1

FROM PRINCIPLES TO MEASUREMENT

Theory-Based Tips on Writing Better Questions

Hart Blanton and James Jaccard

Self-reports are the dominant assessment method in the social sciences and a large part of their appeal is the ease with which questions can be generated and administered. In our view, however, this apparent ease obscures the care that is needed to produce questions that generate meaningful data. In this chapter, we introduce and review basic principles of measurement, which we then use as a foundation to offer specific advice ("tips") on how to write more effective questions.

Principles of Measurement

A Measurement Model

Suppose a researcher wanted to measure consumers' judgments of the quality of a product. Perceptions of product quality cannot be observed directly—perceived quality is a latent, theoretical psychological construct, assumed to be continuous in character, such that it can only be inferred indirectly through observable actions. One such action can be ratings a consumer makes on a rating scale. Suppose consumers are asked to rate the perceived quality of a product on a scale that ranges from 0 ("very low quality") to 6 ("very high quality"). By seeking to quantify product perceptions in this manner—and whether the researcher has realized it or not—a formal measurement model has been embraced. This model is depicted in Figure 1.1.

The rectangle labeled "Q" in Figure 1.1 represents the rating on the 0-to-6 scale. This rating does not, by fiat, reveal "true" quality perceptions of the respondent, which is conceptualized as an unobservable latent construct and represented in Figure 1.1 by the circle with the word "quality" in it. The researcher assumes that the observed "Q" is influenced by true, latent quality perceptions, but that

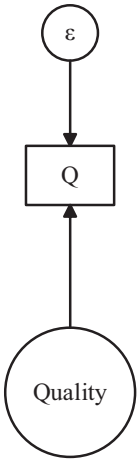

FIGURE 1.1 Measurement Model

the correspondence between latent and observed constructs is less than perfect. Ratings on Q are thus a function of both the consumers' true evaluations and measurement error (represented as "ε" in Figure 1.1). This can be expressed algebraically in the form of a linear model:

$$Q = \alpha + \lambda Quality + \varepsilon \qquad [1]$$

where α is an intercept, λ is a regression coefficient (also frequently called a *loading*), and ε is measurement error. When the relationship is linear, as assumed in Equation 1, then Q is an interval-level measure of the latent construct of perceived quality. If the relationship is non-liner but monotonic, Q is an ordinal measure of the latent construct. Articulation of this formal model focuses attention on one of the primary challenges facing researchers who wish to create self-report questions—the need to reduce the influence of error on observed ratings. We next consider two sources of error, random and systematic, as well as their implications for characterizing the reliability and validity of self-report items.

Random Error and Reliability

Random error represents random influences, known or unknown, that arbitrarily bias numeric self-reports upward or downward. Often referred to as "noise," random error can be generated by such factors as momentary distractions, fluke misunderstandings, transient moods, and so on. This form of error is commonplace, but its relative magnitude can vary considerably from one question to the next. As such, it is meaningful to think about the degree to which a given question

or set of questions is susceptible to random error. This represents the concept of *reliability*.

The reliability of observed scores conveys the extent to which they are free of random error. Statistically, a reliability estimate communicates the percentage of variance in the observed scores that is due to unknown, random influences as opposed to systematic influences. Thus, if the reliability of a set of scores is 0.80, then 80% of their variation is systematic and 20% is random. The presence of random error in measures can bias statistical parameter estimates, potentially attenuating correlations and causing researchers to think they have sufficiently controlled for constructs in an analysis, when they have not.

Systematic Error and Validity

Another form of measurement error is called *systematic error*. This source of error often introduces variance into observed self-report items that is non-random; i.e., that is a function of one or more psychological constructs that are something different than the construct of interest. Consider the model in Figure 1.2. Here a researcher hopes to measure both drug use and grade-point average (GPA) via self-report. Each of these constructs are influenced by the true latent constructs that are of interest (as in Figure 1.2), but another latent construct is also exerting influence on the two measures, social desirability.

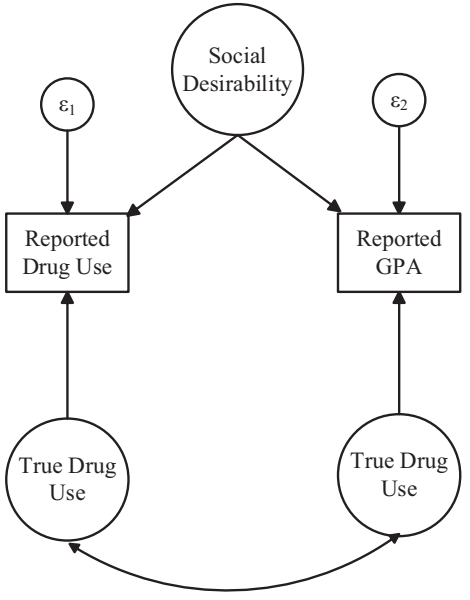

FIGURE 1.2 Example of Systematic Error

The dynamic in Figure 1.2 can arise if those most concerned with projecting a positive image are under reporting their true drug use and over reporting their true GPA. The systematic influence of this "third variable," social desirability, might cause a researcher to overestimate (or underestimate) the strength of the relationship between drug use and academic performance.

Systematic error of the type in Figure 1.2 is a threat to the *validity* of a measure. Ratings on a self-report are valid to the extent that they accurately reflect the construct that is of interest, as opposed to constructs that are not of interest. In the above example, the two measures are partly influenced by the constructs that were of interest (drug use and GPA) but they also are partly influenced by a construct that was not (social desirability), and so the validity of these measures was undermined. In more extreme cases, a measure might be so strongly biased by systematic, confounding influences that it is best viewed as invalid; i.e., it should be viewed as a measure assessing something other than the construct of interest.

Statistical and Methodological Approaches to Measurement Error

One way of handling the presence of measurement error is to embrace modern analytic methods that can correct for biasing influences. Structural equation modeling (SEM) is a particularly useful analytic tool, well-suited to estimating statistical parameters while adjusting for both systematic and random sources of measurement error (Kline, 2016). Incorporated into these analytic approaches might also be attempts to formally measure known or anticipated sources of systematic error ("confounds"), so that their influences can be statistically controlled (or "covaried"). For instance, if a researcher has the concern that social desirability tendencies will influence ratings, a separate measure socially desirability can be administered (e.g., Fleming, 2012; Uziel, 2010), so that its influence on ratings can be formally estimated and statistically controlled during parameter estimation.

The Aggregation Approach to Measurement Error

A common approach to reducing the impact of random measurement error is aggregation. Because one can rarely expect to create a single perfect self-report item that captures all of the variance in a complex psychological construct, researchers often construct multi-item instruments to measure constructs. The logic of aggregation is that, even if a given item is influenced to a consequential degree by random error, different items will be influenced in different and largely idiosyncratic ways. The result is that when ratings on multiple items are summed or averaged, the error in specific items will "wash out" in the aggregate score, resulting in a more reliable and valid estimate of the latent construct of interest. This is a generally sound and accepted practice but there are common misperceptions and misapplications of aggregation. We explore these shortly, but we first

consider the different types of constructs one might wish to assess, and how this might affect aggregation.

Defining the Construct of Interest

Broad versus Narrow Constructs

Some constructs are fairly concrete and easily referenced in self-report items (e.g., age, height, income). With such "narrow" constructs, there will often be little gained from constructing multiple items to represent them and then aggregating the questions because they will yield identical information about a respondent (e.g., "How old are you in years and months?" "How old are you in months?"). In contrast, many concepts that are of basic and applied interest in psychology are by their nature abstract and hard to translate into a single question (e.g., intelligence, depression, social support). With such "broad" or abstract constructs, aggregation can have value, as multiple questions give respondents imperfect but nonredundant ways of expressing their standing.

Breadth versus Dimensionality

A construct might be broad in multiple senses. One way is that it can take a myriad of roughly equivalent, interrelated forms, such that a larger number of items might capture more of distinct ways it can manifest itself, leading to improved measurement. Consider extraversion. Highly extraverted people might evidence this quality by seeking to interact with new people, by seeking to have many friends, through their comfort speaking in groups, by their willingness to tell jokes, and so on. Our understanding of this construct is simply too big to be captured by any single item. That said, extraversion needn't necessarily be expressed by any one of these specific behaviors. Some extraverts are known for their joy of speaking in public and others for their love of telling jokes. When aggregating across these and many other distinct expressions, a general tendency to be extraverted can emerge in an aggregate scale total, resulting in a meaningful unitary score that captures relative standing on this broad dimension.

A second way in which a construct might be broad is that it might be multidimensional, in that it is made up of interrelated but distinct facets. As examples, the construct of depression is often thought to be represented by four different (and also broad) facets: a cognitive dimension, an affective dimension, a somatic dimension, and an apathy dimension. Anxiety is thought to have three facets: social anxiety, generalized anxiety, and panic-related anxiety. Social support is thought to have three facets: tangible support, emotional support, and informational support.

To measure a broad construct, it is thus incumbent on researchers to clearly define it, specifying its dimensional structure based on theory or on past research. In the case of extraversion, where a researcher assumes a broad but unidimensional

attribute, the goal in generating items will be to approximate a selection of items, drawn from a theoretical and infinitely large pool of equally good items, each of which is influenced by a person's true extraversion (in a manner consistent with Equation 1). In contrast, in the case of depression, the goal will be to first define four facets of depression, and to repeat this same process of item generation four different times. In truth, whether pursuing items to capture a broad unidimensional construct or multiple broad dimensions of a multidimensional construct, some items will almost assuredly be better than others (as expressed by the relative size of λ and ε in Equation 1). However, through the creation of multiple imperfect items that vary in their quality, the resulting aggregate score can produce an observed estimate that is far better than can be generated by the pursuit of the single "best" self-report item.

An Iterative Process

How successful one will be at generating sets of questions that combine to estimate a broad construct should be viewed as an empirical question, one that often can be evaluated through reference to the results of analyses performed on the test items themselves. It is beyond the scope of this chapter to detail this process other than to note that scale construction is often an iterative process, one in which many potential items are generated, the "bad" items are revised or rejected, and the "good" items are retained. In the course of evaluating items, assumptions about the dimensionality of the construct should be scrutinized and perhaps revised, in light of empirical results. A construct that was first conceptualized as unidimensional might through trial and error reveal itself to be multidimensional, and vice versa. There are many useful texts to offer guidance on this iterative process and the standards one should apply to reevaluating measurement assumptions (see Furr & Bacharach, 2018; Nunally & Bernstein, 2004). Rather than review this well-covered material, we seek in the following sections to point to some of the more common misperceptions surrounding multi-item scales.

Internal Consistency versus Homogeneity

One common source of confusion is the distinction between the *internal consistency* of a multi-item scale and its degree of *homogeneity*. Internal consistency refers to the degree of interrelatedness of items, whereas homogeneity refers to their dimensionality or the extent to which the covariance structure among items can be accounted for by a single latent factor. These properties are not isomorphic. For example, if 10 items are all intercorrelated at $r = 0.20$, the correlational pattern among them can be accounted for by a single latent variable (i.e., they are unidimensional), but their internal consistency is relatively modest. As the intercorrelation between items increases, so too will the internal consistency, everything else being equal. Coefficient alpha is a common index thought to reflect the internal consistency of items, the homogeneity of items, or both. However,

despite the isomorphism this example highlights, reliability estimates do not make for good homogeneity estimates. Consider the correlation patterns for two six-item scales, each with an alpha of 0.86:

Item	*1*	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*
1	–						–					
2	.8	–					.5	–				
3	.8	.8	–				.5	.5	–			
4	.3	.3	.3	–			.5	.5	.5	–		
5	.3	.3	.3	.8	–		.5	.5	.5	.5	–	
6	.3	.3	.3	.8	.8	–	.5	.5	.5	.5	.5	–

The scale on the left clearly is not unidimensional despite its large coefficient alpha. The scale on the right is unidmensional. To determine unidimensionality, one should assess it directly and not infer it from reliability (see Cortina, 1993).

Assessing Unidimensionality

So how is homogeneity to be assessed? One useful strategy is to conduct a confirmatory factor analysis on items. If a one-factor model fits the data well, then one can assume unidimensionality. A common practice after a factor analysis of items (be it confirmatory or exploratory) is to select only items that load on the same factor and the use these as the core items for aggregation in a final scale. Unfortunately, there are no clear standards for what constitutes a large enough factor loading for an item to be said to adequately represent the underlying factor. Loadings in the 0.30 to 0.40 range are often suggested, but closer inspection of what these values mean suggest that one might want higher standards. For example, in a traditional confirmatory factor analysis, the square of a standardized factor loading is the proportion of variation in an indicator that is due to the underlying factor, and one minus this value is the proportion of unique variance associated with the indicator. A factor loading of 0.50, for example, implies that just 25% of the variation in the indicator is due to the underlying factor whereas 75% of its variation is unique and has nothing to do with the factor. With this in mind, suppose a researcher created a four-item scale measuring perceived stigma of having a mental health problem, finding that all four items load on a single factor as follows:

Item	*Loading*
Sometimes I am talked down to because of my mental health problems	0.60
I believe I would be discriminated against by my employers because of my mental health problems	0.50
I would have had better chances in life if I had not had a mental illness	0.52
People's reactions to my mental health problems make me keep to myself	0.55

A global index of perceived stigma can be obtained by aggregating across the four items but as a result, attention is drawn away from the unique variance of each item—even though each item is dominated by unique variance. Perhaps this unique variance is most relevant to predicting an outcome rather than the common variance among the items. Suppose that a researcher wished to determine the extent to which perceived stigma predicts discrimination in an employment setting. The second item has the lowest loading, 0.50, and this means that it has about 75% unique variance relative to the underlying generalized stigma factor. However, this item is the only item focused on perceptions of stigma in employment settings. As a general rule, the accuracy of prediction generally will increase to the extent features of the judgment closely correspond to features of the criterion one wishes to predict (for review, see Fishbein & Ajzen, 2010). Perhaps as a result, this particular researcher—given the nature of the research question at hand—should focus attention on this one item, not the scale total. There are many other contexts where one might not want to be too quick to focus exclusively on common variance but instead work with *both* the common and unique sources of variance. We caution researchers to consider both the ways that aggregate estimates of broad concepts (pro-environmentalism) might predict broad behavioral outcomes (e.g., carbon footprint), and how they might be separated out into more narrow constructs (e.g., aluminum recycling attitudes) to predict specific behavioral tendencies (e.g., aluminum recycling; see Davidson & Jaccard, 1979; Kallgren & Wood, 1986).

Assessing the Reliability of a Composite Through Item Analysis

As noted, items are often aggregated to capitalize on the fact that random error in individual items will tend to cancel out, yielding a more reliable composite. We often want to estimate the reliability of a composite, with coefficient alpha being the most frequently used index for doing so. However, psychometricians argue against its use (Sijtsma, 2009), and recommend an alternative index that makes fewer assumptions, called *composite reliability*. Both composite reliability and coefficient alpha assume unidimensionality, but only coefficient alpha also assumes (a) that the factor loadings for items are all equal and (b) there is no correlation between any of the errors of individual items. Such assumptions are often violated and, as such, composite reliabilities are generally preferred to coefficient alphas as an index of the reliability of a scale composite (see Raykov, 2001). By the same token, item elimination from a scale based on the value of the "coefficient alpha if item is eliminated" has been shown to be flawed and is better approached in terms of how the composite reliability is affected if a given item is eliminated (Raykov & Marcaloudis, 2015).

Making Your Scaling Function Explicit

Equation 1 presented the operative measurement model used throughout this chapter. Although a linear function between a measure and a latent construct is

by far the most common function assumed by researchers, it is just one of many possible functions that can be operating. This is particularly important to keep in mind when conducting item analyses for multi-item scales. In traditional scaling models, the researcher assumes that a given response to an item is generated as a linear function of the latent construct (per Equation 1). However, psychometricians have elaborated other functions that have implications for how one writes items. To this end, it is useful to consider a construct central to psychometrics, *item-operating characteristic* (IOC). An IOC specifies the relationship between true score on the construct of interest and how the probability of endorsement of an item changes as the true score increases (see Green, 1954). Consider as an example a researcher interested in measuring someone's attitude towards a given attitude object, X. There exist a number of plausible IOCs for measures designed with this purpose. One type of IOC derives from the logic of Thurstone scaling and states that the probability of endorsing an item should be highest for an individual whose attitude toward X matches the "scale value" of the item with respect to X. For example, an individual with a neutral attitude toward X should be most likely to endorse an item that conveys neutrality with respect to X; an individual with a moderately positive attitude toward X should be most likely to endorse items that express moderately positive favorability towards X; and a person with an extremely unfavorable attitude toward X should be most likely to endorse items that express extreme unfavorability towards X. The more discrepant an individual's attitude is from the particular scale value of the item, in either a positive or a negative direction, the less likely the individual should be to endorse the item.

Figure 1.3 presents the IOCs based on this logic for three items that differ in their scale values. The scale values, in principle, vary from 0 to 10, with higher scores indicating higher degrees of favorability and 5 representing a neutral point. The first item in Figure 1.3 has an extreme positive scale value (of 10), and it can be seen that the IOC for this behavior is linear in form: The more positive the person's attitude towards X, the more likely the person will be to endorse the item. Consider the second item. This item has a scale value of 5, which represents neutral affect. In this case, individuals with neutral attitudes are most likely to endorse the item and the probability of endorsement decreases as one's attitude becomes more negative or more positive. This IOC is curvilinear in form and one would expect a low correlation between item endorsement and a person's attitude, because a correlation coefficient is primarily sensitive to linear relationships. Thus, using Thurstonian logic, one cannot identify "good" items purely by examining item-total correlations. Rather, one needs to use analytic strategies that allow for non-linearity in the probability of endorsement of an item and the total score, depending on the item's scale value.

An alternative conceptualization of the IOC derives from the basic logic of Guttman scaling (Edwards, 1957). Guttman assumed step-shaped IOCs: If an individual's attitude is less favorable than the degree of favorability implied by an item (i.e., its scale value), then the probability of endorsing the item is zero. However, if the individual's attitude is as favorable or more favorable than the scale value of

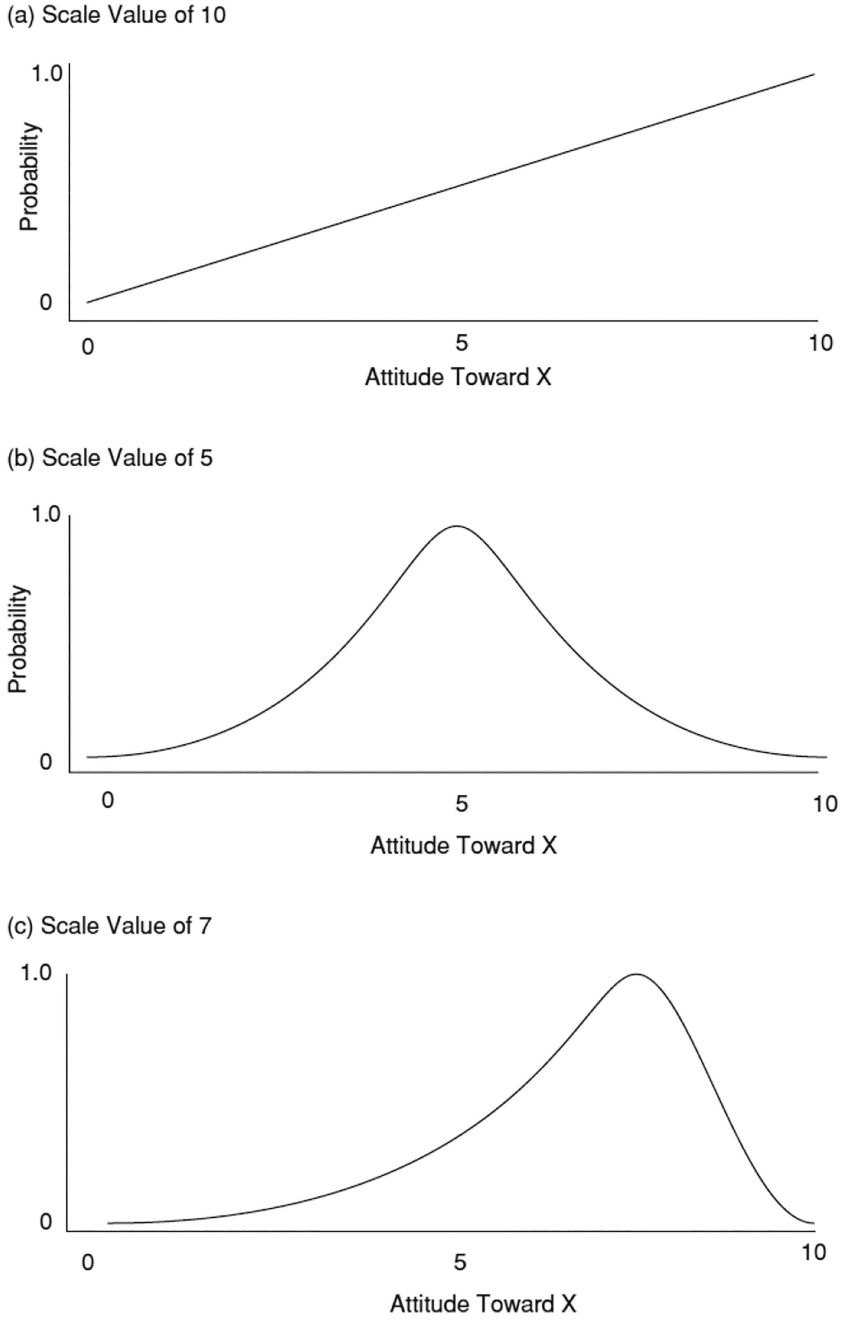

FIGURE 1.3 IOCs for Thurstone Scaling

the item, the probability of endorsement is 1.0. Figure 1.4 presents IOCs for the same three items using Guttman's logic (see Edwards, 1957, for elaboration of this rationale). Again, item-total correlations will not be helpful in identifying strong items under this form of measurement model. Rather, we require analytic strategies that are sensitive to step-shaped functions.

The general point is that the way we write items and the analyses we use to identify strong items for a multi-item scale are highly dependent on the measurement model we assume and the presumed item-operating characteristics for that scale. A measurement model that assumes simple linear IOCs (which is typical of Likert scaling) is but one model that can be adopted. It is important to be explicit about the measurement model one seeks to use.

Writing Self-Report Items

In this next section, we translate the measurement principles discussed above to provide concrete advice about writing self-report items and measures. The first step in generating questions is to step back and define the construct one wishes to measure. We start there and move through a wide range of tips to writing stronger questions for quantitative analyses.

Defining the Scope of the Construct

If a construct is conceived as narrow in scope, then a single, straightforward question might be sufficient to produce a sufficiently valid and reliable estimate. When trying to estimate the intention to vote for Candidate X, for instance, a single rating scale measuring perceived likelihood of voting for Candidate X will cover a lot of ground. If multiple items are attempted (e.g., *intention* to vote, *willingness* to vote, and *expectation* of voting), the inter-correlations will likely be so high that little information is gained, although the cancelation of random errors may increase measure reliability.

In contrast, if the construct is conceived as broad and manifested in many ways, greater thought must to be given to the nature and types of items needed to fully sample the construct universe of interest. Is the construct broad but unidimensional? If so, unidimensionality should be a priority when generating questions that might load on a single factor. With a broad, unidimensional construct, one should be to produce items that sample liberally from a larger pool of potential expressions. Although random (and systematic) sources of error might affect each individual item to some extent, potentially resulting in lower inter-item correlations, higher reliability can be produced through aggregation.

In the process of generating items to assess broad constructs, however, one should give consideration to the unique variance introduced by specific items and whether any given item taps unique facets of the construct that have value in their own right, as stand-alone items. When generating items to estimate a person's overall

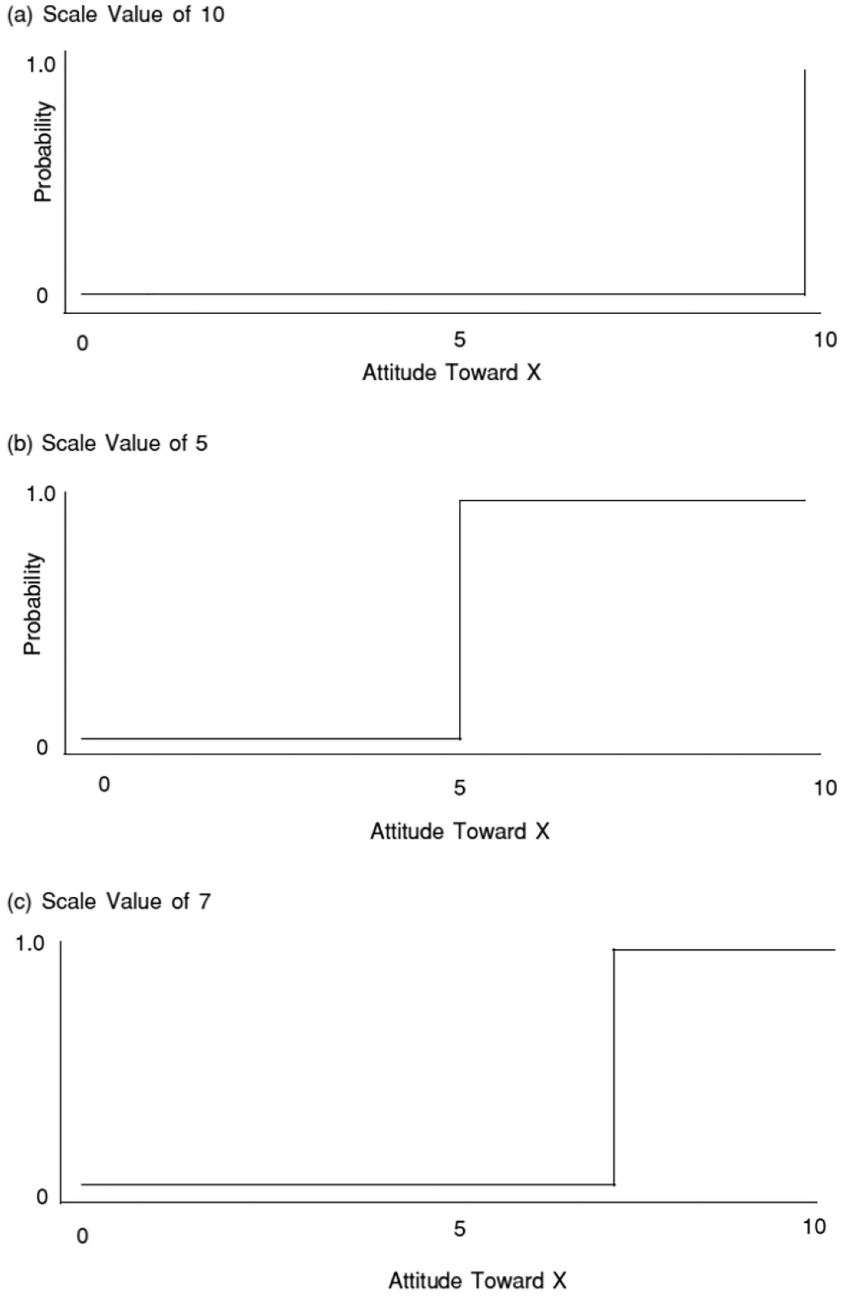

FIGURE 1.4 IOCs for Guttman Scaling

level of extraversion, for instance, a researcher might see different applications for items that tap sociability (e.g., being outgoing) social assertiveness (e.g., likes taking charge), each of which is an expression of extraversion. As attention turns from common to unique item variance, the researcher might consider if there is cause to generate sets of multiple, interchangeable items taping different distinct types expressions of the original construct (see, for instance, Soto & John, 2017).

Articulating the Item-Operating Characteristic

We noted earlier that the traditional and most common approach to measurement is to assume linear IOCs for items comprising a multi-item scale. In such cases, one should write questions with an eye for generating items that will have high inter-correlations and high item-total correlations. Consider for instance a researcher interested in measuring attitudes towards smoking marijuana. Respondents are asked to rate their endorsement of the statement "I can be friends with a person who has smoked marijuana in the past" but this might be problematic. Most anyone with even the most modest of pro-marijuana attitudes will endorse this statement highly, restricting the range of responses and yielding a data pattern that is at odds with a linear IOC. When working with non-linear IOCs, as much as possible, items should be constructed such that across the full set of items, they capture incremental, linear movement along the full range of potential scale values that might occur along the theoretical metric of interest. One should avoid ending up with a scale where the items, as a collective, truncate the range of scale values.

Tip 1: Consider a "Linear Wording" Approach to Asking Questions

When working with linear IOCs (as is typically the case), one should generate items whose probability of endorsement will clearly vary linearly as the underlying construct changes. Be particularly aware of base rate issues surrounding ceiling and floor effects. Suppose, for instance, a researcher wishes to assess attitudes towards getting pregnant among high school seniors. An item like "Getting pregnant now would be bad" would probably be of limited use, because almost all high school girls will agree strongly with the statement; it will not discriminate those with highly negative attitudes from those with moderately negative attitudes. However, this item can be modified to read, "Getting pregnant now would be one of the worst things that could happen to me." This likely would avoid the ceiling effect that the prior version of the item exhibits. Even a subtle shift in phrasing from "I was sad last week" versus "I was very sad last week" can affect response distributions for items in ways that improve the range of responses one obtains as a function of the underlying construct.

In some cases, one also can adopt a strategy of simply asking people directly how they stand on the construct in question. To sort individuals in terms of their

attitudes towards smoking marijuana, for instance, one might simply ask respondents "how do you feel about smoking marijuana?" where answers are made on a numeric scale that ranges on a dimension from negative to positive, with a neutral midpoint. Or, one might ask respondents to complete the phrase "I feel _____ about smoking marijuana," by choosing from a set of options that range from "extremely negative" to "extremely positive" (and see below where we list rating scale options to use in such a scenario). Such questions can be framed in a way that they reflect a linear relationship between responses to the item and the underlying latent construct. It is not always possible to articulate a construct in this manner, especially for constructs that are broad and require assessments of a wide range of interrelated manifestations. But, in many instances, a good way to obtain an index of a latent construct is to clearly articulate to the respondent the type of judgment (or IOC) that is desired and then to ask them to provide a rating accordingly.

Reducing Random Error

Random error is an unfortunate fact of life and researchers should expect it to influence responses to some degree. However, there are many ways to minimize it through the design of questions and in this section, we provide tips that might help.

Tip 2: Keep Items Short, Simple, and Understandable

The more cognitively demanding a question, the greater likelihood that transient differences in motivation, attention, and interest will affect responding. As much as possible, avoid long sentences, large or obscure words, complex phrasing, and unnecessary words. In most common instances, try to keep the reading level to about the fourth or fifth grade.

Tip 3: Make Sure the Item Measures Only One Concept

Items that are open to multiple interpretations will be more prone to error. An item like "My therapist was expert and sincere" is double-barreled and thus inherently ambiguous. Respondents who view their therapist as expert but not sincere have no valid response. Similarly, an item like "I intend to go to my appointment because it will help me get better" might be difficult for respondents who intend to attend their appointment, despite holding doubt it will help.

Tip 4: Avoid Negations, Particularly Double Negations

Inclusion of a negation in a question can be confusing, particularly if the item offers an opportunity to reject the original negation. For instance, respondents

who like to dance might fail to notice or skim over the word "not" when asked to "agree or disagree" with the statement, "I do not like to dance." Problems only increase if a question is a double negation. An item like "Students should not fail to go to school" can be cognitively demanding (especially for respondents given the option to "disagree" with the statement). As a general rule, it is best to avoid the word "not" altogether, as people often misinterpret or fail to notice the negation and misreport.

Tip 5: Look for and Remove Potentially Ambiguous Terms and Phrases

One problem where some psychologists have difficulty is with the use of jargon. We become so fluent as "psychologizers," that we forget how confusing we can be to many of the non-psychologists we wish to study. It would probably be a bad idea, for instance, to ask respondents how "reactant" they felt while reading a health message. Lack of understanding can be subtle, however, and can also arise from familiar, non-jargon words. Even a simple item, like, "I smoked marijuana last month" can introduce ambiguities, because some respondents will interpret the "last month" as some time in the last 30 days, whereas others will interpret it based on a calendar month. It is an unfortunate fact that some ambiguities only become obvious to researchers after the data have been collected but someone good at writing questions will put considerable energy into identifying any potential source of confusion a priori.

Tip 6: Personalize the Item and Provide Contextual Information and Time Frames

If not made explicit in a question, respondents will often impute their own time frames and other contextual information into questions, leading to item unreliability. The item "Joining a gang would be good" can elicit a very different response than the item "For me, joining a gang in my neighborhood at this time would be good." The first statement not only fails to indicate a time period, it fails to clarify whether the respondent is being asked about his or her own gang-related decisions or the decisions of people in general. As much as is possible, clarify the "who," the "what," the "where," and the "when" as well as the effect that is of interest.

Tip 7: Avoid Slang and Abbreviations

Our earlier warning against jargon points to the importance of writing in familiar and accessible language, but pursuit of the colloquial can misfire when the researcher drifts into the use of slang. Although many respondents might refer

to marijuana as "weed," it would be unwise to assume that this term is universally understood. Abbreviations carry related problems. An item like "I know the whereabouts of my child 24/7" might fit the way many parents talk, but it might also be confusing to parents unfamiliar with the phrase. As "square" as it might seem to clearly define your terms and stick to dry, clinical language, this approach to writing questions will often reduce the influence of random error.

Reducing Systematic Error

As with random error, the potential sources of systematic error might extend far beyond a researcher's ability to anticipate. There are some common culprits, however, and researchers should be on the lookout for them. Chief among these are sources of systematic error that can come about as a function of respondent demographics: gender, age, race, education, and socioeconomic status, to name a few. Self-report items written by a researcher from the viewpoint of his or her own social groups and life experiences might have far different meanings to respondents reading them from the vantage of different groups and experiences. For instance, a female researcher asking questions concerning "sexual harassment" might fail to realize that her male (but not female) respondents bring far different interpretations to this term than she had in mind while she was constructing her questions. Similarly, questions assessing "attitudes towards education" might be interpreted differently by children whose parents are college graduates, compared to those whose parents are high school dropouts. Much as with random error, one way of reducing systematic error is to define one's terms and write questions clearly, such that a single, unambiguous meaning dominates.

Importantly, however, demographic differences are not the only factors that can exert systematic influences on ratings. Error can also be introduced as a result of any number of psychological attributes that exert influence on ratings. Earlier (Figure 1.2), we pointed to one potential source of systematic bias, social desirability. Concern for one's public and private image can undermine self-reports on a wide range of sensitive topics. Practices that have been shown to reduce the effects of social desirability on self-reports include:

- Use of self-administered as opposed to face-to-face reports, such that respondents do not have to report sensitive behaviors directly to another person.
- Use of anonymous or confidential conditions, offering respondents reassurance that identifying information will not be associated with their data.
- Delivery of motivational instructions at the outset, encouraging honest reporting.
- Instructing respondents not to answer question at all, if they are not going to be truthful in their response (and using state of the art analytic methods to handle the missing data that results).

- Obtaining a measure of social desirability tendencies and using it as a statistical covariate when modeling the data.

Any or all of these methods might be applied to lessen the impact of desirability concerns on reporting, but it also is important to consider ways to eliminate bias through the design of better self-report items. This leads us to two new tips:

Tip 8: Avoid Leading Questions

Sometimes while writing questions, we reveal our own assumptions and values. The linguistic cues that lead a research participant to respond in certain ways can appear subtle but still exert influences on the ratings given. An item phrased as "To what extent does your mother disapprove of marijuana?" might elicit different answers than an item phrased as "To what extent does your mother approve or disapprove of marijuana?" The former item might lead or encourage respondents to communicate disapproval, as it fails to acknowledge that some mothers do hold favorable views towards marijuana.

Tip 9: Convey Your Acceptance of Potentially Undesirable Answers

Questions can be worded such that they reduce the sting of providing socially undesirable (but truthful) responses. For example, research suggests that older adults are less comfortable reporting their age than they are reporting the year they were born. They also are at times more comfortable checking off age categories than listing out their own specific age. One can also write questions in a manner that conveys acceptance. For instance, it is often the case that far more people indicate to pollsters that they voted in previous elections than voting rolls would indicate. People who did not vote might feel embarrassed to admit this, but some degree of embarrassment might be removed with careful questioning (e.g., "There are many reasons why people don't get a chance to vote. Sometimes they have an emergency, or are ill, or simply can't get to the polls. Did you vote in the last election?"). This strategy might make what was undesirable feel acceptable, but one has to be careful when using it not to be leading.

There are many other common forms of systematic error that one might also consider. For instance, psychometricians have identified a range of specific response styles, including (a) acquiescence response sets (i.e., the tendency make ratings indicating agreement), (b) disacquiescence response sets (i.e., the tendency to make ratings indicating disagreement), and (c) a middle-category response set (i.e., the tendency to move the midpoint of rating scales). The empirical evidence for prevalence of the contaminating influence of these artifacts is somewhat inconsistent, but it is clear they operate for some populations, in some contexts

(see Conway & Lance, 2010; Podsakovv, MacKenzie, & Podsakovv, 2012; Rorer, 1965; Wiggins, 1973). These possibilities do point to another tip:

Tip 10: Write Both Positively and Negatively Keyed Items, as Appropriate

One approach to dealing with acquiescence and disacquiescence response sets is to pursue a balance of positively and negatively keyed items. If one is seeking to measure extraversion, for instance, it might be a good idea to include positively keyed items (assessing such things as comfort talking to people), as well as negatively keyed items (assessing such things as interest in being alone). This advice comes with two large caveats, however.

First, it is important to treat as an empirical question the factor structure of a multi-item scale containing both positively and negatively keyed items. It may be that as a result of one general factor, extraversion, the greater comfort someone has talking to people, the less interest that person has in being alone. Or, it may be that the construct measured by positively keyed items (extraversion) is empirically distinct from the constructs measured by negatively keyed items (introversion). In research on attitude structure, for instance, researchers often find evidence that positive and negative evaluations of the same object are empirically distinct. Positive and negative evaluations can have distinct cognitive and emotional antecedents, as well as distinct consequences for judgment, decisions, and behavior (Cacioppo, Gardner, & Bernston, 1997), and so unidimensionality should not be assumed.

Second, when writing questions, it is important to generate positively and negatively keyed items that are non-redundant and equally sensible (see Weijters & Baumgartner, 2012). People can run afoul on both counts if their strategy for generating negatively keyed items is to try to "reverse" other, positively keyed items in an inventory. By simply reversing a sensible question, a nonsensible sentence might result. This is particularly likely if the new item is created through negation, which we noted earlier can introduce error. Whereas respondents might find it easy to answer "to what extent does your mother approve or disapprove of marijuana," they might react with confusion when asked "to what extent does your mother NOT approve or disapprove of marijuana?"

Another problematic approach to producing reverse-keyed items is to include reverse-oriented, counterintuitive scales. Whereas this response metric is intuitive:

How much do you like going to parties?								
Not at All	0	1	2	3	4	5	6	*Extremely*

This metric might seem odd (and highly confusing) to many respondents:

How much do you like going to parties?								
Extremely	0	1	2	3	4	5	6	*Not at all*

It is reasonable to expect that some respondents answering the second question will wonder why enjoyment implies a lower number. Are they misunderstanding the question? Does the researcher have some trick up a sleeve? By introducing provocative and counterintuitive metrics into the mix, respondents might slow and perhaps become confused, producing misreporting.

Designing Item Metrics

The Pursuit of Rating Precision

Items are often rated on metrics using judgments such as agree-disagree, true-false, approve-disapprove, or favorable-unfavorable. Such metrics can be dichotomous ("yes" versus "no") or many-valued (such as "strongly agree," "moderately agree," "neither," "moderately disagree," and "strongly agree"). The ***precision*** of a metric or scale refers to the number of discriminations it allows the respondent to make. Earlier we showed how precision might be reduced if questions are worded in a way that yields ceiling or floor effects, but the metrics one employs can have similar effects, if they force respondents who have meaningfully different evaluations to use the same category to describe their states of mind. Consider an item and response scale like this:

How much do you approve or disapprove of the Affordable Care Act?
____ *Disapprove* ____ *Approve*

This question creates a reality in which respondents who "slightly disapprove" of the Affordable Care Act will receive the same score as those who "strongly oppose" it. Treating such people as if they are the same when analyzing data can introduce bias into parameter estimates and adversely affect statistical power. A simulation study by Bollen and Barb (1981) is informative. These authors created data, such that the true population correlations between two continuous variables were either 0.20, 0.60, 0.80, or 0.90. They 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 continuous variable that ranges from -3 to +3 can be turned into a two-point scale by assigning anyone with a score of 0 or less a "0" and anyone with a score greater than 0 a "1."

They found that true correlations were relatively well reproduced by coarse measures, as long as the coarse measures had 5 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 at or below 0.60. They concluded that five categories were probably sufficient for many research applications, and this recommendation has been borne out in many other simulation studies (although some research suggests seven or more categories may be best in some contexts; see Green, Akey, Fleming,

Hershberger, & Marquis, 1997; Lozano, García-Cueto, & Muñiz, 2008; Lubke & Muthén, 2004; Taylor, West, & Aiken, 2006). Thus, coarse measurement is not necessarily problematic, unless it is very coarse, namely less than five categories, leading us to the next tip:

Tip 11: In Most Instances, Orient Questions Around Five or More Response Categories

There are some caveats to this tip as well. First, this only applies to psychological attributes that are continuous in form. For ratings that orient around nominal categories (e.g., country of origin) the number of categories are dictated by the substantive content of the construct. For populations where researchers believe the cognitive demands of using a rating scale with five or more points is problematic, precision often can be had by delivering responses orally and in multiple steps. For example, one might ask respondents if they "agree," "disagree," or have "no opinion" about a given statement. Those who agree can then be asked in a follow-up if they "strongly" or "moderately" agree, just as those who disagree can be asked if they "strongly" or "moderately" disagree. Across the two questions, the researcher can then classify the respondent as having chosen one of the five categories ("strongly disagree," "moderately disagree," "neither," "moderately agree," or "strongly agree").

Inclusion of a "Don't Know" Response?

A common criticism of ratings scales is that they structure answers to such a degree that respondents are able to report evaluations that mean nothing to them (Sniderman, Tetlock, & Elms, 2001). One strategy that is sometimes used to combat this is to offer respondents a "don't know" or "no opinion" response option. With this option, a respondent does not have to answer a question. According to some theorists, people indeed often have "no opinion" on a topic and if forced to respond to an item without them allowing to indicate "don't know," they will either respond randomly or in a non-meaningful way based on situational features in the testing context or their mood. If we include "don't know" options, however, we may end up with a large number of answers that must be coded as non-responses. Despite plausible predictions to the contrary, extant research on this matter does not support the universal assertion that inclusion of a "don't know" response category increases the reliability or validity of a measure, although there are some exceptions. In our view, a better strategy for generating meaningful data is to conduct qualitative research before questions are created to gain a better understanding of what questions are or are not meaningful to those in the population of interest, and to write questions accordingly (see Fisher, Fisher, & Aberizk, this volume).

Choosing Adverb Qualifiers

Data analyses promote stronger conclusions when a researcher works with measures that have interval or ratio-level properties, rather than nominal or ordinal properties. Interval properties often can be better approximated with rating scales using adverb qualifiers, as when respondents are asked if they agree with a statement "a little" or "a lot." Interval-level properties can prove elusive if one utilizes a discrete set of adverbs that create unequal intervals or "spacing" between them, as with this question and response:

How much do you love puppies?			
Not at All	*A Little*	*Somewhat*	*Completely*

The difference between puppy-loving "a little" and "somewhat" seems slight, especially compared to the difference between a puppy-loving level of "somewhat" versus "completely." This set of response categories illustrates the importance of pursuing anchors that create equal-appearing intervals, covering the full range of possible evaluations from low to high. There are large literatures in psychometrics that researchers can consult to identify adverb sets that help produce equal-appearing intervals (Budescu & Wallsten, 1994; Cox, 1980; Czaja & Blair, 2005; Dawes & Smith, 1985; Krosnick & Fabrigar, 1998; Tourangeau & Rasinski, 1988). As one example, an early study in psychophysics attempted to determine the modifying value of different adverbs. Cliff (1959) found that describing something as "slightly good" was perceived to be about 0.33 times as "good" as the simple, unmodified "good" (also see 1966a, 1966b). By consulting research on modifying values of adverbs, one can choose adverb qualifiers to more closely approximate equal-appearing intervals, thus producing ratings that more closely approximate interval-level properties. To be sure, some care must be taken in doing so, because qualifying values have been found to vary somewhat as a function of the population being studied and the type of judgment being made. However, more often than not, use of carefully selected adverb qualifiers will produce data that reasonably approximate interval-level properties. As a practical aid to readers, the Appendix provides sets of adverb qualifiers that produce roughly interval-level data for a wide range of judgments (and see Vagias, 2006).

Combining Numeric Ratings With Adverb Anchors

Data analyses generally are more straightforward and efficient if one works with measures that have at least interval-level properties. In this pursuit, self-report scales often orient around simple numeric rating scales. However, the mere fact that respondents can answer your question within provided numeric sequences

does not mean that you have achieved interval measurement. As an example, consider the following:

How do you feel?	*Sad*	1	2	3	4	*Happy*

This researcher is interested in measuring mood but there is a mismatch between the numbering system and the anchors. The anchors suggest interest in a bipolar construct, anchored at one pole with "Sad" and the other by "Happy" but the rating system is unipolar, moving from 1 to 4. How would a respondent indicate a neutral mood? It also is unclear why "Sad" is associated with a small quantity (the value of 1), compared to "Happy"—can't sad be felt with intensity? This example leads us to introduce a number of additional tips.

Tip 12: Communicate a "Zero Judgment" on Your Scale and Give It the Value of Zero

Some researchers include midpoints in rating scales (e.g., a "neutral" or "neither agree nor disagree" category), whereas other researchers omit them in order to force a respondent to "take a stand." Use of a midpoint is theoretically warranted if it represents a valid psychological response for the judgment in question. Indeed, respondents may become irritated if they are not allowed to express their true feelings or opinions. There has been considerable research on the use or non-use of midpoints and, although somewhat mixed, overall the research tends to favor the use of midpoints as long as they are theoretically meaningful.

Tip 13: Communicate Bipolar Dimensions With Bipolar Rating Systems, Centered on Zero

Consider a researcher interested in measuring mood on a scale that ranges from extreme sadness to extreme happiness, with a midpoint of neutrality. This can be expressed as:

How do you feel?	–3 *Extremely Sad*	–2	–1	0 *Neutral*	1	2	3 *Extremely Happy*

A scale such as this communicates clearly the researcher's conceptualization of mood as a bipolar evaluative dimension. It also utilizes a middle anchor to clarify the meaning of the zero-point; i.e., as the absence of either sadness or happiness.

Tip 14: Place Anchors That Approximate Interval-Level Distinctions at Equal-Appearing Numeric Intervals on the Response Scale

Error can be introduced in rating scales if they lack anchors at key points on the scale. Consider the following:

How happy are you?	*Not at All*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	*Extremely*

Researchers might use scales such as these in the hopes of increasing precision, but error can be introduced by such a numbering system, because it requests discriminations in terms of magnitude that likely go beyond the respondents' abilities to discern and/or communicate. What is the difference between happiness of 5 and 7 or between 7 and 12? Verbal anchors can help eliminate confusion about such rating systems. Earlier we discussed adverb modifiers that can be used to approximate interval-level rating systems (see also the Appendix). One fruitful approach is to combine these with the rating systems just discussed, reducing "number valleys" requesting fine discriminations. Here are two such examples:

How happy are you?										
0	1	2	3	4	5	6	7	8	9	10
Not at All Happy			*Slightly Happy*			*Quite Happy*			*Extremely Happy*	

How do you feel?						
–3	–2	–1	0	1	2	3
Extremely Sad	*Quite Sad*	*Slightly Sad*	*Neutral*	*Slightly Happy*	*Quite Happy*	*Extremely Happy*

With each of these ratings scales, the evaluative dimension is communicated through the use of a sensible numbering system and well-chosen anchors.

Tip 15: Add Extreme Anchors, When There Is Meaningful Variability at the Extremes

Sometimes opinions of interest will be endorsed extremely and with a high degree of consensus in populations of interest. As one example, Sweeney, Blanton, and Thompson (2009) sought to measure soldier's trust in their "most trusted leader," in the days before they participated in the launch of the second Gulf War. Needless to say, soldiers in these instances tended to have exceptionally high trust in this individual—so much so that one could reasonably anticipate that a ceiling effect would make this a meaningless rating. However, these researchers were able

to avoid this problem adding an extreme anchor and expanding precision around the extreme. The resulting question thus read:

To what extent do you trust your most trusted leader?

0	1	2	3	4	5	6	7	8	9	10	11	12
Not at All			*Slightly*			*Quite a Bit*			*Extremely*			*Completely*

Although this scale might appear to request a wide range of evaluations from respondents, the researchers effectively administered a 4-point scale to this group of soldiers, as all but a handful of made ratings that ranged from 9 to 12. Similar strategies can be useful for predicting such things as adolescent health-risk tendencies, as even those likely to engage in risky behaviors tend to express negative evaluations, but to varying (and predictive) degrees (Gibbons, Gerrard, Blanton, & Russell, 1998). Burrows and Blanton (2015) reported results of a pilot study where they successfully predicted the likelihood of driving under the influence of alcohol (DUI), using a response scale that asked respondents to discriminate whether they were "completely" unwilling to drive under the influence or just "extremely" unwilling to DUI. In our view, one of the arguments for utilizing implicit measures rather than self-reports—i.e., that people often will not report socially undesirable attitudes—might in some instances be more easily addressed by giving respondents the option of making extreme ratings. Consider the measurement of racial bias, for instance, where it is often argued that respondents will not report socially undesirable attitudes they possess. Rather than pursuing implicit measurement strategies as a response, however, one might seek to measure how "completely" or "absolutely" individuals reject prejudicial beliefs and attitudes, using where the more moderate position is simply to reject prejudicial attitudes "extremely" (see Blanton & Jaccard, 2015).

Conclusion

Self-report is and will likely remain the most ubiquitous method of psychological assessment, in part because self-report items are easy to construct. Often missed, however, is the ease with which self-report items might be constructed, badly. We hope this chapter illustrates that the likelihood of writing strong questions can be increased through rigorous application of measurement principles. Researchers can improve their questions by clearly defining their constructs in terms of breadth and dimensionality, articulating scaling functions desired of questions, and paying close attention to sources of random and systematic error, such that they write stronger questions, and provide more informative ratings scales, and combine multiple items when conditions suggest this will improve measurement of the construct of interest.

APPENDIX

Across a wide range of psychometric studies, the following two sets of adverb qualifiers tend to produce roughly interval-level data:

For agreement judgments, two sets of reasonable qualifiers are:	
Strongly agree	*Strongly agree*
Moderately agree	*Agree*
Neither	*Neither*
Moderately disagree	*Disagree*
Strongly disagree	*Strongly disagree*
For frequency judgments, two sets of reasonable qualifiers are	
Very frequently	*Always or almost always*
Frequently	*Usually*
Occasionally	*About half the time*
Rarely	*Sometimes*
Never	*Never or almost never*
For importance judgments, two useful sets of adverb qualifiers are	
Extremely important	*Very important*
Quite important	*Moderately important*
Slightly important	*Slightly important*
Not at all important	*Unimportant*
For bipolar affective judgments, two sets of reasonable adverb qualifiers are	
Extremely favorable	*Very good*
Quite favorable	*Quite good*
Slightly favorable	*Slightly good*
Neither	*Neither*
Slightly unfavorable	*Slightly bad*
Quite unfavorable	*Quite bad*
Extremely unfavorable	*Very bad*

For extreme ratings, where ceiling or floor effects appear likely, consider adding extreme options (e.g., "absolutely" or "completely") to expend beyond traditional endpoint (e.g., "extremely").

References

Beckstead, J. (2014). On measurements and their quality. Paper 4: Verbal anchors and the number of response options in rating scales. *International Journal of Nursing Studies*, *51*(5), 807–814. doi:10.1016/j.ijnurstu.2013.09.004

Blanton, H., & Jaccard, J. (2015). Not so fast: Ten challenges to importing implicit attitude measures to media psychology. *Media Psychology*, *18*(3), 338–369.

Bollen, K. A., & Barb, K. (1981). Pearson's *r* and coarsely categorized measures. *American Sociological Review*, *46*, 232–239.

Budescu, D. V., & Wallsten, T. S. (1994). Processing linguistic probabilities: General principles and empirical evidence. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Decision making from the perspective of cognitive psychology*. New York, NY: Academic Press.

Burrows, C. N., & Blanton, H. (2015). Real-world persuasion from virtual world campaigns: How transportation into virtual worlds moderates in-game influence. *Communication Research*, *43*(4), 542–570.

Cacioppo, J. T., Gardner, W. L., & Bernston, C. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*(1), 3–25.

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review*, *66*, 27–44.

Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business Psychology*, *25*, 325–334.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*, 402–422.

Czaja, R., & Blair, J. (2005). *Designing surveys: A guide to decisions and procedures*. Thousand Oaks, CA: Pine Forge Press.

Davidson, A. R., & Jaccard, J. J. (1979). Variables that moderate the attitude—behavior relation: Results of a longitudinal survey. *Journal of Personality and Social Psychology*, *37*(8), 1364–1376. doi:10.1037/0022-3514.37.8.1364

Dawes, R. M., & Smith, T. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (pp. 509–566). New York, NY: Random House.

Edwards, A. L. (1957). *Techniques of attitude scale construction*. East Norwalk, CT: Appleton-Century-Crofts.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York, NY: Psychology Press (Taylor & Francis).

Fleming, P. (2012). Social desirability, not what it seems: A review of the implications for self-reports. *The International Journal of Educational and Psychological Assessment*, *11*(1), 3–22.

Furr, M., & Bacharach, V. (2018). *Psychometrics: An introduction* (2nd ed.). Newbury Park: Sage Publications.

Gibbons, F. X., Gerrard, M., Blanton, H., & Russell, D. (1998). Reasoned action and social reaction: Intention and willingness as independent predictors of health risk. *Journal of Personality and Social Psychology*, *74*(5), 1164–1180.

Green, B. B. (1954). Attitude measurement. In G. Lindzey (Ed.), *Handbook of social psychology* (pp. 335–469). Cambridge, MA: Addison-Wesley.

Green, S. B., Akey, T., Fleming, K., Hershberger, & Marquis, J. (1997). Effect of the number of scale points on chi square fit indices in confirmatory factor analysis. *Structural Equation Modeling*, *4*, 108–120.

Howe, E. S. (1966a). Verb tense, negatives and other determinants of the intensity of evaluative meaning. *Journal of Verbal Learning and Verbal Behavior*, *5*, 147–155.

Howe, E. S. (1966b). Associative structure of quantifiers. *Journal of Verbal Learning and Verbal Behavior*, *5*, 156–162.

Kallgren, C. A., & Wood, W. (1986). Access to attitude-relevant information in memory as a determinant of attitude-behavior consistency. *Journal of Experimental Social Psychology*, *22*(4), 328–338. doi:10.1016/0022-1031(86)90018-90011

Kline, R. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford.

Krosnick, J. A., & Fabrigar, L. R. (1998). *Designing good questionnaires: Insights from psychology*. New York, NY: Oxford University Press.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of ratings scales. *Methodology*, *4*, 73–79.

Lubke, G., & Muthén, B. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*, 514–534. Retrieved from https://doi.org/10.1027/1614-2241.4.2.73

Nunally, J., & Bernstein, I. (2004). *Psychometric theory*. New York, NY: McGraw-Hill.

Podsakovv, P. M., MacKenzie, S. B., & Podsakovv, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539–569.

Raykov, T. (2001). Bias of Cronbach's coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76.

Raykov, T., & Marcoulides, G. A., (2015). Scale reliability evaluation under multiple assumption violations. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 302–313. doi:10.1080/10705511.2014.938597

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, *63*(3), 129–156. http://dx.doi.org/10.1037/h0021888

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120.

Sniderman, P. M., Tetlock, P. E., & Elms, L. (2001). Public opinion and democratic politics: The problem of nonattitudes and the social construction of political judgment. In J. H. Kuklinski & J. H. Kuklinski (Eds.), *Citizens and politics: Perspectives from political psychology* (pp. 254–288). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511896941.013

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*, 117–143.

Sweeney, P. J., Thompson, V., & Blanton, H. (2009). Trust in combat: A test of an interdependence model and the links to leadership in Iraq. *Journal of Applied Social Psychology*, *39*(1), 235–264.

Taylor, A., West, S., & Aiken, L. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement*, *66*, 228–239.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*, 299–314.

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5*(3), 243–262. doi:10.1177/1745691610369465

Vagias, W. M. (2006). *Likert-type scale response anchors*. Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson University. Retrieved November, 2017, from www.uc.edu/content/dam/uc/sas/docs/Assessment/likert-type%20response%20anchors.pdf

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737–747. doi:10.1509/jmr.11.0368

Wiggins, J.S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley