1

Overview

Measurement is of vital concern across a broad range of social research contexts. For example, consider the following hypothetical situations:

1. A health psychologist faces a common dilemma: The measurement scale she needs apparently does not exist. Her study requires that she have a measure that can differentiate between what individuals *want* to happen and what they *expect* to happen when they see a physician. Her research shows that previous studies used scales that inadvertently confounded these two ideas. No existing scales appear to make this distinction in precisely the way that she would like. Although she could fabricate a few questions that seem to tap the distinction between what one wants and expects, she worries that "made-up" items might not be reliable or valid indicators of these concepts.
2. An epidemiologist is unsure how to proceed. He is performing secondary analyses on a large data set based on a national health survey. He would like to examine the relationship between certain aspects of perceived psychological stress and health status. Although no set of items intended as a stress measure was included in the original survey, several items originally intended to measure other variables appear to tap content related to stress. It might be possible to pool these items into a reliable and valid measure of psychological stress. However, if the pooled items constitute a poor measure of stress, the investigator might reach erroneous conclusions.
3. A marketing team is frustrated in its attempts to plan a campaign for a new line of high-priced infant toys. Focus groups have suggested that parents' purchasing decisions are strongly influenced by the

apparent educational relevance of toys of this sort. The team suspects that parents who have high educational and career aspirations for their infants will be more attracted to this new line of toys. Therefore, the team would like to assess these aspirations among a large and geographically dispersed sample of parents. Additional focus groups are judged to be too cumbersome for reaching a sufficiently large sample of consumers.

In each of these situations, people interested in some substantive area have come head to head with a measurement problem. None of these researchers is interested primarily in measurement per se. However, each must find a way to quantify a particular phenomenon before tackling the main research objective. In each case, "off-the-shelf" measurement tools are either inappropriate or unavailable. All the researchers recognize that adopting haphazard measurement approaches runs the risk of yielding inaccurate data. Developing their own measurement instruments seems to be the only remaining option.

Many behavioral and social science researchers have encountered similar problems. One all-too-common response to these types of problems is reliance on existing instruments of questionable suitability. Another is to assume that newly developed questionnaire items that "look right" will do an adequate measurement job. Uneasiness or unfamiliarity with methods for developing reliable and valid instruments and the inaccessibility of practical information on this topic are common excuses for weak measurement strategies. Attempts at acquiring scale development skills may lead a researcher either to arcane sources intended primarily for measurement specialists or to information too general to be useful. This volume is intended as an alternative to those choices.

General Perspectives on Measurement

Measurement is a fundamental activity of science. We acquire knowledge about people, objects, events, and processes by observing them. Making sense of these observations frequently requires that we quantify them (i.e., that we measure the things in which we have a scientific interest). The process of measurement and the broader scientific questions it serves interact with each other; the boundaries between them are often imperceptible. This happens, for example, when a new entity is detected or refined in the course of measurement or when the reasoning involved in determining how to quantify a phenomenon of interest sheds new light on the phenomenon itself. For example, Smith et al. (1995) investigated women's perceptions of battering. An a priori conceptual model based on theoretical analysis suggested six distinct components to these perceptions. Empirical work aimed at developing a scale to measure these perceptions indicated that, among both battered and nonbattered women, a much simpler conceptualization prevailed: A single concept thoroughly explained how study participants responded to 37 of 40 items administered. This finding suggests that what researchers saw as a complex constellation of variables

was actually perceived by women living in the community as a single, broader phenomenon. Thus, in the course of devising a means of measuring women's perceptions about battering, we discovered something new about the structure of those perceptions.

Duncan (1984) argues that the roots of measurement lie in social processes and that these processes and their measurement actually precede science: "All measurement . . . is social measurement. Physical measures are made for social purposes" (p. 35). In reference to the earliest formal social measurement processes, such as voting, census taking, and systems of job advancement, Duncan notes that "their origins seem to represent attempts to meet everyday human needs, not merely experiments undertaken to satisfy scientific curiosity." He goes on to say that similar processes

> can be drawn in the history of physics: the measurement of length or distance, area, volume, weight, and time was achieved by ancient peoples in the course of solving practical, social problems; and physical science was built on the foundations of those achievements. (p. 106)

Whatever the initial motives, each area of science develops its own set of measurement procedures. Physics, for example, has developed specialized methods and equipment for detecting subatomic particles. Within the behavioral/social sciences, *psychometrics* has evolved as the subspecialty concerned with measuring psychological and social phenomena. Typically, the measurement procedure used is the questionnaire and the variables of interest are part of a broader theoretical framework.

Historical Origins of Measurement in Social Science

Early Examples

Common sense and the historical record support Duncan's claim that social necessity led to the development of measurement before science emerged. No doubt, some form of measurement has been a part of our species' repertoire since prehistoric times. The earliest humans must have evaluated objects, possessions, and opponents on the basis of characteristics such as size. Duncan (1984) cites biblical references to concerns with measurement (e.g., "A false balance is an abomination to the Lord, but a just weight is a delight," Proverbs 11:1) and notes that the writings of Aristotle refer to officials charged with checking weights and measures. Anastasi (1968) notes that the Socratic method employed in ancient Greece involved probing for understanding in a manner that might be regarded as knowledge testing. In his 1964 essay, P. H. DuBois (reprinted in Barnette, 1976) describes the use of civil service testing as early as 2200 BCE in China. Wright (1999) cites other examples of the importance ascribed in antiquity to accurate measurement, including the "weight of seven"

on which seventh-century Muslim taxation was based. He also notes that some have linked the French Revolution, in part, to peasants being fed up with unfair measurement practices.

The notion that measurement can entail error and that certain steps might be taken to reduce that error is a more recent insight. Buchwald (2006), in his review of measurement discrepancies and their impact on knowledge, notes that, while still in his twenties during the late 1660s and early 1670s, Isaac Newton was apparently the first to use an average of multiple observations. His intent was to produce a more accurate measurement when his observations of astronomical phenomena yielded discrepant values. Interestingly, he did not document the use of averages in his initial reports but concealed his reliance on them for decades. This concealment may have stemmed less from a lack of integrity than from a limited understanding of error and its role in measurement. Commenting on another astronomer's similar disdain for discrepant observations, Alder (2002) argues that even in the late 1700s, concealment of discrepancies in observation "were not only common, they were considered a savant's prerogative. It was an error that was seen as a moral failing" (p. 301). Buchwald (2006) makes a similar observation:

> [17th- and early 18th-century scientists'] way of working regarded differences not as the inevitable byproducts of the measuring process itself, but as evidence of failed or inadequate skill. Error in measurement was potentially little different from faulty behavior of any kind: it could have moral consequences, and it had to be managed in appropriate ways. (p. 566)

Astronomers were not the only scientists making systematic observations of natural phenomena in the late 1600s and early 1700s. In the 1660s, John Graunt was compiling birth and death rates from christening and burial records in Hampshire, England. Graunt used an averaging procedure (though not the one in common use today) to summarize his findings. According to Buchwald (2006), Graunt's motivation for this averaging was to capture an ephemeral "true" value. The notion was that the ratio of births to deaths obeyed some law of nature but that unpredictable events that might occur in any given year would mask that fundamental truth. This view of observation as an imperfect window into nature's truths suggests a growing sophistication in how the measurement was viewed: In addition to the observer's limitations, other factors could also corrupt empirically gathered information, and some adjustments of those values might more accurately reveal the true nature of the phenomenon of interest.

Despite these early insights, it was a century after Newton's first use of the average before scientists more widely recognized that all measurements were prone to error and that an average would minimize such error (Buchwald, 2006). According to physicist and author Leonard Mlodinow (2008), in the late 18th and early 19th centuries, developments in astronomy and physics forced

scientists to approach random error more systematically, which led to the emergence of mathematical statistics. By 1777, Daniel Bernoulli (nephew of the more famous Jakob Bernoulli) compared the distributions of values obtained from astronomical observations to the path of an archer's arrows, clumping around a central point with progressively fewer at increasingly greater distances from that center. Although the theoretical treatment that accompanied that observation was wrong in certain respects, it marks the beginning of a formal analysis of error in measurement (Mlodinow, 2008). Buchwald (2006) argues that a fundamental shortcoming of 18th-century interpretations of measurement error was a failure to distinguish between random and systematic error. Not until the dawning of the next century would a more incisive understanding of randomness emerge. With this growing understanding of randomness came advances in measurement; and, as measurement advanced, so did science.

Emergence of Statistical Methods and the Role of Mental Testing

Nunnally's (1978) perspective supports the view that a more sophisticated understanding of randomness, probability, and statistics, was necessary for measurement to flourish. He argues that, although systematic observations may have been going on, the absence of more formal statistical methods hindered the development of a science of measuring human abilities until the latter half of the 19th century. The eventual development of suitable statistical methods in the 19th century was set in motion by Darwin's work on evolution and his observation and measurement of systematic variation across species. Darwin's cousin, Sir Francis Galton, extended the systematic observation of differences to humans. A chief concern of Galton was the inheritance of anatomical and intellectual traits. Karl Pearson, regarded by many as the "founder of statistics" (e.g., Allen & Yen, 1979, p. 3), was a junior colleague of Galton's. Pearson developed the mathematical tools—including the Product-Moment Correlation Coefficient bearing his name—needed to systematically examine relationships among variables. Scientists could then quantify the extent to which measurable characteristics were interrelated. Charles Spearman continued in the tradition of his predecessors and set the stage for the subsequent development and popularization of factor analysis in the early 20th century. It is noteworthy that many of the early contributors to formal measurement (including Alfred Binet, who developed tests of mental ability in France in the early 1900s) shared an interest in intellectual abilities. Hence, much of the early work in psychometrics was applied to "mental testing."

The Role of Psychophysics

Another historical root of modern psychometrics arose from psychophysics. As we have seen, measurement problems were common in astronomy and other physical sciences and were a source of concern for Sir Isaac Newton (Buchwald, 2006). Psychophysics exists at the juncture of psychology and physics and

concerns the linkages between the physical properties of stimuli and how they are perceived by humans. Attempts to apply the measurement procedures of physics to the study of sensations led to a protracted debate regarding the nature of measurement. Narens and Luce (1986) have summarized the issues. They note that in the late 19th century, Helmholtz observed that physical attributes, such as length and mass, possessed the same intrinsic mathematical structure as did positive real numbers. For example, units of length or mass could be ordered and added as could ordinary numbers. In the early 1900s, the debate continued. The Commission of the British Association for the Advancement of Science regarded fundamental measurement of psychological variables to be impossible because of the problems inherent in ordering or adding sensory perceptions. S. S. Stevens argued that strict additivity, as would apply to length or mass, was not necessary and pointed out that individuals could make fairly consistent ratio judgments of sound intensity. For example, they could judge one sound to be twice or half as loud as another. He argued that this ratio property enabled the data from such measurements to be subjected to mathematical manipulation. Stevens is credited with classifying measurements into nominal, ordinal, interval, and ratio scales. Loudness judgments, he argued, conformed to a ratio scale (Duncan, 1984). At about the time that Stevens was presenting his arguments on the legitimacy of scaling psychophysical measures, L. L. Thurstone was developing the mathematical foundations of factor analysis (Nunnally, 1978). Thurstone's interests spanned both psychophysics and mental abilities. According to Duncan (1984), Stevens credited Thurstone with applying psychophysical methods to the scaling of social stimuli. Thus, his work represents a convergence of what had been separate historical roots.

Later Developments in Measurement

Evolution of Basic Concepts

As influential as Stevens has been, his conceptualization of measurement is by no means the final word. He defined measurement as the "assignment of numerals to objects or events according to rules" (Duncan, 1984). Duncan challenged this definition as

> incomplete in the same way that "playing the piano is striking the keys of the instrument according to some pattern" is incomplete. Measurement is not only the assignment of numerals, etc. It is also the assignment of numerals in such a way as to correspond to *different degrees of a quality* . . . or property of some object or event. (p. 126)

Narens and Luce (1986) also identified limitations in Stevens's original conceptualization of measurement and illustrated a number of subsequent refinements. However, their work underscores a basic point made by Stevens: Measurement models other than the type endorsed by the Commission (of the

British Association for the Advancement of Science) exist, and these lead to measurement methods applicable to the nonphysical as well as physical sciences. In essence, this work on the fundamental properties of measures has established the scientific legitimacy of the types of measurement procedures used in the social sciences.

Evolution of Mental Testing

Although, traditionally, mental testing (or ability testing, as it is now more commonly known) has been an active area of psychometrics, it is not a primary focus of this volume. Nonetheless, it bears mention as a source of significant contributions to measurement theory and methods. A landmark publication, *Statistical Theories of Mental Test Scores,* by Frederic M. Lord and Melvin R. Novick, first appeared in 1968 and has recently been reissued (Lord & Novick, 2008). This volume grew out of the rich intellectual activities of the Psychometric Research Group of the Educational Testing Service, where Lord and Novick were based. This impressive text summarized much of what was known in the area of ability testing at the time and was among the first cogent descriptions of what has become known as *item response theory*. The latter approach was especially well suited to an area as broad as mental testing. Many of the advances in that branch of psychometrics are less common and perhaps less easily applied when the goal is to measure characteristics other than mental abilities. Over time, the applicability of these methods to measurement contexts other than ability assessment has become more apparent, and we will discuss them in a later chapter. Primarily, however, I will emphasize the "classical" methods that largely have dominated the measurement of social and psychological phenomena other than abilities. These methods are generally more tractable for nonspecialists and can yield excellent results.

Assessment of Mental Illness

The evolution of descriptions of mental illness has a separate history that provides a useful case study in how the lack of a guiding measurement model can complicate assessment. Over the centuries, society's ability to recognize different types of mental illness has evolved from completely unsystematic observation toward efforts to understand relationships among symptoms, causes, and treatments that are compatible with more formal measurement. It has been a challenging journey.

Early Roman, Greek, and Egyptian writings equated what we now recognize as symptoms of mental illness with demonic possession or other supernatural circumstances (e.g., PBS, 2002). By 400 BCE, the Greek physician Hippocrates was trying to understand mental conditions as arising from the physiological processes that were the primary focus of his scholarly work (PBS, 2002). His efforts may have been among the earliest to think of the overt indicators of mental illness in terms of their latent causes. However, even at that stage and well beyond, mental illnesses were described phenomenologically; that is, the

manifestations associated with mental illness were merely catalogued descriptively rather than understood as endpoints in a sequence with one or more clear, underlying causes.

Fairly crude methods of categorization continued for more than a millennium. Tartakovsky (2011) has summarized how mental illness was categorized for U.S. Census purposes as early as the mid-1800s. In the 1840 census, a single category, "idiocy/insanity," indicated the presence of a mental problem. By 1880, the census classification scheme had expanded to the following categories: mania, melancholia, monomania, paresis, dementia, dipsomania, and epilepsy. These are essentially descriptions of abnormal states or behaviors (e.g., persistent sadness, excessive drinking, muscle weakness, or convulsions) rather than etiological classifications.

Early in the 1880s, German psychiatrist Emil Kraepelin began to differentiate more systematically among mental disorders. A student of Wilhelm Wundt, who is credited as the founder of experimental psychology, Kraepelin was also a physician (Eysenck, 1968). Thus he brought two different perspectives to his classifications of mental illness. In 1883, he published *Compendium der Psychiatric* (Kraepelin, 1883), a seminal text arguing for a more scientific classification of psychiatric illnesses and differentiating between dementia praecox and manic depressive psychosis. But, again, despite his efforts to invoke explanations for these illnesses, his early diagnostic categories primarily are summary descriptions of manifest symptoms that tend to co-occur rather than cogent etiological explanations (Decker, 2007). Although Kraepelin advanced the scientific approach to understanding mental illness, the tools at his disposal were primitive, and in the end, his nosological categories were still largely descriptive. Decker (2007) assesses his legacy as follows: "To sum up: by today's research standards, Kraepelin's record-keeping and deductions would raise questions about preconceived notions and observer bias. The scientific shortcomings can be seen in Kraepelin's own description of his methods. For all his brilliance in categorical formulations, his legacy is balanced on shaky empirical foundations" (p. 341).

In the mid-20th century, American psychiatry tried to impose greater order on the assessment of mental illness. By the time of the appearance of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*; American Psychiatric Association [APA], 1952), the prevailing categorization systems attempted to classify mental illnesses based on both their manifestations and their etiologies, as in the case of acute brain trauma or alcoholism. However, more subtle notions of etiologies for conditions not linked to an obvious exogenous cause were not yet well developed and psychodynamic causes were often assumed. The term applied to such conditions was *reactions*, presumably to psychic stressors of unspecified origins. Again, the categorizations primarily were descriptions of manifest symptoms. Although *DSM*'s system of classification represented clear progress beyond earlier systems, it still fell short of conforming to standards of modern measurement. Even four decades later, when *DSM-IV* (American Psychiatric Association, 2000) appeared, there was considerable

dissatisfaction with the classification system. Psychologist Paul Meehl (1999) noted that the problem was not necessarily with the use of categories (some hard and fast, belong or don't belong, categories probably did exist, he argued) but the absence of a clear rationale for assigning people to them. To quote Meehl (1999), "For that minority of DSM rubrics that do denote real taxonic entities, the procedure for identifying them and the criteria for applying them lack an adequate scientific basis" (p. 166).

The prelude to and eventual appearance of *DSM-V* in 2013 (American Psychiatric Association, 2013) created an opportunity for the reexamination of mental health classification. Some feel that the team working on the revision failed to capitalize fully on that opportunity. As noted, a feature of mental health classification historically is that it has sought to categorize rather than scale. That is, the goal has been to describe the presence or absence rather than the degree of a particular condition. Experience suggests that, even for conditions, such as schizophrenia, that Meehl (1999) was willing to recognize as "taxonic" (i.e., being discrete disorders either present or absent), there is a continuum of impairment rather than an all-or-none state. Yet a reliance on categorization rather than scaling persists. In many cases, this has involved arbitrary thresholds for signs and symptoms, such that crossing some imaginary line of severity constituted the presence of a condition whereas falling just short of that line did not. Also, classifications have been based almost exclusively on observations of manifest symptoms rather than assessments of key signifiers of the conditions, such as the presence of causal pathogens, a genetic marker, or an abnormal state of internal chemistry that may be a basis for assigning a physical diagnosis. When work began (outside of public view) on *DSM-V*, many hoped it would be a bolder revision than the earlier editions and would apply more modern assessment approaches. In 2005, after plans for a revised *DSM* (which would become *DSM-V*) were announced, the mental health scientific community began to voice its concerns. A special issue of the *Journal of Abnormal Psychology*, for example, focused on the importance and utility of a reconceptualization of psychopathology based on identifying fundamental dimensions, such as disordered thought, affect, and behavior, that gives rise to specific mental health problems (Kreuger et al., 2005). Kreuger et al. (2005) argued that this approach could address two fundamental empirical shortcomings of category-based classification systems: the wide prevalence of comorbidity (i.e., individual symptom clusters fitting multiple diagnoses) and the extreme heterogeneity within diagnoses (i.e., individuals assigned the same diagnosis sharing few or perhaps no symptoms). Researchers, theoreticians, and even philosophers (e.g., Aragona, 2009) pressed for a reconceptualization of the diagnosis of mental illness that was more in line with empirical work, such as modern measurement approaches. Despite these efforts, however, the American Psychiatric Association issued *DSM-V* in a form that retained the basic categorization system used in earlier editions. This prompted Thomas Insel, Director of the National Institute of Mental Health (NIMH), to issue a statement on his blog (Insel, 2013) saying that NIMH would

no longer structure its research efforts around *DSM* categories and was undertaking a 10-year effort, the Research Domain Criteria (RDoC) project, to reconceptualize mental illness. Insel (2013) characterized this effort by saying that "RDoC is a framework for collecting the data needed for a new nosology. But it is critical to realize that we cannot succeed if we use *DSM* categories as the 'gold standard.'" The following month Insel issued a joint press release with the then-president elect of the American Psychiatric Association, Jeffrey A. Lieberman. In that release, they observed the following:

> Today, the American Psychiatric Association's (APA) Diagnostic and Statistical Manual of Mental Disorders (DSM), along with the International Classification of Diseases (ICD) represents the best information currently available for clinical diagnosis of mental disorders. . . .
>
> Yet, what may be realistically feasible today for practitioners is no longer sufficient for researchers. Looking forward, laying the groundwork for a future diagnostic system that more directly reflects modern brain science will require openness to rethinking traditional categories. It is increasingly evident that mental illness will be best understood as disorders of brain structure and function that implicate specific domains of cognition, emotion, and behavior. This is the focus of the NIMH's Research Domain Criteria (RDoC) project. (Insel & Lieberman, 2013)

In October 2015, Insel resigned his post at NIMH (Insel, 2015) to accept a position at the Life Sciences division (subsequently renamed Verily) of Alphabet, the umbrella company formed as part of Google's structural reorganization. One of the factors Insel mentioned as influencing his decision was his hope of bringing a more organized approach to mental health classification. As he stated in an interview for *MIT Technology Review*, his move to Alphabet, in part, represented his "trying to figure out a better way to bring data analytics to psychiatry. The diagnostic system we have is entirely symptom based and fairly subjective" (Regalado, 2015). Many hope the work Insel does at Alphabet will promote modernization of psychiatric assessment to make it more compatible with modern measurement standards.

The argument in favor of a more evidence-based classification of mental illness continues. Insel himself cofounded a company whose mission includes a greater focus on measurement. One of their principles is that, "Measurement-based care is fundamental to improving mental health care outcomes" (Mindstrong, 2020).

Broadening the Domain of Psychometrics

Duncan (1984) notes that the impact of psychometrics in the social sciences has transcended its origins in the measurement of sensations and intellectual abilities. Psychometrics clearly has emerged as a methodological paradigm in its own right. Duncan supports this argument with three examples of the

impact of psychometrics: (1) the widespread use of psychometric definitions of reliability and validity, (2) the popularity of factor analysis in social science research, and (3) the adoption of psychometric methods for developing scales measuring an array of variables far broader than those with which psychometrics was initially concerned (p. 203). Although Duncan made those assertions almost 40 years ago, they still apply today. The applicability of psychometric concepts and methods to the measurement of diverse psychological and social phenomena will occupy our attention for the remainder of this volume.

The Role of Measurement in the Social Sciences

The Relationship of Theory to Measurement

The phenomena we try to measure in social science research often derive from theory. Consequently, theory plays a key role in how we conceptualize our measurement problems. In fact, Lord and Novick (2008) ascribe theoretical issues an important role in the development of measurement theory. Theoreticians were concerned that estimates of relationships between constructs of interest were generally obtained by correlating *indicators* of those constructs. Because those indicators contained error, the resultant correlations were an underestimate of the actual relationship between the constructs. This motivated the development of methods of adjusting correlations for error-induced attenuation and stimulated the development of measurement theory as a distinct area of concentration (p. 69).

Of course, many areas of science measure things derived from theory. Until a subatomic particle is confirmed through measurement, it too is merely a theoretical construct. However, theory in psychology and other social sciences is different from theory in the physical sciences. Social scientists tend to rely on numerous theoretical models that concern rather narrowly circumscribed phenomena, whereas theories in the physical sciences are fewer in number and more comprehensive in scope. Festinger's (1954) social comparison theory, for example, focuses on a rather narrow range of human experience: the way people evaluate their own abilities or opinions by comparing themselves with others. In contrast, physicists continue to work toward a grand unified field theory that will embrace all the fundamental forces of nature within a single conceptual framework. Also, the social sciences are less mature than the physical sciences, and their theories are evolving more rapidly. Measuring elusive, intangible phenomena derived from multiple, evolving theories poses a clear challenge to social science researchers. Therefore, it is especially important to be mindful of measurement procedures and to fully recognize their strengths and shortcomings.

The more researchers know about the phenomena in which they are interested, the abstract relationships that exist among hypothetical constructs, and the quantitative tools available to them, the better equipped they are to develop

reliable, valid, and usable scales. Detailed knowledge of the specific phenomenon of interest is probably the most important of these considerations. For example, social comparison theory has many aspects that may imply different measurement strategies. One research question might require operationalizing social comparisons as relative preference for information about higher- or lower-status others, while another might dictate ratings of self relative to the "typical person" on various dimensions. Different measures capturing distinct aspects of the same general phenomenon (e.g., social comparison) thus may not yield convergent results (DeVellis et al., 1990). In essence, the measures are assessing different variables despite the use of a common variable name in their descriptions. Consequently, developing a measure that is optimally suited to the research question requires understanding the subtleties of the theory.

Different variables call for different assessment strategies. Number of tokens taken from a container, for example, can be observed directly. Many—arguably, most—of the variables of interest to social and behavioral scientists are not directly observable; beliefs, motivational states, expectancies, needs, emotions, and social role perceptions are but a few examples. Certain variables cannot be directly observed but can be determined by research procedures other than questionnaires. For example, although cognitive researchers cannot directly observe how individuals organize information about ethnicity into their self schemas, they may be able to use recall procedures to make inferences about how individuals structure their thoughts about self and ethnicity. There are many instances, however, in which it is impossible or impractical to assess social science variables with any method other than a self-administered measurement scale. This is often but not always the case when we are interested in measuring theoretical constructs. Thus, an investigator interested in measuring empathy may find it far easier to do so by means of a carefully developed questionnaire than by some alternative procedure.

Theoretical and Atheoretical Measures

At this point, we should acknowledge that although this book focuses on measures of theoretical constructs, not all self-report assessments need be theoretical. Education and age, for example, can be ascertained from self-report by means of a questionnaire. Depending on the research question, these two variables can be components of a theoretical model or simply part of a description of a study's participants. Some contexts in which people are asked to respond to a list of questions using a self-report format, such as an assessment of hospital patient meal preferences, have no theoretical foundation. In other cases, a study may begin atheoretically but result in the formulation of theory. For example, a market researcher might ask parents to list the types of toys they have bought for their children. Subsequently, the researcher might explore these listings for patterns of relationships. Based on the observed patterns of toy purchases, the researcher may develop a model of purchasing behavior. Public opinion questionnaires are another example of relatively atheoretical measurement. Asking people which brand of soap they use or for whom they

intend to vote seldom involves any attempt to tap an underlying theoretical construct. Rather, the interest is in the subject's response per se, not in some characteristic of the person it is presumed to reflect.

Distinguishing between theoretical and atheoretical measurement situations can be difficult at times. For example, seeking a voter's preference in presidential candidates as a means of predicting the outcome of an election amounts to asking a respondent to report his or her behavioral intention. An investigator may ask people how they plan to vote not out of an interest in voter decision-making processes but merely to anticipate the eventual election results. If, on the other hand, the same question is asked in the context of examining how attitudes toward specific issues affect candidate preference, a well-elaborated theory may underlie the research. The information about voting is not intended in this case to reveal how the respondent will vote but to shed light on individual characteristics. In these two instances, the relevance or irrelevance of the measure to theory is a matter of the investigator's intent, not the procedures used. Readers interested in learning more about constructing survey questionnaires that are not primarily concerned with measuring hypothetical constructs are referred to Converse and Presser (1986); Czaja and Blair (1996); Dillman (2007); Fink (1995); Fowler (2009); and Weisberg, Krosnick, and Bowen (1996).

Composite Measurement Tools

Measurement instruments that are collections of items combined into a composite score and intended to reveal levels of theoretical variables not readily observable by direct means are often referred to as *composite* or *aggregate measurement tools*. In this book, we further subdivide aggregate measurement tools into two classes, scales and indices. We typically develop composite tools when we want to measure phenomena that we believe to exist because of our theoretical understanding of the world but that we cannot assess directly. For example, we may invoke depression or anxiety as explanations for behaviors we observe. Most theoreticians would agree that depression or anxiety is not equivalent to the behavior we see but underlies it. Our theories suggest that these phenomena exist and that they influence behavior but that they are intangible. Sometimes, it may be appropriate to infer their existence from their behavioral consequences. However, at other times, we may not have access to behavioral information (as when we are restricted to mail survey methodologies), may not be sure how to interpret available samples of behavior (as when a person remains passive in the face of an event that most others would react to strongly), or may be unwilling to assume that behavior is isomorphic with the underlying construct of interest (as when we suspect that crying is the result of joy rather than sadness). In instances when we cannot rely on behavior as an indication of a phenomenon, it may be more useful to assess the construct by means of a carefully constructed and validated scale.

Even among theoretically derived variables, there is an implicit continuum ranging from relatively concrete and accessible to relatively abstract and

inaccessible phenomena. Not all will require multi-item scales. Age and education certainly have relevance to many theories but rarely require a multi-item scale for accurate assessment. People know their age and level of education. These variables, for the most part, are linked to concrete, relatively unambiguous events (e.g., date of birth and years of schooling, respectively). Unless some special circumstance, such as a neurological impairment is present, respondents can retrieve information about their age and education from memory quite easily. They can respond with a high degree of accuracy to a single question assessing a variable such as these. Ethnicity arguably is more complex and abstract than age or education. It typically involves a combination of physical, cultural, and historical factors. As a result, it is less tangible—more of a social construction—than age or education. Although the mechanisms involved in defining one's ethnicity may be complex and unfold over an extended period of time, most individuals have arrived at a personal definition and can report their ethnicity with little reflection or introspection. Thus, a single variable may suffice for assessing ethnicity under most circumstances. (This may change, however, as our society becomes progressively more multiethnic and as individuals define their personal ethnicity in terms of multiple ethnic groups reflecting their ancestry. A similar change has taken place with respect to gender identity, with a wider array of self-definitions than the traditional male-female distinction now in wider use.) Many other theoretical variables, however, require a respondent to reconstruct, interpret, judge, compare, or evaluate less accessible information. For example, measuring how married people believe their lives would be different if they had chosen a different spouse probably would require substantial mental effort, and one item may not capture the complexity of the phenomenon of interest. Under conditions such as these, using an aggregate measurement tool may be a more appropriate assessment strategy. Multiple items may capture the essence of such a variable with a degree of precision that a single item could not attain. It is precisely this type of variable—one that is not directly observable and that involves thought on the part of the respondent—that is most appropriately assessed by means of some form of an aggregate measurement tool.

It is important to differentiate among types of multi-item measures that yield a composite score. The distinctions among these different types of aggregate measures are of both theoretical and practical importance, as later chapters will reveal. The two principal types on which we will focus are a *scale* and an *index*. As the terms are used in this volume, a scale consists of what Bollen (1989, pp. 64–65; see also Loehlin, 1998, pp. 200–202) refers to as "effect indicators"—that is, items whose values are caused by an underlying construct (or *latent variable*, as we shall refer to it in the next chapter). A measure of depression often conforms to the characteristics of a scale, with the responses to individual items sharing a common cause—namely, the affective state of the respondent. Thus, how someone responds to items such as "I feel sad" and "My life is joyless" probably is largely determined by that person's feelings at the time. I will use the term *index*, on the other hand, to describe sets

of items that are cause indicators—that is, items that determine the level of a construct. A measure of presidential candidate electability, for example, might fit the characteristics of an index. The items might assess a candidate's public speaking effectiveness, record of military service, physical attractiveness, ability to inspire campaign workers, and potential financial resources. Although these characteristics probably do not share any common cause, they might all share an effect—increasing the likelihood of a successful presidential campaign. The items are not the result of any one thing, but they determine the same outcome. A more general term for a collection of items that one might aggregate into a composite score is *emergent variable* (e.g., Cohen, Cohen, Teresi, Marchi, & Velez, 1990), which includes collections of entities that share certain characteristics and can be grouped under a common category heading. Grouping them together, however, does not necessarily imply any causal linkage. Sentences beginning with a word having fewer than five letters, for example, can easily be categorized together although they share neither a common cause nor a common effect. An emergent variable "pops up" merely because someone or something (such as a data analytic program) perceives some type of similarity among the items in question. In Chapter 7, we will discuss differences between scales and indices and consider the latter in greater detail. Most of our discussion in earlier chapters, however, will focus on scales.

All Scales Are Not Created Equal

Regrettably, not all item composites are developed carefully. For many, *assembly* may be a more appropriate term than *development*. Researchers often throw together or dredge up items and assume they constitute a suitable scale. These researchers may give no thought to whether the items share a common cause (thus constituting a scale), share a common consequence (thus constituting an index), or merely are examples of a shared superordinate category that does not imply either a common causal antecedent or consequence (thus constituting an emergent variable).

A researcher not only may fail to exploit theory in developing a scale but also may reach erroneous conclusions about a theory by misinterpreting what a scale measures. An unfortunate but distressingly common occurrence is the conclusion that some *construct* is unimportant or that some *theory* is inconsistent based on the performance of a *measure* that may not reflect the variable assumed by the investigator. Why might this happen? Rarely in research do we directly examine relationships among variables. As noted earlier, many interesting variables are not directly observable, a fact we can easily forget. More often, we assess relationships among proxies (such as scales) that are intended to represent the variables of interest. The observable proxy and the unobservable variable may become confused. For example, variables such as blood pressure and body temperature, at first consideration, appear to be directly observable, but what we actually observe are proxies, such as a column of mercury or a digital readout. Our conclusions about the variables assume

that the observable proxies are closely linked to the underlying variables they are intended to represent. Such is the case for a thermometer; we may describe the level of mercury in a thermometer as "the temperature," even though, strictly speaking, it is merely a visible manifestation of temperature (i.e., thermal energy). In this case, where the two closely correspond, the consequences of referring to the measurement (scale value that the mercury attains) as the variable (amount of thermal energy) are nearly always inconsequential. When the relationship between the variable and its indicator is weaker than in the thermometer example, confusing the measure with the phenomenon it is intended to reveal can lead to erroneous conclusions. Consider a hypothetical situation in which an investigator wishes to perform a secondary analysis on an existing data set. Let us assume that our investigator is interested in the role of social support on subsequent professional attainment. The investigator observes that the available data set contains a wealth of information on subjects' professional statuses over an extended period of time and that subjects were asked whether they were married. In fact, there may be several items, collected at various times, that pertain to marriage. Let us further assume that, in the absence of any data providing a more detailed assessment of social support, the investigator decides to sum these marriage items into a "scale" and to use this as a measure of support. Most social scientists would agree that equating social support with marital status is not justified. The latter both omits important aspects of social support (e.g., the perceived quality of support received) and includes potentially irrelevant factors (e.g., status as a child too young to have married versus an adult of an age suitable for marriage at the time of measurement). If this hypothetical investigator concluded, on the basis of this assessment method, that social support played no role in professional attainment, that conclusion might be completely wrong. In fact, the comparison was between marital status and professional attainment (or more precisely, indicators of these variables). Only if marriage actually indicated level of support would the conclusion about support and professional attainment be valid.

Costs of Poor Measurement

Even if a poor measure is the only one available, the costs of using it may be greater than any benefits attained. Situations are rare in the social sciences in which an immediate decision must be made in order to avoid dire consequences and one has no other choice but to make do with the best instruments available. Even in these rare instances, however, the inherent problems of using poor measures to assess constructs do not vanish. Using a measure that does not assess what one presumes can lead to wrong decisions. Does this mean that we should use only measurement tools that have undergone rigorous development and extensive validation testing? Although imperfect measurement may be better than no measurement at all in some situations, we should *recognize* when our measurement procedures are flawed and temper our conclusions accordingly.

Often, an investigator will consider measurement as secondary to more important scientific issues that motivate a study and, thus, the researcher will attempt to economize by skimping on measurement. However, adequate measures are a necessary condition for valid research. Investigators should strive for an isomorphism between the theoretical constructs in which they have an interest and the methods of measurement they use to operationalize them. Poor measurement imposes an absolute limit on the validity of the conclusions one can reach. For an investigator who prefers to pay as little attention to measurement and as much to substantive issues as possible, an appropriate strategy might be to get the measurement part of the investigation correct from the very beginning so that it can be taken more or less for granted thereafter.

A researcher also can falsely economize by using instruments that are too brief in the hope of reducing the burden on respondents. Although several systematic reviews have shown that longer questionnaire length tends to be associated with somewhat lower response rates, this association is modest overall and absent in some studies (Rolstad et al., 2011; Edwards et al., 2002; Sitzia & Wood, 1998). Respondents' willingness to complete longer instruments may also be heavily influenced by the study's context and their level of interest in the content. When surveyed about a topic of high personal relevance (e.g., personal health status or experience with illness), respondents may tolerate or even prefer longer measures that allow them to better convey their perspective (Rolstad et al., 2011; Sitzia & Wood, 1998). Furthermore, choosing a questionnaire that is too brief to be reliable is a bad idea no matter how much respondents prefer its brevity. A reliable questionnaire that is completed by half of the respondents yields more information than an unreliable questionnaire completed by all respondents. If you cannot determine what the data mean, the amount of information collected is irrelevant. Consequently, completing "convenient" questionnaires that cannot yield meaningful information is a poorer use of respondents' time and effort than completing a somewhat longer version that produces valid data. Thus, using inadequately brief assessment methods may have ethical as well as scientific implications.

Summary and Preview

This chapter stresses that measurement is a fundamental activity in all branches of science, including the behavioral and social sciences. Psychometrics, the specialty area of the social sciences that is concerned with measuring social and psychological phenomena, has historical antecedents extending back to ancient times. In the social sciences, theory plays a vital role in the development of composite measurement instruments, which are collections of items that reveal the level of an underlying variable. Often, in the behavioral and social sciences, such a measurement tool will fit the definition of a scale. However, not all collections of items constitute scales. Developing composite measurement tools may be more demanding than selecting items casually; however, the costs of using casually constructed measures usually greatly outweigh the benefits.

The following chapters cover the rationale and methods of scale development in greater detail. Chapter 2 explores the latent variable, the underlying construct that a scale attempts to quantify, and presents the theoretical bases for the methods described in later chapters. Chapter 3 provides a conceptual foundation for understanding reliability and the logic underlying the reliability coefficient. Chapter 4 reviews validity, while Chapter 5 is a practical guide to the steps involved in scale development. Chapter 6 introduces factor analytic concepts and describes their use in scale development. Chapter 7 is an exploration of an alternative type of aggregate measure, the index. Chapter 8 is a conceptual overview of an alternative approach to scale development—item response theory. Finally, Chapter 9 briefly discusses how scales fit into the broader research process.

Exercises

1. What are the key differences between a *scale* and an *index* as we have described them?
2. Two professions that have long histories of assessment are education (through the development and use of standardized ability tests) and psychiatry (through the specification and application of standardized diagnostic criteria). What are some of the key differences between how these two fields of inquiry have approached assessment?