Scale Construction and Psychometrics for Social and Personality Psychology

R. Michael Furr
This chapter presents principles and practices that are among the broadest and most fundamental issues for scale construction, modification, use, evaluation, and interpretation. The points are rather straightforward but are vitally important in conducting informative research. Thus, this chapter provides nontechnical overviews of each point, to be complemented by greater exploration and depth later in this volume.

Several of these principles and practices strike me, as an editor, reviewer, and reader of social/personality research, as being somewhat under-appreciated. To retain focus on those issues, this chapter bypasses issues that, though fundamental to scale construction and psychometrics, seem generally well-known and well-implemented. Indeed, much social/personality research is based upon measurement that is well-conceived and appropriately-executed. This discussion is intended to raise awareness and understanding of issues that, if appreciated even more widely, will enhance the generally good conduct and interpretation of research. The issues are summarized in Table 2.1.

Most facets of the process and principles covered in this chapter apply to all forms of psychological measurement. For example, this chapter addresses the need to articulate the construct and context of a measurement strategy, the need to evaluate psychometric properties, and the need to revise the measurement strategy if necessary—all of which apply to measurement strategies such as “tests” of maximal performance, reaction time, behavioral observations, physiological measures, choices and decisions, informant-reports, and so on.

In addition, this chapter outlines scale construction in terms of four steps (Figure 2.1). Reflecting contemporary social/personality psychology (John & Benet-Martinez, 2000), this chapter (and this volume more generally) blends several approaches to scale construction. It involves rationally-focused item-writing, attention to scale dimensionality and internal coherence, and empirical examination of the scale’s psychological meaning.
## Table 2.1 Under-appreciated principles and practices in scale construction, use, evaluation, and interpretation

<table>
<thead>
<tr>
<th>Principles and practices</th>
<th>1</th>
<th>Examination and interpretation of reliability and validity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a)</td>
<td>Psychometric properties and quality of the current data should be evaluated</td>
</tr>
<tr>
<td></td>
<td>b)</td>
<td>Psychometric properties and quality of the current data should be considered when using scales and when drawing psychological implications from observed effects</td>
</tr>
<tr>
<td>2</td>
<td>Dimensionality (i.e., factor structure)</td>
<td></td>
</tr>
<tr>
<td>a)</td>
<td>Dimensionality should be evaluated and considered in scale construction, use, and evaluation</td>
<td></td>
</tr>
<tr>
<td>b)</td>
<td>Coefficient alpha should be not an index of unidimensionality, the “eigenvalue greater than one” rule should be avoided, and oblique rotations are preferable to orthogonal</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Ad-hoc scales</td>
<td></td>
</tr>
<tr>
<td>a)</td>
<td>Previously-validated scales are preferable to ad hoc scales</td>
<td></td>
</tr>
<tr>
<td>b)</td>
<td>Ad hoc scales should be examined psychometrically, requiring more than assumed face validity</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Modified scales</td>
<td></td>
</tr>
<tr>
<td>a)</td>
<td>Previously-validated original scales are preferable to modified scales</td>
<td></td>
</tr>
<tr>
<td>b)</td>
<td>Modified scales should be examined psychometrically, not assumed to have the same dimensionality, reliability, and validity of original scales</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Brief/single-item scales</td>
<td></td>
</tr>
<tr>
<td>a)</td>
<td>Sufficiently-long scales are preferable to overly-brief scales</td>
<td></td>
</tr>
<tr>
<td>b)</td>
<td>Brief/single-item scales should be examined psychometrically, and their psychometric quality and potential limitations should be appropriately discussed</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>The use of scales across groups</td>
<td></td>
</tr>
<tr>
<td>a)</td>
<td>If scales are used in groups that might differ psychologically from those within which the scales were developed and validated, then the psychometric properties/differences should be examined, understood, and potentially rectified.</td>
<td></td>
</tr>
<tr>
<td>b)</td>
<td>Careful translation does not guarantee psychometric stability across linguistically-differing groups</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Difference scores (i.e., gain scores, change scores, discrepancy scores)</td>
<td></td>
</tr>
<tr>
<td>a)</td>
<td>Alternatives to difference scores should be considered strongly.</td>
<td></td>
</tr>
<tr>
<td>b)</td>
<td>Difference scores should be used with attention to their components and their psychometric quality.</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Advanced psychometric perspectives (e.g., Confirmatory Factor Analysis, Generalizability Theory, and Item Response Theory)</td>
<td></td>
</tr>
<tr>
<td>a)</td>
<td>Advanced perspectives offer some important differences/advantages, as compared to traditional psychometric theory; thus, they might be highly-appropriate methods of understanding and evaluating the psychometric properties of a given scale</td>
<td></td>
</tr>
<tr>
<td>b)</td>
<td>Producers and consumers of research should be prepared to provide and/or interpret information obtained from these perspectives when necessary</td>
<td></td>
</tr>
</tbody>
</table>

### Examination and Interpretation of Reliability and Validity

First and most broadly, the psychometric quality of the current data should be evaluated and considered when interpreting results. Reliability and validity are
crucial in understanding statistical results and their psychological implications. Roughly stated, reliability is the precision of scores—the degree to which scores accurately reflect some psychological variable in a given sample. Validity, then, concerns the “some variable” reflected by those scores—specifically, validity is the degree to which scores can be interpreted in terms of a specific psychological construct. Note that scores can be reliable (i.e., they can be good indicators of something), but—at the same time—they can be interpreted invalidly (i.e., they can be interpreted in terms of a construct that they do not truly reflect).

Thus, there at least two issues that should be addressed in any psychological study. The first is the psychometric properties and qualities of the measures used in the study. Reliability and validity are fundamental facets of psychometric quality, and as such, researchers should provide evidence regarding the nature and strength of the reliability and validity of any scale, test, assessment, or dependent variable used in a study—is the scale performing well in the sample being studied, do the scale’s scores truly reflect the construct that researchers wish to measure? The second issue is the implications that scales’ reliability and validity have for analysis and psychological implications. Among their effects, reliability affects one’s statistical results, and validity affects one’s ability to interpret results in terms of specific psychological phenomena.
Without proper understanding of the psychometric properties of the measures in a given study, researchers and readers cannot be sure that those measures were used and interpreted appropriately. Despite this importance, fundamental psychometric information is sometimes omitted from research reports. Unfortunately, we cannot assume confidently that the reliability of a scale’s scores in one study or with one sample of participants generalizes to all studies or all samples. Thus, each time a scale is used, its scores should be evaluated in terms of psychometric quality. This volume and others (Nunnally & Bernstein, 1994; Furr & Bacharach, 2008) provide broad backgrounds for addressing psychometric quality.

Reliability can often be estimated quite easily for multi-item scales, and researchers usually assume that validity evidence generalizes across similar samples of participants. However, both of these practices are limited, as discussed later.

**Dimensionality**

A scale’s dimensionality, or factor structure, reflects the number and nature of variables assessed by its items. Some questionnaires, tests, and inventories are unidimensional, with items cohering and reflecting a single psychological variable. Other questionnaires are multidimensional, with sets of items reflecting different psychological variables.

Usually based upon factor analysis, an accurate understanding of the number and nature of a scale’s dimensionality directly affects its scoring, psychometric quality, and psychological meaning. Dimensionality dictates the number of meaningful scores that a scale produces for each participant. If a scale includes two independent dimensions, its items should be scored to reflect those dimensions. For example, the Positive Affect Negative Affect Schedule (PANAS; Watson et al., 1988) is a multidimensional questionnaire that produces one score for Positive Affect (PA) and another for Negative Affect (NA). Researchers should not combine items across the two dimensions, as it would produce scores reflecting no coherent psychological variable. By dictating the number of meaningful scores derived from a questionnaire, dimensionality also directs researchers’ evaluations of reliability and validity. That is, researchers must understand the psychometric quality of each score obtained from a questionnaire. For example, the PANAS has been developed and used with psychometric attention to each of its two “subscales.” Thus, researchers who develop and use psychological scales must understand the dimensionality of those scales. This is true even for short scales that might appear to reflect a single psychological variable. Inaccurate understanding of dimensionality can produce scores that are empirically and psychologically meaningless.

An important related point is that a scale’s dimensionality is not clearly reflected by the familiar Cronbach’s coefficient alpha. Based upon a scale’s internal
consistency, alpha is an estimate of a scale’s reliability; however, it is not an index of unidimensionality. That is, a large alpha value cannot be interpreted as clear evidence of unidimensionality (see Chapter 4).

Finally, there are several recommendations that contradict many applications of factor analysis in evaluating the dimensionality of psychological scales. One is that the “eigenvalue greater than one” rule is a poor way to evaluate the number of dimensions underlying a scale’s items; other procedures, such as scree plots, are preferable. A second recommendation is that oblique rotations are preferable to orthogonal rotations. These recommendations are detailed in Chapter 4.

Ad Hoc Scales

Occasionally, researchers create scales to measure specific constructs for a study. Of course, scale-development is important for psychological research, and there are good reasons to create new scales (e.g., one might wish to measure a construct for which a scale has not yet been developed). However, there are two caveats related to ad hoc scales.

The first caveat is that previously-validated scales are generally preferable to ad hoc scales. That is, when well-validated scales exist for a given construct, researchers should strongly consider using those scales rather than a new scale. For example, there are many well-validated self-esteem scales differing in length, psychological breadth, and dimensionality. With such diversity and known psychometric quality within easy reach, there seems little reason to assemble a new ad hoc self-esteem scale.

The second caveat is that, if ad hoc scales are created, they require psychometric evaluation, including validity evidence that goes beyond face validity. Ideally, scales that are intended to measure important psychological variables are developed through a rigorous process emphasizing psychometric quality. However, ad hoc scales sometimes seem to be applied without much evaluation, with researchers apparently relying upon on face validity and assuming that items obviously reflect the intended construct. Not only should researchers examine the dimensionality and reliability of ad hoc scales, but they should strive to obtain and report independent evidence of validity. For example, researchers might recruit independent raters (e.g., colleagues or students) to read the items along with items intended to reflect other variables, ask the raters to rate the clarity with which each item reflects each of several variables, and evaluate the degree to which each item was rated as clearly reflecting its intended variable. Such examinations produce validity-related evidence that goes beyond the researcher’s opinions, and they might convince readers that a scale is sufficiently valid for narrowly-focused application.
Modified Scales

In addition to creating ad hoc scales, researchers sometimes modify existing scales for new purposes. For example, researchers might shorten an existing scale or might revise a measure of one variable (e.g., health locus of control) to “fit” a different variable (e.g., financial locus of control). Again, there can be good reason for such modifications (e.g., existing scales are perceived as too lengthy, or there is no existing measure of the variable in question).

However, there are caveats that arise with modified scales, paralleling those associated with ad hoc scales and arising from the fact that modified scales might not have the psychometric properties or quality of an original scale. Indeed, the more substantially-modified a scale is, the less it can be assumed to have psychometric quality similar to the original. Thus, one caveat is that well-validated, original scales are preferable to modified scales. Because a modified scale’s psychometric properties and quality might differ from those of the original scale, the modified scale is—to some degree—an ad hoc scale. As such, its psychometric properties and quality are unclear and suspect. Consequently, a second caveat is that modified scales require psychometric evaluation and interpretation. Because the psychometric properties of a modified scale likely differ from those of the original, researchers should carefully examine the scale’s dimensionality, reliability, and validity.

Brief or Single-item Scales

An issue related to the creation or modification of scales is the use of brief scales—scales with very few items, perhaps only a single item. Indeed, brief scales are appealing, particularly when participants cannot be burdened with long scales.

Unfortunately, brief scales have important psychometric costs—their psychometric quality might be, likely is, poor or even unknown. As we shall see, traditional reliability theory suggests that reliability is relatively weak for brief scales, all else being equal. For example, recent research (van Dierendonck, 2005) compared three versions of a “Purpose in Life” scale—a 14-item version, a 9-item version, and a 3-item version. Results showed weaker reliability for shorter versions, with reliability estimates of $\alpha = .84$, $\alpha = .73$, and $\alpha = .17$ for the three versions, respectively. The particularly-poor reliability of the 3-item version suggests that “it is troublesome if the scales are to be used as variables in correlational analysis. Low reliability diminishes the chance of finding significant correlations” (van Dierendonck, 2005, p. 634). In fact, low reliability is problematic not only for typical “correlational analyses” but for any analysis of group differences or associations (see Chapter 4). A second difficulty particular to single-item scales is that
they preclude the use of internal consistency methods for estimating reliability (e.g.,
coefficient alpha). Because internal consistency methods are easier than most other
methods of estimating reliability, single-item scales are often used without attention
to psychometric quality. This is a serious problem, preventing researchers,
reviewers, editors, and readers from knowing the quality of a measurement that,
being based upon single-item scale, is inherently suspect. Thus, researchers who
use single-item scales might examine test–retest studies in order to estimate the
reliability of the scales.

Importantly, brief or even single-item scales can have psychometric properties
sufficient for researchers facing strict constraints in measurement strategies. For
example, the Single-Item Self-Esteem Scale has good test–retest reliability esti-
mates, and it has strong convergent validity correlations with the widely-used
Rosenberg Self-Esteem Inventory (Robins et al., 2001). Similarly, the Ten-Item
Personality Inventory has good test–retest estimates of reliability for each of its
2-item scales, which have strong convergent validity correlations with longer
measures of their constructs (Gosling et al., 2003). The important point is that
brief scales are appropriate and useful when their psychometric properties are
adequate, as demonstrated by solid empirical estimates.

Scale Use Across Psychologically-differing Groups

Researchers often assume that psychometric properties generalize across sam-
ples of participants, but this assumption is not always valid. Indeed, a scale’s
psychometric properties might be importantly different in differing groups, and
this might be particularly problematic for research in which scales are trans-
ported across cultural groups. Whether due to group differences in the interpreta-
tions of words, question/items, instructions, or the general meaning of a set of
items, it is possible that “the items of the scale do not similarly represent the
same latent construct … across groups” (Tucker et al., 2006, p. 343). In such
situations, the scale’s psychological meaning differs across groups and “the
accuracy of interpretations about group differences on the latent construct is
compromised” (ibid.).

Several considerations are important when a scale is used in groups that might
differ psychologically from the group in which it was initially developed and
validated. First, psychometric properties should be examined within each new
group, and psychometric differences should be examined, understood, and poten-
tially rectified before scores are used in those groups. If the new group might
differ in the interpretation of a scale or in terms of the scale’s link to the construct
of interest, then researchers should explore this possibility. Results will either
reveal that the scale is similarly meaningful in the new group, or they will reveal
psychologically-interesting differences.
To establish the comparability of a scale across groups, researchers examine “test bias” (see Chapter 6), “measurement invariance,” or differential item functioning (see Chapter 10). For example, Tucker et al. (2006) examined the Satisfaction With Life Scale (SWLS) in North Americans and Russians, finding that scores were not strongly comparable across groups. Such results suggest that comparison of groups’ averages might produce misleading conclusions.

The second consideration is that careful translation does not guarantee psychometric stability across groups. Many researchers should be commended for careful attention in translating scales into a new language. Indeed, such attention is crucial for good cross-cultural work; however, it does not guarantee that the scale’s psychometric properties or meaning are comparable. Again, the comparability of scale’s properties, and ultimately of its psychological meaning, are revealed through psychometric analysis of responses to the scale.

**Difference Scores**

One important topic addressed in this volume is the use of difference scores as indicators of psychological phenomena. Difference scores—or “change scores” or “discrepancy scores”—are obtained by measuring two component variables and computing the difference between the two. For example, a participant’s intergroup bias might be measured by having her rate the positivity of an ingroup and of an outgroup, and then calculating the difference between the two ratings. The difference might be interpreted as the degree to which she has a more favorable attitude toward the ingroup than toward the outgroup.

Difference scores are appealing, but their intuitive appeal masks complexities that have been debated for decades. Although much debate highlights the supposed unreliability of difference scores, there is also concern that difference scores can lack discriminant validity, in terms of simply reflecting one of their component variables. Both issues can compromise the psychological conclusions based upon difference scores.

This volume discusses these complexities, including psychometrically-based and statistically-based recommendations for handling them. First, and most generally, alternatives to difference scores should be considered seriously. That is, difficulties might be handled best by avoiding difference scores altogether, focusing instead on their component variables. Second, if difference scores are used, they should be used with attention to their components and to their psychometric quality. There seems to be a tendency to ignore the fact that difference scores are variables with psychometric properties that must be understood and considered when drawing psychological conclusions. Without such understanding and consideration, research based upon difference scores is ambiguous in terms of measurement quality, statistical validity, and psychological meaning.
Advanced Psychometric Perspectives

Advanced psychometric perspectives or tools such as Confirmatory Factor Analysis, Generalizability Theory, and Item Response Theory are increasingly accessible. As they become more well-known and well-integrated into user-friendly statistical software, they may become more important for scale development and evaluation.

Such perspectives offer important differences from and advantages over traditional psychometric theory, and they can be useful for understanding and evaluating the psychometric properties of psychological measures. Furthermore, producers and consumers of psychological research should be prepared to provide and/or interpret information obtained from these perspectives. Given the advantages of these perspectives, they may be the optimal choices for some—perhaps much—of the work in scale development and evaluation. This volume presents important principles of these perspectives, with examples that lay foundations for conducting and interpreting these important psychometric perspectives.

Steps in Scale Construction

Scale construction can be seen as a four-step process that is often iterative (Figure 2.1). Although each step is important, some are ignored in some scale construction procedures. Unfortunately, bypassing any of these steps might produce a scale with unknown psychometric quality and ambiguous meaning. High-quality research requires serious attention to scale construction and evaluation.

Step 1: Articulate the Construct and Context

The first, and perhaps most deceptively-simple, facet of scale construction is articulating the construct(s) to be measured. Whether the construct (one or more) is viewed as an attitude, a perception, an attribution, a trait, an emotional response, a behavioral response, a cognitive response, or a physiological response, or—more generally, a psychological response, tendency, or disposition of any kind—it must be carefully articulated and differentiated from similar constructs. Is more than one construct to be measured? What is the exact psychological definition of each construct? Is each construct narrow or broad? Does the construct have subcomponents or dimensions that should be differentiated and measured? What are the likely associations and differences between the intended construct and other relevant psychological constructs? Such questions guide subsequent steps in scale construction and evaluation, ultimately determining the scale’s meaning and quality. For example, if an intended construct is not clearly differentiated from other constructs, then
subsequent steps might produce a scale with poor validity and ambiguous meaning.

In addition, researchers creating a new scale must articulate the context in which it is likely to be used. The context includes at least two elements—the likely target population and the likely administration context. Obviously, the intended population will direct subsequent steps of item writing. For example, the response formats, items, or instructions for a scale intended to be used with adults would differ dramatically from those for one intended to be used with children. As discussed earlier, researchers cannot assume that a scale developed for and validated within one population is psychometrically comparable or similarly meaningful in different populations. Similarly, the likely administration context(s) must be considered carefully. For example, if the scale will be used primarily in research contexts that are time-sensitive, then subsequent steps will likely focus on brevity. Or, for example, if the scale will be administered via an online survey, then researchers should consider implementing online strategies in Step 3 of the construction process.1

Step 2: Choose Response Format and Assemble Initial Item Pool

In the second step of scale construction, researchers choose a response format and assemble an initial item pool. Guided by considerations from the first step, researchers write or seek out items that seem psychologically relevant to the intended construct. Of course, this depends on factors such as the number of constructs to be measured, the intended length of the scale, and the clarity of the construct’s definition.

As discussed in Chapter 3, this step often includes iterative sub-steps in which items are discussed, considered in terms of conceptual relevance and linguistic clarity, and discarded or revised. In addition, this work may lead researchers to revisit the first step—potentially re-conceptualizing the focal construct(s). Indeed, a thoughtful item-writing process can reveal shortcomings in a scale’s conceptual basis.

Step 3: Collect Data

After one or more constructs have been articulated, the likely assessment context has been determined, and items have been assembled, the items should be administered to respondents representing the likely target population, in a manner reflecting the likely administration context. This step has at least two purposes. First, it can reveal obvious problems through respondent feedback or observation. For example, respondents might require more time than initially supposed, or they might express confusion or frustration. Such issues might require revision of the scale. Second, this step produces data for the next step of scale construction—evaluation of the item pool’s psychometric properties and quality.
Step 4: Psychometric Analysis

Scale construction requires attention to the psychometric properties of proposed items and of the proposed scale as a whole. By collecting data in a representative administration context and attending to dimensionality, reliability, and validity, researchers enhance the possibility that the scale will be useful and psychologically informative. Without such attention, even scales with the most straightforward appearance might be psychologically ambiguous or meaningless. Subsequent chapters articulate principles and processes important for psychometric evaluation.

The results of psychometric analyses determine subsequent phases of scale construction. If analyses reveal clear psychometric properties and strong psychometric quality, then researchers might confidently complete scale construction. However, psychometric analyses often reveal ways in which scales could be improved, leading researchers back to item (re)writing. In addition, psychometric analyses occasionally even lead researchers to re-conceptualize the nature of the construct(s) at the heart of the scale. Upon rewriting, the newly-revised scale should be evaluated in terms of its psychometric properties. This back-and-forth process of writing, analysis, and re-writing might require several iterations, but the result should be a scale with good psychometric quality and clear psychological meaning.

General Issue: Scale of Measurement

One general issue that sometimes escapes scrutiny is whether a scale produces scores at an interval level of measurement. At an interval level of measurement, the underlying psychological difference between scores is constant across the entire range of scores. Consider a hypothetical 1-item scale measuring homophobia: “I avoid homosexual people,” with response options of 1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always. To interpret scores at an interval level of measurement, researchers must believe that the size of the psychological difference (in terms of underlying homophobic attitudes) between “Never” and “Rarely” avoiding homosexual people is identical to the size of the psychological difference between “Rarely” and “Sometimes” avoiding homosexual people. That is, the psychological difference between a score of 1 and a score of 2 is identical to the psychological difference between a score of 2 and a score of 3. There is serious debate about whether this is true for many psychological scales.

Level of measurement can have implications for the meaningfulness of specific forms of statistical analysis, though there is disagreement about this. Strictly speaking, a scale that is not at least at the interval level of measurement is difficult to interpret in terms of analyses based upon linear regression (as are ANOVA-based procedures). For example, an unstandardized regression slope reflects the difference in an outcome variable associated with a one-unit difference in a predictor
variable. The psychological meaning of this is clear only when “a one-unit difference in a predictor variable” is consistent across levels of that variable. This is true for interval scales, but not for ordinal or nominal scales.

Regardless of ambiguities and disagreements, researchers generally treat Likert-type scales (such as the hypothetical homophobia scale) as an interval level of measurement. Particularly for aggregated scores obtained from multi-item scales, researchers assume that scores are “reasonably” interval-level. For very brief or single-item scales, this assumption is more tenuous. In such cases, researchers should either consider alternative analytic strategies or acknowledge the potential problem.

**Summary**

This chapter has reviewed principles and recommendation for scale construction, evaluation, and use, and has summarized the scale construction process. The remainder of this volume provides greater depth into the process, principles, and practices, hopefully enhancing motivation and ability to pursue effective measurement procedures.

**Note**

1 The possibility that online surveys differ meaningfully from traditional methods has received some empirical attention. For example, Gosling et al. (2004) found that web-based surveys produce results similar to those produced by traditional methods; however, they note that “this question has yet to be resolved conclusively” (p. 102).