The operation of these biases may differ depending on characteristics of the observer, the target person, the research context, the behavioral context, and the judgment dimensions. As you plan your measurement strategy, think through these theoretical facets.

A concept of some importance in studies using trained observers is *observer drift* (Smith, 1986). Most definitions of drift state that it occurs when observer understandings of the behavioral codes or coding criteria they are to use change over time. Other researchers use the term *drift* to refer to such changes between the time observers finish training and the time they begin observing. In either case, it is important to ensure that observers maintain consistency across time, that is, to theorize about factors that might impact drift in one's research.

Sometimes observers are used who have not undergone formal training in the target observational tasks. For example, teachers might be asked to rate or make judgments about individual students for research purposes, and parents might be asked to rate or make judgments about their children. In such cases, one must be particularly sensitive to observer bias. Perhaps one can ask questions of these observers or use instructional sets in ways that minimize bias. As stated earlier, lack of convergence of reports by these different "observers" may not reflect observer bias, but instead may indicate that the observers have access to different behavior samples of the target person. Parents who are asked to rate the anxiety levels of their children see their children in different contexts than teachers do. Disparities in ratings between multiple observers often are regarded as problematic, but in some cases, they are meaningful.

Just as cognitive response testing can be used to good effect for self-reports (see Chapter 13), it also can be used for observer reports. After observers have been trained and execute a few observations in a training session, verbal probes about the interpretation of behavioral codes, judgment strategies used, and response translation can be helpful in building your measurement theory and provide insights into factors that need to be addressed.

CONSTRUCTING A THEORY OF "OBJECTIVE" MEASURES

In addition to self-reports and observer reports, another measurement approach uses "objective" measures that directly indicate one's standing on a construct. For example, there are biomarkers to indicate if a woman is pregnant, if a person has cancer, viral loads for HIV, if a person has consumed alcohol in the past 30 days, and the amount of physical exertion during an activity, to name but a few. Formal records of people's text messaging or search patterns on the Internet also do not require self-reporting or direct observer reports. Examination of the accuracy of "objective" measures like these often reveals the possibility of bias; there can be false positives or false negatives in them, that is, measurement error. For example, urine tests for opioid use can yield a false positive if the person taking the test consumes a large amount of poppy seeds prior to the test. False positive urine tests for some illicit drugs also can occur if people taking the test consume cold medications containing pseudoephedrine. False negatives for urine tests of illicit drug use can occur through tampering (e.g., adding water to dilute the sample or adding soap to the sample), taking diuretics, or drinking large amounts of water before the test. A urine test for cannabis use will test positive for 1 to 3 days after a single use of marijuana, but not longer, indicating the time-restricted nature of these tests. Given that "objective" measures are subject to measurement error, it follows that a theory construction mindset surrounding that error is required. Accurate use of biomarkers for research purposes often requires considerable expertise.

"Objective" measures are not always sensitive to the behavioral dimension of theoretical interest to a social scientist. For example, there is no biomarker of the number of days in the past 30 days that someone has smoked marijuana, so if one is interested in understanding the frequency or patterns of marijuana use, biomarkers are of little help. A theory of "objective" measurement also should articulate such boundary conditions.

Researchers often use official records as "objective" measures of constructs, such as school records to document grade point averages of students, arrest records for drunk driving or some form of criminal activity, and death certificates to index mortality. In some cases, these indices will contain relatively little measurement error, but in other cases, this will not be so. For example, many people drive drunk but never get arrested for it. Official arrest records for drunk driving are thus poor measures of it. As another example, in 2015, official death certificates in the United States were found to overlook more than half of the people killed by police (Feldman, Gruskin, Coull, & Krieger, 2017), indicating a source of bias in using such certificates as an index of mortality rates.

Scientists are skeptics, and it is best to approach any "objective" measure with a mindset of potential fallibility. A theory construction mindset of potential sources of error and how to rectify them is as applicable to "objective" measures as it is to self-reports and observer reports.

THE IMPORTANCE OF A THEORY CONSTRUCTION MINDSET TO MEASUREMENT

In sum, as you approach measurement matters in your research, we strongly urge you to adopt a theory construction mindset, namely, strive to construct theories of measurement for every question you ask or every measure you use by making use of the thinking strategies and heuristics elaborated in the current and previous chapters. As you evaluate measures for possible use, bring that measurement theory to bear. To the extent that important psychometric issues are theoretically unresolved, you may need to conduct preliminary research to gain perspectives on them before you move forward with your main study. If you encounter poor psychometric practice for a standard scale (e.g., the scale fails to provide adequate anchors, it does not use appropriate adverb qualifiers, or it pays inadequate attention to precision), consider altering the measure if you believe the changes will improve its psychometric properties.

As per our discussion of measurement theory in Chapter 13, the theories you construct about self-reports, observer reports, and objective reports can be subjected to empirical tests and published in scientific journals for other researchers to benefit from. However, as noted in Chapter 13, we usually do not go about the task of empirically testing every core (untested) theoretical expression in our measurement theory because doing so would sidetrack us from our main purpose of building substantive theory in our main study. We instead rely heavily on common sense and past measurement research for measure evaluation, but we also invariably end up conducting a preliminary measurement-oriented research project to address measurement issues (and to test some of the theoretical expressions we generated) to improve the primary study.

MEASUREMENT AND QUALITATIVE RESEARCH

Our discussion thus far has emphasized quantitative measurement, but the issues of comprehension, judgment, and response translation apply just as strongly to qualitative research, either when asking people questions, when making your own observations, or when training others to make observations for you. In qualitative research, as in quantitative research, it is important that the questions we pose to people are comprehended. In addition, if we have some sense of the cognitive and affective processes people use when processing and answering our questions, or if we can discover what those processes are, we might be able to better structure questions and probes to yield answers that provide a deeper understanding of the phenomena we are studying. Finally, it is important for qualitative researchers to appreciate the difference between

people's understanding of their environment, as represented by the concepts they have in mind, and their description of that environment per se, as represented by the symbols or words they use to describe their thoughts, that is, response translation.

Qualitative researchers, like quantitative researchers, probably can benefit from the use of cognitive response testing of their interview-based questions to ensure that their questions will be understood properly when posed to informants. As well, many of the factors discussed above that impact comprehension, judgment, and response translation directly apply to qualitative research contexts. For example, how might issues of statisficing and task motivation be addressed when questions are posed in unstructured interviews in qualitative research? Are questions phrased in ways that encourage bipolar or unipolar responses, and if bipolar, are the opposites truly opposites from the respondent's perspective? Is it necessary to provide examples of "anchors" for people to establish reference points for certain open-ended questions? What should those anchors be? A theory construction mindset for measurement is as relevant to qualitative researchers as it is to quantitative researchers.

Qualitative researchers often engage in a strategy known as *informal interviewing* (Bernard, 2017). For example, Connolly (1990) studied children who live, eat, and sleep on the streets in Guatemala City by just hanging out with them and talking with them informally during his everyday interactions to learn about their lives. At the end of each day, Connolly wrote extensive field notes based on his observations and his conversations with the children. Our discussion of theory construction surrounding observer reports directly applies to this scenario.

Unstructured interviewing also is common in qualitative research. Such interviews have a different quality than informal interviewing because the researcher typically sits with the informant and asks specific questions. To be sure, unstructured interviews are "structured" in the researcher's mind in that he or she invariably has certain goals that are constantly salient during the course of the interview. However, the general spirit of unstructured interviews is to get people talking about a topic and then stay out of the way as they express their thoughts, feelings, recollections, and hopes.

Key to unstructured interviewing is the act of *probing*, that is, knowing when, how, and how often to ask for clarification and elaboration, or knowing how to re-focus interviewees to get them back on topic. Is there a "theory of probing" you can use to help you approach this important facet of unstructured interviewing? If you were to construct a "theory of probing" for your particular research project, what would it be? Are there different types of probes one can use? What is the typology describing probe differences? Are there nonverbal probes in this typology (e.g., the nod of one's head, a look of surprise)? Are some types of probes better in some contexts and for certain types of people and certain topics? How do you effectively probe without interjecting your own views into the interviewee's thinking? How do you know people are quiet because they are reflecting on the topic at hand as opposed to being quiet because they have finished expressing their thoughts? Are the three processes of comprehension, judgment, and response translation relevant to probes?

Although formal measurement theory has been dominated by psychologists, anthropologists have offered important insights into such theories as well. Instead of thinking

of a measure as a static "object" or "thing," anthropological perspectives often think of the act of completing a measure as a behavior in its own right that is subject to the same types of ethnographic-based descriptions and explanations as any other behavior (Hubley & Zumbo, 2017; Maddox, Zumbo, Tay-Lim, & Qu, 2015). Theorizing is not restricted to measurement principles brought to bear per traditional psychometric theory; it also expands such analysis to include broader situational, cultural, and ecological facets pertinent to assessment. Would doing so help in your theory construction efforts?

SUMMARY AND CONCLUDING COMMENTS

Testing creative and novel theories often requires that you develop your own measures of constructs. These measures can take the form of self-reports, observer reports, or "objective" measures. When constructing (or evaluating) measures, you need to invoke a measurement theory. As such, theory construction at the level of measurement is important.

Three fundamental processes are involved when people provide self-reports: (1) they must comprehend the questions you are asking, (2) they must form answers and judgments relevant to those questions, and (3) they must report those answers and judgments to you. There is considerable room for theory construction surrounding these processes of comprehension, judgment, and response translation. Question comprehension is potentially impacted by characteristics of the population you are studying, the structure and nature of the questions you are asking, and the broader assessment context in which the questions are asked. Factors within each of these domains that may come into play are summarized in Figure 14.1. These factors may be nuanced and augmented depending on your substantive application and research questions and, as such, you need to theory construct relative to them. Judgment and answer formulation ultimately involves cognitive and affective processes that operate in working memory of individuals. The content of working memory is determined by a person's cognitive and affective appraisals of the assessment context and questions being posed as well as information retrieved from long-term memory, which is then used to construct one's answers. Careful analysis of these judgment processes can often help researchers frame questions in ways that will maximize their information yield, reliability, and validity. Response translation takes many forms, but a ubiquitous format is that of the rating scale. Rating scales take many different forms (e.g., requiring unipolar versus bipolar judgments), and how people interpret and use rating scales is critical. Two people may make identical judgments cognitively but may differ in the responses they make on a rating scale, depending on scale interpretation. Similarly, individuals may make different cognitive judgments but give the same response on the rating scale depending on scale interpretation. Factors that impact rating scale interpretation and utility include metric precision, anchoring, the choice of adverb qualifiers, the use of satisficing response strategies, and scale familiarity. As you approach measurement in your research, you will need to theory construct about judgment processes and response translation, just as you theory construct about question comprehension, and then, based on these theories, formulate

optimal measurement strategies to achieve your broader research goals. Measurement is complicated, and so it demands high-quality theory construction to do it right.

Observer reports typically use trained observers to record behavioral observations of a target person or to make judgments about the person's observed behaviors. External observers fundamentally use the same processes described for self-reports. First, observers must have a clear understanding of the behaviors/constructs they are to observe (comprehension). They then must make judgments about the person's behavior relative to those constructs (judgment). Finally, they must communicate those judgments to the researcher by translating them onto a rating scale or some other response format (response translation). As such, the theoretical issues addressed for self-reports often apply with equal vigor to observer reports. However, the observers' biases can impact the accuracy of their recorded observations, including halo effects, horn effects, central tendency bias, leniency bias, strictness bias, contrast effects, assimilation effects, recency bias, primacy bias, stereotyping, and expectancy effects. Observer drift, through which observer interpretation of behavioral codes changes over time, also is of concern. "Objective" measures do not require observation in the sense that self-report and observer reports do, and they include such approaches as biomarkers, formal search indices on the Internet, and a wide range of unobtrusive measures or behavioral traces left behind by people. Close examination of these measures often reveals sources of measurement error in them, requiring that scientists maintain a skeptical attitude when considering them as measures to include in their research. For both observer-based and "objective" measures, you will need to theory construct about potential sources of error and then adopt practices to counter them.

SUGGESTED READINGS

Abrams, W. (2000). *Observational research handbook: Understanding how consumers live with your product.* New York: McGraw-Hill; and Yoder, P., & Symons, F. (2018). *Observational measurement of behavior.* Baltimore: Brookes.

—Two useful books on a wide range of issues to consider when conducting observational research.

Aiken, L. (1996). *Rating scales and checklists: Evaluating behavior, personality, and attitudes.* New York: Wiley.

—An informal and engaging writing on rating scales in the social sciences.

Bernard, H. R. (2017). *Research methods in anthropology: Qualitative and quantitative approaches.* Lanham, MD: Rowman & Littlefield.

—Includes chapters that provide practical and thoughtful accounts of measurement principles from an anthropological perspective.

Edwards, A. L. (1957). *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts.

—A dated but incredibly clear exposition of the classic scaling methods of Guttman, Likert, and Thurstone.

Marsden, P., & Wright, J. (Eds.). (2010). *Handbook of survey research*. Bingley, UK: Emerald Group.

—A thorough reference on survey research and questionnaire design.

Oakhill, J., Cain, K., & Elbro, C. (2015). *Understanding and teaching reading comprehension: A handbook*. New York: Routledge.

—Although this book is focused on teaching reading skills to children, it is filled with useful concepts and principles that can benefit comprehension issues in question design.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61,* 576–602.

—An overview of the advantages of conversational interviewing. See also Lavrakas, P. (2008). Conversational interviewing. In *Encyclopedia of survey research methods*. Thousand Oaks, CA: Sage.

Weldon, R. B., Corbin, J. C., Garavito, D. M. N., & Reyna, V. F. (2017). The gist is sophisticated yet simple: Fuzzy-trace theory's developmental approach to individual differences in judgment and decision making. In M. Toplak & J. Weller (Eds.), *Individual differences in judgment and decision making from a developmental context* (pp. 67–84). New York: Routledge.

—A nice introduction to how people extract gists from information and the role of gists in memory and judgment.

Wingfield, A., & Byrnes, D. L. (1981). *The psychology of human memory*. New York: Academic Press.

—A good introduction to the basics of short-term memory, working memory, and long-term memory.

KEY TERMS

comprehension (p. 403)
judgment (p. 403)
response translation (p. 403)
literacy (p. 404)
functional literacy (p. 404)
functional illiteracy (p. 404)
numeracy (p. 405)
short-term memory (p. 405)
working memory (p. 405)
long-term memory (p. 405)
forward–back translation (p. 406)
lexical level (p. 407)
semantic level (p. 407)
standardized interviewing (p. 408)
conversational interviewing (p. 408)
think backward strategy (p. 410)
think forward strategy (p. 410)
cognitive rounding/heaping (p. 412)
discrete variable (p. 415)
continuous variable (p. 415)
metric precision (p. 415)
cognitive anchors (p. 417)

visual analog scale (p. 417)
adverb qualifiers (p. 418)
semantic differential (p. 419)
satisficing (p. 420)
halo effects (p. 424)
horn effects (p. 424)
central tendency bias (p. 424)
leniency bias (p. 424)
strictness bias (p. 424)
contrast effect (p. 424)
assimilation effect (p. 424)
recency bias (p. 424)
primacy bias (p. 424)
stereotyping (p. 424)
expectancy effects (p. 424)
observer drift (p. 424)
informal interviewing (p. 427)
unstructured interviewing (p. 427)
probing (p. 427)

EXERCISES

Exercises to Reinforce Concepts

1. What are the three cognitive processes that underlie self-reports? Characterize each of them.

2. What are literacy and functional literacy? Why is it important to consider them in question construction?

3. What is numeracy? How is it important for question design?

4. What is the role of working memory in the question construction process? How does it work with short-term and long-term memory to affect question comprehension and the formulation of answers to questions?

5. What is the method of forward–back translation?

6. What is the difference between standardized interviewing and conversational interviewing? What are the strengths and weaknesses of each?

7. You are going to ask people to recall the number of times they ate dinner out at a restaurant during the past 3 months. What factors would you take into consideration in framing this question to them?

8. What is the difference between a think forward and think backward strategy? Which one tends to work best?

9. Give some examples of factors that represent systematic measurement error. Name at least three. How might each of them be addressed to reduce their impact?

10. What is the validity of a measure?

11. What is cognitive heaping? How might you deal with it?

12. What is metric precision? In general, what is the minimum number of scale categories you need for a measure of a continuous variable?

13. What is a visual analog scale? Describe some of the different variants of it.

14. What is anchoring? Why is it important psychometrically?

15. How might you choose adverbs for rating scales to help approximate interval-level properties? What principles would you take into account in choosing adverbs?

16. What is satisficing? Why is it important in question design? How can you prevent it?

17. Why is it important to remove practice effects when you are using rating scales?

18. Name five biases that can bias observer reports, and define each of them.

19. What is observer drift? How can you prevent it?

20. Why are the processes of comprehension, judgment, and response translation important in qualitative research?

Exercises to Apply Concepts

1. Find a study of interest to you in the literature and read the section on the primary measures used in that study. Critique the measurement section as if you were reviewing the article for a journal. Write out constructive suggestions to the author for how to improve his or her measurement. If you want to draw on principles from Chapter 13 as well, do so.
2. Locate a copy of a dissertation that collected original data, from either your school library or on the Internet. Role-play that you are on the student's dissertation committee and that the student gave you the section on measurement in the dissertation during his or her *proposal* defense. What recommendations would you make to the student to improve his or her measurement?
3. Pick out an extant measure of a construct you are interested in that uses a rating scale. Find someone you know and do a cognitive response test on the measure using verbal probes. Use the material you learned in this chapter to shape the probes you ask.
4. Find a study that used trained observers to do behavioral observations. Critique the observer report strategies used in the study. How might you improve what the study did?
5. Construct a theory of probing for a research topic and population of interest to you.