Scale Construction and Psychometrics for Social and Personality Psychology

R. Michael Furr
Good research in social/personality psychology depends upon good measurement, which often depends upon careful and effective scale construction. This chapter focusses on the second step of the scale-construction process—choosing a response format and assembling items. Due to space limitations, it discusses key issues, but additional details are available in other excellent sources. Some sources provide general perspectives (e.g., John & Benet-Martinez, 2000; Wegener & Fabrigar, 2004; Dunn, 2009), with others focussing on specific domains (e.g., Jackson, 1971; Burisch, 1984; Fabrigar et al., 2006).

This chapter addresses scale construction in terms of scales common to lab-based research—multi-item questionnaires or inventories reflecting one or a few psychological variables. Such scales usually include either questions to be answered by respondents or statements to which respondents rate their level of agreement or endorsement, with items often framed in terms of agreement (e.g., with a statement of opinion or fact), frequency (e.g., of a behavior or event), quality (e.g., of an object), likelihood (e.g., of an event), truth (e.g., of an assertion), or importance (e.g., of a belief). To produce wide-ranging surveys for large-scale social or political research, such narrowly-targeted scales can be combined with other scales of varying content and format.

More specifically, this chapter focusses on scales with closed-ended items, which provide a limited number of response options. For example, true–false items, multiple-choice items, Likert-type items, and semantic differential items are all closed-ended items because respondents must choose from a set of pre-specified response options. Such items are common in psychological research, and they differ from open-ended items that place no or few constraints on responses (e.g., questions that elicit narrative responses). Other sources discuss considerations in designing open-ended items or wide-ranging social surveys (e.g., Chapter 6 in Dunn, 2009; Wegener & Fabrigar, 2004; Chapter 11 in Whitley, 2002; Visser et al., 2000).
Table 3.1  Considerations and recommendations in choosing a response format and assembling items

<table>
<thead>
<tr>
<th>Considerations and recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Response format</strong></td>
</tr>
<tr>
<td>1 General type of format (e.g., likert, semantic-differential)</td>
</tr>
<tr>
<td>2 Number of options</td>
</tr>
<tr>
<td>3 Labels/anchors</td>
</tr>
<tr>
<td>4 Mid-points</td>
</tr>
<tr>
<td>5 &quot;No opinion&quot; and &quot;Don’t know&quot;</td>
</tr>
<tr>
<td>6 Consistency across items</td>
</tr>
<tr>
<td><strong>Item-writing</strong></td>
</tr>
<tr>
<td>1 Relevant content</td>
</tr>
<tr>
<td>2 Number of items</td>
</tr>
<tr>
<td>3 Clear language (e.g., jargon, double-negatives, double-barreled)</td>
</tr>
<tr>
<td>4 Not leading or presumptive</td>
</tr>
<tr>
<td>5 Balanced scales—positively-keyed and negatively-keyed items</td>
</tr>
</tbody>
</table>

**Response Formats**

Roughly speaking, a scale’s response format refers to the way in which items are presented and responses are obtained. As summarized in Table 3.1, there are several issues associated with response formats.

**General type of format**

As mentioned earlier, this discussion focuses on closed-ended scales. More specifically, it focuses on scales having graded response formats—scales generally described as Likert scales or semantic differential scales. These are probably the most common response formats in social/personality psychology.

For Likert-type scales, respondents read an item’s text, which is usually a question or statement (e.g., “I like to go to parties”). They then choose one of the available response options (e.g., “Strongly Agree”). Usually, each available response option is associated with a specific quantitative value (e.g., 1 = Strongly Disagree; 5 = Strongly Agree), which are summed or averaged across all of a person’s responses relevant to a given dimension.

In scales modeled on a semantic-differential format, each item includes paired adjectives representing the poles of a psychological dimension relevant to the construct of interest. For example, a measure of friendliness might include “cruel” and “kind,” along with terms and/or values representing the breadth of the dimension (e.g., 1 = very cruel, 2 = somewhat cruel, 3 = equally cruel and kind, 4 = somewhat kind, 5 = very kind). Respondents consider the target of the assessment (e.g., themselves) and choose the appropriate term or value (note that, even if the options are not presented with numerical values, they are scored as having specific
numerical values). Again, a participant’s response values are aggregated across all adjective pairs relevant to a given dimension.

Number of response options

Researchers must consider the number of response options available to participants. A minimum of two is required (e.g., Agree/Disagree, True/False), but a larger number has benefits and costs. A potential benefit is that a relatively large number of options allows for finer gradations—potentially revealing subtler psychological differences among respondents than is possible with scales having few options. For example, a scale with response options that include “strongly agree,” “moderately agree,” and “slightly agree” (versus similar options of disagreement) allows participants to express degrees of agreement, whereas a dichotomous agree-versus-disagree scale does not. The potential cost of having many response options is potentially increased random error, as participants attempt to interpret overly-subtle gradations in the response format. In practice, researchers often use five or seven response options, balancing fine-gradation, subtlety, and psychometric quality.

Labels/anchors for response options

A third issue is the labeling of response options. Response options are often presented in terms of numerical values, and researchers must decide whether to supply verbal labels for some or all options. For example, researchers using a 7-point agreement format would, at minimum, label the endpoints (e.g., as strongly disagree and strongly agree). In addition, they might label one or more of the remaining points.

Research supports fully-labeled response options. That is, research reveals that labeling all response options produces better psychometric quality than does labeling only the endpoints, but there are limits to this benefit (Krosnick et al., 2005). For example, researchers will likely have difficulty providing clear, effective, widely-understood labels for a large number of response options. With a large number of response options, researchers might use terms like “somewhat agree,” “mildly agree,” “agree,” and “strongly agree” to differentiate shades of agreement. However, such differentiation might not be clear to respondents, or might not be interpreted similarly by all respondents—one person’s “somewhat agree” might be psychologically equivalent to another person’s “mild agreement.” Such ambiguity suggests that full-labeling of scales with more than six or seven response options might create more confusion than clarity.

Researchers labeling multiple response options should consider several straightforward yet practical issues. Labels should clearly differentiate the psychological meaning of the options. In addition, they should represent psychologically-equal differences among the response options, as much as possible (see
Chapter 2’s comments regarding interval level of measurement). For example, the first three options might be labeled “Very Strongly Disagree = 1,” “Strongly Disagree = 2,” and “Mildly Disagree = 3.” Unfortunately, the psychological difference between the first two options may be smaller than the psychological difference between the second and third options, and thus the labels should be revised. Finally, the labels should represent a broad range of the dimension being assessed. For example, response labels for an agreement scale should represent a broad range of “levels of agreement.”

Mid-points

Following from the previous issues, the use of psychologically-neutral mid-points in a set of response options is a frequent consideration in scale construction. For example, an “agreement” scale might include an option reflecting a neutral, undecided, or non-committed level of agreement (versus disagreement). In terms of labels, mid-points are presented with terms such as “neutral” or “neither agree nor disagree.” Alternatively, a scale might not include a neutral point, providing options referring only to various degrees of agreement or disagreement. A psychological mid-point is often achieved through an odd-number of response options—e.g., a 5-point scale has a natural mid-point, whereas a 4-point scale does not.

There may be costs associated with psychological mid-points. Specifically, mid-points might be an “easy way out” for respondents who are unmotivated or unable to think carefully about the scale. Thus, mid-points might elicit less psychologically-informative or less accurate responses from some participants. Indeed, some researchers avoid mid-points, hoping to thereby “force respondents to go one way or the other”—for example, forcing participants to either agree or disagree with items.

However, the potential costs of mid-points are likely offset by important benefits, and—on balance—there is often good reason to use them. When participants do have genuinely-neutral responses to specific items, mid-points allow those responses, thereby enhancing the psychological accuracy of scale scores. Additionally and perhaps relatedly, some evidence suggests that the use of mid-points enhances scales’ psychometric quality (O’Muircheartaigh et al., 2000). In this research, when mid-points were unavailable, people having truly-neutral psychological positions were forced to choose inaccurate options.

“No opinion” and “Don’t know” response options

In some domains of psychology, researchers might wish to accommodate respondents who have no opinion regarding an item or who might not know their true psychological perspective regarding the item. Thus, researchers might consider “No opinion” or “Don’t know” response options.
Some researchers might thus be tempted to label mid-point response options as “I don’t know,” to allow responses from people who might claim no knowledge or opinion regarding an item; however, this practice is inadvisable. For example, a person taking an attitude scale might have a carefully-considered opinion that is genuinely neutral, reflecting a thoughtful recognition of good and bad qualities of the attitudinal object. Such well-considered, genuinely neutral opinions are meaningfully different from a lack of opinion or a lack of knowledge, as would be reflected in a “No opinion” or “Don’t know” response. Indeed, empirical analysis of mid-points and “Don’t know” responses suggests that treating “Don’t know” responses as a psychological mid-point compromises psychometric quality (O’Muircheartaigh et al., 2000). Thus, researchers should avoid treating a “No opinion” or “Don’t know” response as being midway between two poles of the underlying psychological dimension.

More generally, a recent review of empirical work regarding “No opinion” and “Don’t know” response options suggests that such options (even as separate from mid-points) are inadvisable (Krosnick et al., 2005). When they are available and respondents choose them, those choices seem to reflect issues other than genuine lack of knowledge or opinion. Rather, they reflect issues such as low motivation, genuine ambivalence about an object, or ambiguity in the question itself.

Considering these problems, the best solution is to create scales that have simplicity, clarity, and breadth of the underlying psychological dimension. For example, researchers might construct easily-read scales that minimize respondent fatigue, which might otherwise generate low motivation to respond thoughtfully. Similarly, they can write items that are as clear and simple as possible, reducing confusion or frustration. Finally, researchers should again make sure that response options reflect a broad range of the underlying dimension, potentially capturing nuanced perspectives of people with ambivalence toward the topic of the assessment.

Consistency across items

When constructing a psychological scale, researchers should consider at least two issues regarding the consistency of response options across items. First and perhaps most obviously, a scale’s items should have equal numbers of response options. This is important because typical scoring procedures require researchers to sum or average across scale items, with all items equally weighted. Equal weighting is most likely to occur when items have equal numbers of response options. The second consideration is that the logical order of the response options should be consistent across items. For example, if one item’s responses are ordered “Strongly Disagree,” “Disagree,” “Neutral,” “Agree,” and “Strongly Agree” (from left to right or from top to bottom), then all items’ response options should be ordered similarly. If the order differs across items, then respondents
might misread the scale and, consequently, provide inaccurate responses. Researchers rightly worry about respondents’ motivation, attention, and ability to respond thoughtfully to item content, and the use of consistently-ordered response options eliminates one factor that might compromise these.

Assembling and Writing Items

Along with choosing a response format, researchers must assemble the scale’s items—either by borrowing or modifying items from other scales or by writing new items. The current section presents suggestions for writing effective items (or for borrowing/modifying items).

Relevant content

Most fundamentally and obviously, items’ content must reflect the intended psychological variable. This, of course, requires researchers to have a clear definition and understanding of the variable in question (Step 1 in the scale construction process) and to let this understanding guide item-writing. In addition, across all items, the breadth of the variable must be reflected in the scale’s content. Many important psychological constructs are broad in scope, having several facets or modes of manifestation. For example, researchers intending to measure Extraversion might conceptualize it in terms of behavioral, cognitive, motivational, affective, and physiological components. If so, then their items should reflect all of these components. In contrast, the researchers might intend to measure only the behavioral component of Extraversion. Such a scale would not include item content reflecting the other components; however, it would be a measure of the behavioral component of Extraversion, not of Extraversion more broadly.

Number of Items

A second important issue in assembling items is the number of items. Researchers must consider this issue for each construct to be measured by the scale, with each having its own set of items and receiving its own score. The optimal number of items depends upon several issues. First, traditional psychometric theory suggests that, all else being equal, longer scales have better reliability than shorter scales. Second, scales intended to reflect broadly-defined constructs may require more items than do scales reflecting narrowly-defined constructs. That is, to capture broad constructs such as Extraversion or Psychological Well-Being with good reliability and sufficient content-coverage, scales are likely to require a relatively large number of items. Although brief scales of some broad constructs have been developed, such scales are generally developed with rigorous attention to
psychometric criteria, and they do not provide multidimensional information about the separate facets of the constructs. A third consideration in determining an optimal number of items is the likely context of administration (again, Step 1 of the scale construction process). As discussed earlier, a scale that is likely to be administered in time-sensitive contexts might need to be shorter than would otherwise be preferred.

Procedures presented later in this volume allow researchers to evaluate costs and benefits of different numbers of items, thereby enhancing scales’ efficiency. For example, item-analysis, the Spearman–Brown prophecy formula, and Generalizability Theory allow researchers to identify good or bad items and to forecast the likely psychometric quality of scales of differing lengths. Such information, obtained in Step 4 of the scale construction process, is extremely useful when deciding whether to revise a scale and, if so, how to do so.

Clarity of language

To enhance the likelihood that participants will provide meaningful information, researchers should avoid things that reduce their motivation and ability to do so. Perhaps the most important such task is generating items that are easily-understood by potential respondents. Items and instructions that are relatively clear and simple are likely to be understood by respondents and to require little cognitive effort, enhancing respondents’ ability and motivation to provide psychologically-meaningful responses.

There are several strategies for maintaining clear language in a scale’s items. One fundamental recommendation is to avoid complex words (considering the likely target population), including psychological jargon. A second recommendation is to avoid double-negatives. For example, a Conscientiousness scale intended for students might include either of two items—“I never fail to do my homework” or “I always do my homework.” Of course, most (hopefully all) students could understand the first item; however, doing so likely requires expending more cognitive effort than would be required by the second item. Consequently, the double-negative (i.e., “never” and “fail”) of the first item might begin to diminish respondents’ motivation and ability to complete the scale with care and accuracy, and several such items might cause problems for many respondents. Third, researchers should avoid double-barreled items—items reflecting two separable questions or statements. For example, another problematic potential Conscientiousness item might read, “I feel ashamed when I perform poorly on a test, therefore I avoid telling my parents about it.” Specifically, what is the appropriate response from someone who feels ashamed about performing poorly, but who does tell his or her parents? Or what is the appropriate response from someone who avoids telling his or her parents about a poor performance, but who does so for reasons other than shame? The answers might be clear with some thought,
but once again, such items place unnecessary cognitive demands on respondents. Thus, double-barreled items introduce confusion and, potentially, error into the scale.

Not leading or presumptive

The hypothetical Conscientiousness item ‘I feel ashamed when I perform poorly on a test, therefore I avoid telling my parents about it’ reflects an additional subtle problem. Specifically, it presumes that a respondent performs poorly on tests, at least occasionally. That is, the item reads “… when I perform poorly on a test …,” but some student respondents might not perform poorly on tests. What is the appropriate response from such students—disagreement with the statement, a neutral answer, a refusal to answer? Each of these responses creates ambiguity or missing data. For example, disagreements made by students who never perform poorly would be inseparable from: a) disagreements made by students who perform poorly but do not feel ashamed; or from b) disagreements made by students who perform poorly and who feel ashamed, but who do not avoid telling their parents.

Balanced scales

As a general rule, scales should be ‘balanced’ by including positively-keyed and negatively-keyed items. Positively-keyed items are those for which agreement or endorsement indicates a high level of the psychological variable being assessed. For example, the item “On the whole, I am satisfied with myself” is a positively-keyed item on the Rosenberg Self-esteem Scale—a genuine endorsement (i.e., a response of “Agree” or “Strongly Agree”) reflects a high level of self-esteem, and rejection reflects lower self-esteem. In contrast, negatively-keyed items are those for which agreement or endorsement indicates a low level of the psychological variable. For example, the item “At times, I think I am no good at all” is a negatively-keyed item—genuine endorsement reflects a low level of self-esteem. Responses to negatively-keyed items must be reversed in the scoring process.

Balanced scales reduce the effects of acquiescence bias. As will be discussed (Chapter 6), acquiescence (or yea-saying) bias occurs when respondents agree to or endorse items without regard for content, and nay-saying occurs when respondents disagree to or refuse to endorse items without regard for content. Both biases can compromise the psychometric quality of scale scores, but balanced scales reduce the effects. Specifically, balanced scales reduce the possibility that relatively high (or low) scores will be obtained purely on the basis of acquiescence (or nay-saying). On balanced scales, participants responding purely with an acquiescent bias will obtain scores near the scale mid-point, which likely means that their scores will not create spuriously-large differences or positive (or negative) correlations. Without the use of balanced scales, researchers risk drawing inaccurate psychological conclusions from spuriously-large effects.
Summary

This chapter discussed two crucial facets of scale construction. When choosing response formats, researchers must consider issues such as the appropriate number of response options, the labeling of options, and the use of psychological mid-points. When writing items, they should consider issues such as the clarity of language, relevant content, and the use of balanced scales. Well-considered scale construction can increase the chance that respondents will complete the scale with proper motivation, good attention, and sufficient ability.