ONE

INTRODUCTION AND OVERVIEW

PURPOSE OF THE BOOK

Measurement is at the heart of virtually all scientific endeavors. Measures, and the psychometric properties used to evaluate them, will vary by the type of measurement undertaken and the context and goals of the scientific endeavor. There are several strategies for developing and refining measurement/assessment instruments, and the relevance of a given strategy will depend on what type of scientific phenomenon is being assessed and its underlying measurement model. In this book, we focus on developing and validating paper-and-pencil measures of latent social-psychological constructs. Furthermore, the constructs focused upon are perceptual; a respondent rates himself or herself or others on constructs that are subjective or opinion-based. Mental/ability testing and classification measurement for clinical diagnosis are not an emphasis of this text. Although several of the principles of mental/ability testing and measurement for clinical diagnosis are applied in developing and validating social-psychological measures, such as classical test theory and generalizability theory, other principles more akin to ability/mental testing and clinical diagnosis are not emphasized. (The interested reader is referred to Anastasi and Urbina [1998], Crocker and Algina [1986], Hambleton, Swaminathin, and Rogers [1991], Haynes, Nelson, and Blaine [1999], and Kaplan and Saccuzzo [1997] for a discussion of mental/ability testing and clinical diagnosis.)

Given their latent nature, the constructs we focus on represent abstractions that can be assessed only indirectly. The indirect assessment of these constructs is accomplished via self-report/paper-and-pencil measures on which multiple items or indicators are used to measure the construct, that is, "scaling" a construct. Although measurement of psychological constructs also involves "classification," that is, defining whether objects fall in the same or different categories with respect to a given attribute, this text is on scaling. As such, the purpose of this book is to discuss the issues involved in developing and validating multi-item scales of self-report/paper-and-pencil measures.

PERSPECTIVES ON MEASUREMENT IN THE SOCIAL SCIENCES

What Is Measurement?

At its core, measurement consists of rules for assigning symbols to objects to numerically represent quantities of attributes. Measurement includes evaluating numbers such that they reflect the differing degrees of the attribute being assessed (DeVellis, 1991; Haynes et al., 1999; Nunnally & Bernstein, 1994). In the social sciences, most of the time the "objects" are people, "rules" involve the explicitly stated assignment of numbers, and "attributes" are particular features of the objects being measured. As such, it is important to note that objects (e.g., people) are not measured; their attributes are measured (e.g., self-esteem).

Rules of measurement require a bit more explanation. Some rules are obvious and universal, such as measuring weight in pounds or kilograms. Rules for measuring social-psychological constructs are not so obvious. For example, what are appropriate rules for measuring constructs such as self-esteem, job satisfaction, and consumer self-confidence? Although there are no "universal" rules for measuring such constructs, developing rules that are eventually accepted is important for standardization and establishing norms. A measure is standardized when (a) rules of measurement are clear, (b) it is practical to apply, (c) it is not demanding of the administrator or respondent, and (d) results do not depend on the administrator (Nunnally & Bernstein, 1994). Such a measure yields similar results across applications (i.e., the measure is reliable) and offers scores that can be easily interpreted as low, medium, and high.

The focus on measuring attributes also requires clarification. As stated, we are not measuring a person per se; we are measuring his or her attributes. This distinction is important because it emphasizes the abstract nature of social-psychological measurement. That is, we must "abstract" the attribute from the person. Many studies in the social sciences attempt to determine the relationship between two attributes (e.g., self-esteem and need for achievement). To avoid confounding among related attributes, the exact nature of the attribute must be carefully determined and specified. Furthermore, an assessment must be made if the attribute can be measured at all. As noted by Nunnally and Bernstein (1994), some attributes are so abstract that they may not be amenable to measurement (e.g., clairvoyance).

Usefulness of Social Science Measures

What properties constitute a measure's usefulness? As previously stated, there are multiple criteria (psychometric properties) that are used to evaluate measures. The criteria that are most relevant depend on the goals of the assessment and the scientific endeavor undertaken. Given that our text focuses on scaling latent social-psychological constructs, we focus on those psychometric properties most applicable to such constructs.

Although differing opinions exist, one view that seems to be shared by most social scientists is that results based on a measure should be repeatable and that the measure itself is standardized. *Repeatability* and *standardization* are related concepts. Under similar circumstances, a research finding based on the same measure should replicate. This is the basic tenet of repeatability—that the measure performs reliably under similar testing conditions.

Sound psychometric procedures for scale development include establishing norms. When these norms can be interpreted as describing a person as low, medium, or high on an attribute, the measure is felt to be standardized. Standardization has several advantages.

First, although we measure perceptions that by their very nature are subjective, a standardized measure enhances social science objectivity. When one researcher can independently verify a relation between two constructs that was found by another researcher, objectivity is enhanced—given the measures used are the same and are standardized. If disagreement exists as to the appropriateness of the measures used in obtaining the finding, objectivity is compromised. In the social sciences, we often test theories, but a theory can be

tested adequately only to the extent that the attributes of the theory (constructs) are adequately measured. When agreed upon procedures exist for measuring the attributes of interest, the objectivity of theory tests is enhanced.

Second, standardization produces quantifiable numerical results. Again, though this text does not address classification per se, such quantification does allow for the creation of categories (e.g., low, medium, high) for mathematical and statistical analyses (ANOVA), or for use as factor levels in experimental designs. Quantification also enhances communication and generalizability of results. Knowledge accumulates in the social sciences when researchers compare their results with the results of previous studies. When the same, standardized measures are used across scientific applications, results termed as "low" in self-esteem or "high" in self-esteem have common meaning across researchers. This enhances both the communication of results and generalizability of findings.

Third, measure/scale development is a time-consuming endeavor. If a measure has been well developed, however, the time spent is also time "well spent." Once standardization occurs, the measure is available for use with little or no time invested because of its agreed upon standards. At the very heart of repeatability and standardization are the measurement properties of *reliability* and *validity*. These two concepts are elaborated upon later in this chapter and extensively discussed and examined in Chapters 3 through 7.

Scaling of Latent Psychological Constructs With Multiple Items

As we have previously stated, social scientists focus on measuring attributes of objects that tend to be abstract. Such abstractions are latent by nature. Latent constructs are not directly observable or quantifiable. A latent construct is also variable; that is, the strength and magnitude for the ratings on a latent construct may change over time. For example, need for achievement represents a latent construct (i.e., a personal attribute). It cannot be observed directly by a researcher and thus requires a scale to estimate its actual magnitude at a given point in time. Furthermore, an individual's level of need for achievement may change over time.

Our text also focuses on latent social-psychological constructs that are theoretical in nature. That is, the construct being measured is embedded in a theoretical framework and/or has theoretical underpinnings. For example, the

construct of job satisfaction is at the center of theories of employee turnover (e.g., Tett & Meyer, 1993) and is concerned with its own theoretical content domain, for example, satisfaction with pay, satisfaction with supervisors, and satisfaction with coworkers. Conversely, some constructs are either empirical or atheoretical in nature. Opinion polls oftentimes assess purely empirical constructs. We are not suggesting that such constructs and measurement approaches have little value in the social sciences. In fact, they may prove useful for theory development. For example, an opinion poll (empirical and likely atheoretical) assessing political liberalism-conservatism and favoring or not favoring gun control could reveal that those who are politically liberal favor gun control more than those who are politically conservative, leading to the theoretical proposition that liberals favor gun control. We are merely stating that constructs based in theory are the focus of this text.

It is generally agreed that measures of latent theoretical constructs require multiple items or statements to more accurately reveal the varying levels of the constructs; that is, they are scaled (Clark & Watson, 1995; DeVellis, 1991; Haynes, Richard, & Kubany, 1995; Nunnally & Bernstein, 1994). Again, given that an object's (person's) level on an attribute that is latent and psychologically abstract cannot be directly measured, a scale must be constructed. Although sometimes it may be possible to infer a level of a latent psychological construct via behavior (e.g., from repeated brand purchase, one infers that an individual believes the brand is a good value for the money [latent construct]), many times a behavior may not be indicative of a latent construct. In such cases, a well-constructed and validated multi-item paper-and-pencil scale of the construct is needed.

Finally, we would like to reiterate that scaling, and not indexing, is our focus. In scaling, scores on items in the scale are theoretically driven by the latent construct; that is, they are "reflected" by the latent construct. With an index, scores on items (indicators) drive the total score of the index; that is, the items/indicators "form" the constructed index score. Although still latent in many respects, formative items/indicators are not considered scales because their scores are not necessarily reflected by the latent construct. An often-used example of formative items/indicators that result in an index is socioeconomic status (SES). Items or indicators might include income, education level, occupation, and dwelling type. Although some of these indicators have the latent property of not being directly observable, their scores are considered "forming" the index of SES and not vice versa (reflective). For a more detailed

discussion of "formative" vs. "reflective" measures, the interested reader is referred to Bollen and Lennox (1991), Diamantopoulus and Winklhofer (2001), MacCallum and Browne (1993), and Smith and McCarthy (1995).

Recent Trends in Scaling Latent Psychological Constructs

As succinctly stated by Clark and Watson (1995), "Scale development remains a growth industry within psychology" (p. 309). This statement applies to all related fields in the social sciences, including business functional areas such as marketing, accounting, management information systems, strategic management, and organizational behavior. In fact, several recent texts that have compiled scales often used in marketing and organizational behavior are now available (e.g., Bearden & Netemeyer, 1998; Bruner & Hensel, 1997; Price & Mueller, 1986; Robinson, Shaver, & Wrightsman, 1991). Two primary factors account for the recent interest in scale development.

First, as theories in the social sciences develop and evolve, so does the need to test them objectively. These theories require operationalizations of the constructs of interest. When the constructs are measured well (reliably and validly), theory testing is enhanced. Furthermore, it is often found that a once-used scale needs to be updated or refined to better reflect a construct of interest. Many measurement articles in the social sciences represent new scales, derived from existing measures, that are felt to more accurately or more efficiently reflect constructs of interest. Second, the advancement of computer and software technology has greatly helped our ability to develop measures. Statistical packages such as SPSS, SAS, BMDP, LISREL, EQS, CALIS, and AMOS, as well as their increased "user friendliness," have made it easier and quicker to perform most of the basic and many of the more advanced analyses recommended in scale development. It should be noted that there is a caveat here. Computer technology and statistical packages may allow for quicker scale development, but not necessarily better scale development. The recommended procedures advocated in this text and other books and articles should be adhered to in developing psychometrically sound scales (Clark & Watson, 1995; DeVellis, 1991; Haynes et al., 1999; Haynes et al., 1995; Nunnally & Bernstein, 1994; Spector, 1992).

Not only has there been a trend toward more scales, but the procedures used to develop and validate scales also have been evolving. Aided by the advances in the previously listed statistical packages, comprehensive and

elaborate tests of scale dimensionality, method effects, and variance partitioning have become apparent (cf. Bearden & Netemeyer, 1998). A more pronounced concern for content and face validity of items early in the development stage, as well as scale length considerations, are evident as well (Clark & Watson, 1995; Haynes et al., 1995). Although around for some time, generalizability theory (G-Theory) is now being more frequently applied in scale development and validation (Marcoulides, 1998; Shavelson & Webb, 1991). All these issues (i.e., dimensionality, face and content validity, scale length, and G-Theory) are addressed in this text.

LATENT CONSTRUCTS

As previously stated, latent constructs are not directly observable or quantifiable, and scores on measures of latent constructs may be variable. That is, the strength and magnitude for the scores may change over time. There is a seemingly endless array of latent constructs in the social sciences, ranging from those that are very broad, such as "extraversion" of the Big Five personality traits, to more narrow constructs that may be considered subcomponents of broader constructs, such as "talkativeness" as a subcomponent of extraversion (Clark & Watson, 1995). As such, latent constructs require a thoughtful elaboration regarding the level of abstraction and specificity.

Theory and Validity

The importance of theory in developing measures of latent constructs cannot be overstated. In their classic works on measurement and validity, Cronbach and Meehl (1955) and Loevinger (1957) eloquently stated the importance of theory in measurement. For measures of a latent construct to have relevance in the social sciences, the latent construct should be grounded in a theoretical framework. Even narrowly abstracted constructs based in theory are more useful as antecedents or consequences of other latent constructs or behaviors when embedded in theory. As such, a latent construct's relevance to the social sciences depends greatly on the theories in which it is couched. In other words, what does the latent construct of interest predict, and/or what predicts the latent construct? These relationships have been referred to as a latent construct's "nomological net" (Cronbach & Meehl, 1955).

Theory is concerned not only with the latent construct of interest but with the validity of the measurement of the construct as well. The two, theory and validity, are intertwined: The relevance of a latent construct largely depends on its "construct validity." Simply stated, construct validity is an assessment of the degree to which a measure actually measures the latent construct it is intended to measure. Cronbach and Meehl (1955) stated that demonstrating construct validity involves at least three steps: (a) specifying a set of theoretical constructs and their relations (a theory), (b) developing methods to measure the constructs of the theory, and (c) empirically testing how well manifest (observable) indicators (items) measure the constructs in the theory and testing the hypothesized relations among the constructs of theory as well (i.e., the nomological net). Furthermore, assessing construct validity is an ongoing process. One study supporting a construct's validity is not enough to conclude that the measure has been validated. Multiple tests and applications over time are required, and some of these may require a refinement of the construct itself, as well as its measure. As stated by Clark and Watson (1995), "The most precise and efficient measures are those with established construct validity; they are manifestations of constructs in an articulated theory that is well supported by empirical data" (p. 310).

Importance of the Literature Review

A well-grounded theory begins with conceptualizations based on a thorough review of the literature. Such literature reviews serve two important purposes. (We will elaborate upon the importance of the literature review later in this text, as it is considered a key issue in scale development. For now, only two broad points are mentioned.) First, a literature review should alert the researcher to previous attempts to conceptualize the construct of interest and theories in which the construct may prove useful as an independent or dependent variable. As such, a more precise conceptualization of the construct, its boundaries and content domain, and potential antecedents and consequences can be uncovered. A rigorous literature review also will indicate past attempts at measuring the construct and the strengths and weaknesses of such attempts.

Second, given that scale development and validation is a time-consuming and sometimes costly endeavor, a thorough literature review should help answer the following question: Is a scale needed at all? If good measures of a construct already exist, the value of a new measure may be small relative to the costs involved in development. A new measure should show some

theoretical or empirical advantage over existing measures of the same construct to be useful. In fact, some authors refer to this as an aspect of "incremental validity." Given the objectives of this text, for a new scale to have incremental validity over existing measures, it should either capture the targeted construct more accurately or be more efficient (e.g., shorter, cheaper, more user friendly, easier to respond to) than existing measures. In sum, a thorough literature review can help avoid the redundancy of developing another scale to assess an already well measured construct.

OVERVIEW OF DIMENSIONALITY, RELIABILITY, AND VALIDITY

Dimensionality, reliability, and validity are all interrelated measurement properties. What follows is a brief overview of these properties and how they are related. As emphasized above, the process of scale development starts with a thorough review of the literature in which a solid theoretical definition of the construct and its domain is delineated and outlined. This definition, and attendant description, should entail what is included in the domain of the construct, what is excluded from the construct's domain, and the a priori dimensionality of the construct's domain. The theoretical definition, the domain of the construct, and dimensionality should be derived from a thorough review of the existing literature and, ideally, expert opinion. In essence, the construct's definition and content domain determine theoretical dimensionality.

Dimensionality

A measure's dimensionality is concerned with the homogeneity of items. Basically, a measure that is considered unidimensional has statistical properties demonstrating that its items underlie a single construct or factor. When the measure is multidimensional, items tap more than one dimension or factor. A construct's domain can be hypothesized as unidimensional, as multidimensional, and/or as a higher-order factor. Thus, the scale (or subscales/factors) used to operationalize the construct should reflect the hypothesized dimensionality. Given that scale (factor) unidimensionality is considered prerequisite to reliability and validity, assessment of unidimensionality should be paramount (Cortina, 1993; Gerbing & Anderson, 1988; Hattie, 1985; Schmitt, 1996).

A number of procedures have been employed to check the dimensionality of a scale (e.g., item analysis and exploratory factor analysis). One somewhat agreed upon technique is confirmatory factor analysis, in which several multi-item factors (and relations among the factors) can be specified and evaluated on criteria used to assess dimensionality (e.g., fit indices, presence of correlated measurement errors, and degree of cross-loading) (Anderson & Gerbing, 1988; Clark & Watson, 1995; Floyd & Widaman, 1995; Gerbing & Anderson, 1988; Hattie, 1985; Kumar & Dillon, 1987). Chapter 2 of this text discusses the concept of dimensionality, and later chapters offer empirical examples.

Reliability

Reliability is concerned with that portion of measurement that is due to permanent effects that persist from sample to sample. There are two broad types of reliability referred to in the psychometric literature: (a) test-retest (temporal stability)—the correlation between the same person's score on the same set of items at two points in time; and (b) internal consistency—the interrelatedness among items or sets of items in the scale.

Test-retest reliability is concerned with the stability of a respondent's item responses over time. A test-retest or "stability" coefficient usually is estimated by the magnitude of the correlation between the same measures (and sample) on different assessment occasions. If the stability coefficient is low in magnitude, with no change in the construct over time, the reliability of the measure is in doubt. Thus, test-retest reliability is useful because it offers information regarding the degree of confidence one has that the measure reflects the construct and is generalizable to other assessment occasions (Haynes et al., 1999). Interestingly, test-retest reliability has not been assessed in scale use or development as frequently as internal consistency (Robinson et al., 1991). It is unfortunate that test-retest estimates are available for so few of the scales in the social sciences, and those planning scale development work should give stronger consideration to assessing test-retest reliability in addition to using other procedures of evaluating reliability and validity.

Internal consistency assesses item interrelatedness. Items composing a scale (or subscale) should show high levels of internal consistency. Some commonly used criteria for assessing internal consistency are individual corrected item-to-total correlations, the average interitem correlation among scale items, and a number of reliability coefficients (Churchill, 1979; Cortina, 1993;

DeVellis, 1991; Nunnally & Bernstein, 1994; Robinson et al., 1991). The most widely used internal consistency reliability coefficient is Cronbach's (1951) coefficient alpha. (Others are briefly discussed later in this text. For now, we limit our discussion to coefficient alpha.) Although a number of rules of thumb also exist concerning what constitutes an acceptable level of coefficient alpha, scale length must be considered. As the number of items increases, alpha will tend to increase. Because parsimony is also a concern in measurement (Clark & Watson, 1995; Cortina, 1993), an important question is "How many items does it take to measure a construct?" The answer to this question depends partially on the domain and dimensions of the construct. Naturally, a construct with a wide domain and multiple dimensions will require more items to adequately tap the domain/dimensions than a construct with a narrow domain and few dimensions. Given that most scales are self-administered and that respondent fatigue and/or noncooperation need to be considered, scale brevity is often advantageous (Churchill & Peter, 1984; Cortina, 1993; DeVellis, 1991; Nunnally & Bernstein, 1994).

With the advent of structural equation modeling, other tests of internal consistency or internal structure/stability became available. Composite reliability (construct reliability), which is similar to coefficient alpha, can be calculated directly from the LISREL, CALIS, EQS, or AMOS output (cf. Fornell & Larcker, 1981). A more stringent test of internal structure/stability involves assessing the amount of variance captured by a construct's measure in relation to the amount of variance due to measurement error—the average variance extracted (AVE). By using a combination of the criteria above (i.e., corrected item-to-total correlations, examining the average interitem correlation, coefficient alpha, composite reliability, and AVE), scales can be developed in an efficient manner without sacrificing internal consistency.

Construct Validity

Construct validity refers to how well a measure actually measures the construct it is intended to measure. Construct validity is the ultimate goal in the development of an assessment instrument and encompasses all evidence bearing on a measure (Haynes et al., 1999). Some disagreement exists as to the classification of and types of validity that fall under the rubric of construct validity. Still, many researchers believe that translation (content and face), convergent, discriminant, criterion-related (or predictive), nomological,

and known-group validity collectively represent the most frequently employed sources of construct validity. Given that all evidence bearing on a measure contributes to establishing construct validity, a measure must also a priori exhibit its theoretical dimensionality and show evidence of reliability to be considered valid. As such, dimensionality and reliability are necessary but insufficient conditions for construct validity. Again, a number of procedures exist for establishing construct validity. Although these are expanded upon throughout the remainder of the text, we offer a brief discussion of each validity type.

Translation validity is concerned with the content of items; two subtypes have been delineated—content and face validity. The term *content validity* has been defined in many ways, with most definitions stressing that a measure's items are a proper sample of the theoretical domain of the construct (Messick, 1993; Nunnally & Bernstein, 1994). Most definitions are consistent with the following: Content validity reflects "the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose" (Haynes et al., 1995, p. 238). "Elements" refer to the content of individual items, response formats, and instructions to respondents, and "representativeness" refers to the degree to which the elements are proportional to the facets (domains) of the targeted construct and to the degree that the entire domain of the targeted construct has been sampled. That is, the items should appear consistent with the theoretical domain of the construct in all respects, including response formats and instructions. In developing scales that are content valid, it is generally recommended that a number of items be generated that "tap the domain of the construct," that the items be screened by judges with expertise in the literature, and that pilot tests on samples from relevant populations be conducted to trim and refine the pool of items (DeVellis, 1991; Robinson et al., 1991). Haynes et al. (1995) offer an excellent description of procedures for establishing content validity that are elaborated upon in Chapters 3 through 7.

Face validity has been referred to as the "mere appearance that a measure has validity" (Kaplan & Saccuzzo, 1997, p. 132). Although the terms "face validity" and "content validity" have been used interchangeably, some argue that face validity should be separate from content validity (Anastasi & Urbina, 1998; Nevo, 1985). Others go a bit further and delineate face from content validity in terms of researcher and respondent. A highly face valid instrument enhances its use in practical situations by (among other things) inducing cooperation of respondents via ease of use, proper reading level, clarity,

easily read instructions, and easy-to-use response formats. A somewhat accepted definition of face validity implies that an instrument or test, when used "in a practical situation should, in addition to having pragmatic or statistical validity, appear practical, pertinent and related to the purposes of the instrument [test] as well; i.e., it should not only *be* valid, but it should *also appear* valid to respondents" (Nevo, 1985, p. 287). Thus, face validity may be more concerned with what respondents from relevant populations infer with respect to what is being measured.

Convergent validity refers to the degree to which two measures designed to measure the same construct are related. Convergence is found if the two different measures of the same construct are highly correlated. *Discriminant* validity assesses the degree to which two measures designed to measure similar, but conceptually different, constructs are related. A low to moderate correlation is often considered evidence of discriminant validity. Multitrait-multimethod matrices (MTMM) have often been used to assess convergent and discriminant validity where maximally different measurement methods (i.e., self-report vs. observational) are required (Campbell & Fiske, 1959; Churchill, 1979; Peter, 1979, 1981). Later in this text, we offer an MTMM example and a structural equation modeling approach to examine discriminant validity (e.g., Anderson & Gerbing, 1988; Fornell & Larcker, 1981).

Nomological validity has been defined as the degree to which predictions from a formal theoretical network containing the concept under scrutiny are confirmed (Campbell, 1960). It assesses the degree to which constructs that are theoretically related are empirically related (i.e., their measures correlate significantly in the predicted direction). Guidelines for establishing nomological validity also exist but have been criticized as well (Peter, 1981). As with internal consistency and convergent and discriminant validation, structural equation packages recently have been used to assess nomological validity of scale measures. Several books (e.g., Bollen, 1989; Byrne, 2001; Hayduk, 1996; Hoyle, 1995; Schumacker & Lomax, 1996) and articles (e.g., Anderson & Gerbing, 1988; Bagozzi, Yi, & Phillips, 1991; Bentler & Chou, 1987) illustrate modeling techniques, evaluative criteria, and guidelines for what constitutes nomological validity.

Definitions of *criterion-related validity* vary, and some definitions are similar to the definitions of other validity types. For example, criterion-related validity has been referred to as the degree to which a measure covaries with previously validated or "gold-standard" measures of the same constructs (Haynes et al., 1999). This definition is similar to that of convergent validity.

Criterion-related validity also has been referred to as the extent to which a measure corresponds to another measure of interest (Kaplan & Saccuzzo, 1997). Some contend that criterion-related validity is the same as *predictive validity*—the functional form or relation between a predictor and a criterion before, during, or after a predictor is applied (Nunnally & Bernstein, 1994). Such an approach is based on the temporal relation of the predictor and its criterion, that is, "post-dictive," "concurrent," and "predictive" validity. What most definitions have in common is that criterion-related validity is assessed by a theoretically specified pattern of relations between a measure and a criterion often referred to as a validity coefficient. Chapter 4 further explores criterion-related validity.

Known-group validity involves the measure's ability to distinguish reliably between groups of people that should score high on the trait and low on the trait. As examples, a person who is truly conservative should score significantly higher on a conservatism scale than a person who is liberal, and salespeople in the retail car business and large computer business should differ in their levels of customer orientation (Saxe & Weitz, 1982). Thus, mean score differences between groups for a given construct can be used as evidence of known-group validity. An excellent application of known-group validity testing can be found in Jarvis and Petty (1996).

OVERVIEW OF RECOMMENDED PROCEDURES AND STEPS IN SCALE DEVELOPMENT

As is clearly evidenced from the preceding pages, numerous articles and books advocate "how" to develop a scale (e.g., Churchill, 1979; Clark & Watson, 1995; DeVellis, 1991; Haynes et al., 1999; Nunnally & Bernstein, 1994; Spector, 1992). Steps and procedures vary from author to author based on the goals and purposes of the measurement. Still, most writings do share a common set of guidelines for scale development. Given our focus, the steps and procedures used to guide this text are based on scaling self-report paper-and-pencil measures of latent social-psychological constructs. Figure 1.1 offers a diagram of the steps we recommend in scale development. Each of these steps is elaborated upon in upcoming chapters. For now, we offer a brief overview of what each step entails.

Step 1: Construct Definition and Content Domain
Issues to Consider:
(a) The importance of clear construct definition, content domain, and the role of theory
(b) The focus on "effect" items/indicators vs. "formative" items/indicators
(c) Construct dimensionality: unidimensional, multidimensional, or a higher-order construct?

Step 2: Generating and Judging Measurement Items
Issues to Consider:
(a) Theoretical assumptions about items (e.g., domain sampling)
(b) Generating potential items and determining the response format
  (1) How many items as an initial pool
  (2) Dichotomous vs. multichotomous response formats
  (3) Item wording issues
(c) The focus on "content" validity in relation to theoretical dimensionality
(d) Item judging (expert and layperson)—the focus on "content" and "face" validity

Step 3: Designing and Conducting Studies to Develop and Refine the Scale
Issues to Consider:
(a) Pilot testing as an item-trimming procedure
(b) The use of several samples from relevant populations for scale development
(c) Designing the studies to test psychometric properties
(d) Initial item analyses via exploratory factor analyses (EFAs)
(e) Initial item analyses and internal consistency estimates
(f) Initial estimates of validity
(g) Retaining items for the next set of studies

Step 4: Finalizing the Scale
Issues to Consider:
(a) The importance of several samples from relevant populations
(b) Designing the studies to test the various types of validity
(c) Item analyses via EFA
  (1) The importance of EFA consistency from Step 3 to Step 4
  (2) Deriving an initial factor structure—dimensionality and theory
(d) Item analyses and confirmatory factor analyses (CFAs)
  (1) Testing the theoretical factor structure and model specification
  (2) Evaluating CFA measurement models
  (3) Factor model invariance across studies (i.e., multiple-group analyses)
(e) Additional item analyses via internal consistency estimates
(f) Additional estimates of validity
(g) Establishing norms across studies
(h Applying G-Theory

Figure 1.1 Steps in Scale Development

Step 1: Construct Definition and Content Domain

As we have stated throughout this introductory chapter, the importance of theory in scale development cannot be overstated, and developing and refining a theory requires a thorough literature review. During the literature review and theory development processes, several issues should be stressed: (a) the importance of clear construct definition, content domain, and the role of theory; (b) a focus on "effect" or "reflective" items rather than "formative" indicators; and (c) construct dimensionality—unidimensional, multidimensional, or a higher-order construct.

Step 2: Generating and Judging Measurement Items

This second step involves generating and judging a pool of items from which the scale will be derived. Several issues must be considered, including the following: (a) theoretical assumptions about items (e.g., domain sampling), (b) generating potential items and determining the response format (i.e., how many items as an initial pool, dichotomous vs. multichotomous response formats, and item wording issues), (c) the focus on "content" validity and its relation to theoretical dimensionality, and (d) item judging (both expert and layperson)—the focus on "content" and "face" validity.

Step 3: Designing and Conducting Studies to Develop and Refine the Scale

Once a suitable pool of items has been generated and judged, empirical testing of the items on relevant samples is the next step. Issues and procedures to be considered include (a) pilot testing as an item-trimming procedure, (b) the use of several samples from relevant populations for scale development, (c) designing studies to test psychometric properties, (d) initial item analyses via exploratory factor analyses (EFAs), (e) initial item analyses and internal consistency estimates, (f) initial estimates of validity, and (g) retaining items for the next set of studies.

Step 4: Finalizing the Scale

Several studies should be used to help finalize the scale. Many of the procedures used and issues involved in refining the scale will also be applicable

to deriving the final form of the scale. These include (a) the importance of several samples from relevant populations, (b) designing the studies to test the various types of validity, (c) item analyses via EFA with a focus on the consistency of EFA results across samples from Step 3 to Step 4 in testing an initial factor structure, (d) item analyses and confirmatory factor analyses (CFAs), (e) additional item analyses via internal consistency estimates, (f) additional estimates of validity, (g) establishing norms across studies, and (h) given that numerous studies have been done across various settings, applying generalizability theory to the final form of the scale.

SUMMARY AND PREVIEW OF THE TEXT

In this opening chapter, we have tried to provide the reader with an overview of the purpose of our text. To reiterate, our purpose is to focus on measuring latent perceptual social-psychological constructs via paper-and-pencil self-reports. For a construct to be valuable, it must have theoretical and/or practical relevance to the social scientist. Thus, a careful consideration must be made of what the construct of interest predicts and/or what predicts the construct of interest. Here, the notion of theory and "knowing" the literature is all-important. Furthermore, given the importance of measurement in the social sciences, any measure must be valid to allow for constructing confident inferences from empirical studies. Such validity rests on how well the latent construct being measured is based in theory.

Also in this opening chapter, we have overviewed the concepts of dimensionality, reliability, and validity, as well as summarized a series of steps for deriving measures with adequate psychometric properties. The remainder of our text elaborates on dimensionality, reliability, and validity, and the four steps in scale construction. Specifically, Chapter 2 discusses dimensionality, its relation to reliability and validity, and procedures for establishing dimensionality. Chapter 3 discusses reliability, its relation to validity, and procedures for establishing reliability, including G-Theory. Chapter 4 discusses validity and procedures for providing evidence of validity. Chapters 5, 6, and 7 provide detailed examples of the four steps in scale development, and Chapter 8 offers concluding remarks, with a focus on the need to constantly reevaluate constructs, their measures, and the validity of the measures.