1

Measurement, Measuring Instruments, and Psychometric Theory

1.1 Constructs and Their Importance in the Behavioral and Social Sciences

Measurement pervades almost every aspect of modern society, and measures of various kinds often accompany us throughout much of our lives. Measurement can be considered an activity consisting of the process of assigning numbers to individuals in a systematic way as a means of representing their studied properties. For example, a great variety of individual characteristics, such as achievement, aptitude, or intelligence, are measured frequently by various persons—e.g., teachers, instructors, clinicians, and administrators. Because the results of these measurements can have a profound influence on an individual's life, it is important to understand how the resulting scores are derived and what the accuracy of the information about examined properties is, which these numbers contain. For the social, behavioral, and educational sciences that this book is mainly directed to, measurement is of paramount relevance. It is indeed very hard for us to imagine how progress in them could evolve without measurement and the appropriate use of measures. Despite its essential importance, however, measurement in these disciplines is plagued by a major problem. This problem lies in the fact that unlike many physical attributes, such as, say, length or mass, behavioral and related attributes cannot be measured directly.

Widely acknowledged is also the fact that most measurement devices are not perfect. Physical scientists have long recognized this and have been concerned with replication of their measurements many times to obtain results in which they can be confident. Replicated measures can provide the average of a set of recurring results, which may be expected to represent a more veridical estimate of what is being appraised than just a single measure. Unfortunately, in the social, behavioral, and educational disciplines, commonly obtained measurements cannot often be replicated as straightforwardly and confidently as in the physical sciences, and there is no instrument like a ruler or weight scale that could be used to directly measure, say, intelligence, ability, depression, attitude, social cohesion, or alcohol dependence, to name only a few of the entities of special interest in these and related disciplines. Instead, these are only indirectly observable entities, oftentimes called constructs, which can merely be inferred from overt behavior (see discussion below for a stricter definition of 'construct'). This overt behavior represents (presumably) the construct manifestation. More specifically, observed behaviors—such as performance on certain tests or items of an inventory or self-report, or responses to particular questions in a questionnaire or ability test—may be assumed to be indicative manifestations of these constructs. That is, each construct is a theoretical entity represented by a number of similar manifested behaviors. It is this feature that allows us to

consider a construct an abstraction from, and synthesis of, the common features of these manifest behaviors.

We can define a construct as an abstract, possibly hypothetical entity that is inferred from a set of similar demonstrated or directly observed behaviors. That is, a construct is abstracted from a cluster of behaviors that are related among themselves. In other words, a construct represents what is common across these manifested behaviors. In this role, a construct is conceptualized as the hidden 'source' of common variability, or covariability, of a set of similar observable behaviors. We note that a construct may as well be a theoretical concept or even a hypothetical entity and may also not be very well defined initially on its own in a substantive area.

There is a set of general references to the notion of construct that have become popular in the social and behavioral sciences. At times constructs are called latent, unobserved, or hidden variables; similarly, a construct may be referred to as an underlying dimension, latent dimension, or latent construct. We will use these terms synonymously throughout the text. Each of them, and in particular the last two mentioned, emphasize a major characteristic feature of constructs used in these disciplines. This is the fact that in contrast with many physical attributes, constructs cannot be directly observed or measured.

In this book, we will treat a construct as a latent continuum, i.e., a latent or unobserved continuous dimension, along which subjects are positioned and in general differ from one another. The process of measurement aims at differentiating between their unknown positions along this dimension and possibly attempting to locate them on it. Because the constructs, i.e., these latent dimensions, are not directly observable or measurable, unlike, say, height or weight, it is easily realized that the above-mentioned major problem of measurement resides in the fact that the individuals' exact locations on this continuum are not known. In addition, as we will have ample opportunities to emphasize in later chapters, the locations of the studied subjects on a latent continuum are not directly and precisely measurable or observable. For this reason, examined individuals cannot be exactly identified on the latent dimension corresponding to a construct under consideration. That is, we can think of each studied subject, whether in a sample or population of interest, as possessing a location—or a score, in quantitative terms—on this dimension, but that location is unknown and in fact may not be possible to determine or evaluate with a high level of accuracy.

Most entities of theoretical and empirical interest in the behavioral and social sciences can be considered latent constructs. Some widely known examples are motivation, ability, aptitude, opinion, anxiety, and general mental ability, as well as extraversion, neuroticism, agreeableness, openness to new experience, and conscientiousness (the so-called Big Five factors of human personality, according to a popular social psychology theory; e.g., McCrae & Costa, 1996). The constructs typically reflect important sides of behavioral and social phenomena that these disciplines are interested in studying. Despite our inability (at least currently) to measure or observe constructs directly in them, these constructs are of special theoretical and empirical relevance. Specifically, the study of their relationships is of particular interest in these sciences. Entire theories in them are based on constructs and the ways in which they relate, or deal with how some constructs could be used to understand better if not predict—within the limits of those theories—other constructs under consideration. Progress in the social, behavioral, and educational disciplines is oftentimes marked by obtaining deeper knowledge about the complexity of relationships among constructs of concern, as well as the conditions under which these relationships occur or take particular forms.

Although there are no instruments available that would allow us to measure or observe constructs directly, we can measure them indirectly. This can be accomplished using proxies of the constructs. These proxies are the above-indicated behavior manifestations, specifically of the behaviors that are related to the constructs. For example, the items in the Beck Depression Inventory (e.g., Beck, Rush, Shaw, & Emery, 1979) can be considered proxies for depression. The subtests comprising an intelligence test battery, such as the Wechsler Adult Intelligence Scale (WAIS; e.g., Chapter 3), can also be viewed as proxies of intelligence. The questions in a scale of college aspiration can be treated as proxies for the unobserved construct of college aspiration. The responses to the problems in a mathematics ability test can similarly be considered proxies for (manifestations of) this ability that is of interest to evaluate.

A widely used reference to these proxies, and in particular in this text, is as indicators of the corresponding latent constructs. We stress that the indicators are not identical to the constructs of actual concern. Instead, the indicators are only manifestations of the constructs. Unlike the constructs, these manifestations are observable and typically reflect only very specific aspects of the constructs. For example, a particular item in an anxiety scale provides information not about the entire construct of anxiety but only about a special aspect of it, such as anxiety about a certain event. An item in an algebra knowledge test does not evaluate the entire body of knowledge a student is expected to acquire throughout a certain period of time (e.g., a school semester). Rather, that item evaluates his or her ability to execute particular operations needed to obtain the correct answer or to use knowledge of a certain fact(s) or relationships that were covered during the pertinent algebra instruction period in order to arrive at that answer.

No less important, an indicator can in general be considered not a perfect measure of the associated construct but only a fallible manifestation (demonstration) or proxy of it. There are many external factors when administering or measuring the indicator that are unrelated to the construct under consideration this indicator is a proxy of, which may also play a role. For instance, when specific items from a geometry test are administered, the examined students' answers are affected not only by the corresponding skills and knowledge possessed by the students but also by a number of unrelated factors, such as time of the day, level of prior fatigue, quality of the printed items or other presentation of the items, and a host of momentary external (environment-related) and internal factors for the students. Later chapters will be concerned in more detail with various sources of the ensuing error of measurement and will provide a much more detailed discussion of this critical issue for behavioral and social measurement (see in particular Chapters 5 and 9 on classical test theory and generalizability theory, respectively).

This discussion demonstrates that the indicators of the studied constructs, as manifestations of the latter, are the actually observed and error-prone variables on which we obtain data informing about these constructs. Yet collecting data on how individuals perform on these indicators is not the end of our endeavors but only a means for accomplishing the goal, which is evaluation of the constructs of concern. Indeed, we are really interested in the underlying constructs and how they relate to one another and/or other studied variables. However, with respect to the constructs, all we obtain data on are their manifestations, i.e., the individual performance on the construct indicators or proxies. On the basis of these data, we wish to make certain inferences about the underlying constructs and their relationships and possibly those of the constructs to other observed measures. This is because, as mentioned, it is the constructs themselves that are of actual interest. They help us better understand studied phenomena and may allow us to control, change, or even optimize these and related phenomena. This lack of identity between the indicators,

on the one hand, and the constructs with which they are associated, on the other hand, is the essence of the earlier-mentioned major problem of measurement in the behavioral and social sciences.

Whereas it is widely appreciated that constructs play particularly important roles in these sciences, the independent existence of the constructs cannot be proved beyond any doubt. Even though there may be dozens of (what one may think are) indicators of a given construct, they do not represent by themselves and in their totality sufficient evidence in favor of concluding firmly that their corresponding latent construct exists on its own. Furthermore, the fact that we can come up with a 'meaningful' interpretation or a name for a construct under consideration does not mean that it exists itself in reality. Nonetheless, consideration of constructs in theories reflecting studied phenomena has proved over the past century to be highly beneficial and has greatly contributed to substantial progress in the behavioral, social, and educational sciences.

1.2 How to Measure a Construct

Inventing a construct is obviously not the same as measuring it and, in addition, is far easier than evaluating it. In order to define a construct, one needs to establish a rule of correspondence between a theoretical or hypothetical concept of interest on the one hand and observable behaviors that are legitimate manifestations of that concept on the other hand. Once this correspondence is established, that concept may be viewed as a construct. This process of defining, or developing, a construct is called operational definition of a construct.

As an example, consider the concept of preschool aggression (cf. Crocker & Algina, 1986). In order to operationally define it, one must first specify what types of behavior in a preschool play setting would be considered aggressive. Once these are specified, in the next stage a plan needs to be devised for obtaining samples of such aggressive behavior in a standard situation. As a following step, one must decide how to record observations, i.e., devise a scheme of data recording for each child in a standard form. When all steps of this process are followed, one can view the result as an instrument, or a 'test' ('scale'), for measuring preschool aggression. That is, operationally defining a construct is a major step toward developing an instrument for measuring it, i.e., a test or scale for that construct.

This short discussion leads us to a definition of a test as a standard procedure for obtaining a sample from a specified set of overt behaviors that pertain to a construct under consideration (cf. Murphy & Davidshofer, 2004). In other words, a test is an instrument or device for sampling behavior pertaining to a construct under study. This measurement is carried out under standardized conditions. Once the test is conducted, established objective rules are used for scoring the results of the test. The purpose of these rules is to help quantify in an objective manner an examined attribute for a sample (group) of studied individuals. Alternative references to 'test' that are widely used in the social and behavioral sciences are scale, multiple-component measuring instrument, composite, behavioral measuring instrument, or measuring instrument (instrument). We will use these references as synonyms for 'test' throughout the remainder of this book.

As is well-known, tests produce scores that correspond to each examined individual. That is, every subject participating in the pertinent study obtains such scores when the

test is administered to him or her. These scores, when resulting from instruments with high measurement quality, contain information that when appropriately extracted could be used for making decisions about people. These may be decisions regarding admission into a certain school or college, a particular diagnosis, therapy, or a remedial action if needed, etc. Because some of these decisions can be very important for the person involved and possibly his or her future, it is of special relevance that the test scores reflect indeed the attributes that are believed (on theoretical and empirical grounds) to be essential for a correct decision. How to develop such tests, or measuring instruments, is an involved activity, and various aspects of it represent the central topics of this book.

The following two examples demonstrate two main types of uses of test scores, which are interrelated. Consider first the number of what could be viewed as aggressive acts displayed by a preschool child at a playground during a 20-minute observation period. Here, the researcher would be interested in evaluating the trait of child aggression. The associated measurement procedure is therefore often referred to as trait evaluation. Its goal is to obtain information regarding the level of aggression in a given child, i.e., about the position of the child along the presumed latent continuum representing child aggression. As a second example, consider the number of correctly solved items (problems, tasks, questions) by a student in a test of algebra knowledge. In order for such a test to serve the purpose for which it has been developed, viz., assess the level of mastery of an academic subject, the test needs to represent well a body of knowledge and skills that students are expected to acquire in the pertinent algebra subject over a certain period (e.g., a school semester or year). Unlike the first example, the second demonstrates a setting where one would be interested in what is often referred to as domain sampling. The latter activity is typically the basis on which achievement tests are constructed. Thereby, a domain is defined as the set of all possible items that would be informative about a studied ability, e.g., abstract thinking ability. Once this definition is complete, a test represents a sample from that domain. We notice here that the relationship of domain to test is similar to that of population to sample in the field of statistics and its applications. We will return to this analogy in later chapters when we will be concerned in more detail with domain sampling and related issues.

We thus see that a test is a carefully developed measuring instrument that allows obtaining meaningful samples of behavior under standardized conditions (Murphy & Davidshofer, 2004). In addition, a test is associated with objective, informative, and optimal assignment of such numerical scores that reflect as well as possible studied characteristics of tested individuals. Thereby, the relationships between the subject attributes, i.e., the degree to which the measured individuals possess the constructs of interest, are expected to be reflected in the relationships between the scores assigned to them after test administration and scoring.

We emphasize that a test is not expected to provide exhaustive measurement of all possible behaviors defining an examined attribute or construct. That is, a test does not encapsulate all behaviors that belong to a pertinent subject-matter area or domain. Rather, a test attempts to 'approximate' that domain by sampling behaviors belonging to it. Quality of the test is determined by the degree to which this sample is representative of those behaviors.

With this in mind, we are led to the following definition of a fundamental concept for the present chapter as well as the rest of this book, that of behavioral measurement. Accordingly, behavioral measurement is the process of assigning in a systematic way quantitative values to the behavior sample collected by using a test (instrument, scale),

which is administered to each member of a studied group (sample) of individuals from a population under consideration.

1.3 Why Measure Constructs?

The preceding discussion did not address specific reasons as to why one would be interested in measuring or be willing to measure constructs in the social and behavioral sciences. In particular, a question that may be posed at this point is the following: Because latent constructs are not directly observable and measurable, why would it be necessary that one still attempt to measure them?

To respond to this question, we first note that behavioral phenomena are exceedingly complex, multifaceted, and multifactorially determined. In order to make it possible to study them, we need special means that allow us to deal with their complexity. As such, the latent constructs can be particularly helpful. Their pragmatic value is that they help classify and describe individual atomistic behaviors. This leads to substantial reduction of complexity and at the same time helps us to understand the common features that interrelated behaviors possess. To appreciate the value of latent constructs, it would also be helpful to try to imagine what the alternative would imply, viz., not to use any latent constructs in the behavioral, social, and educational disciplines. If this alternative would be adopted as a research principle, however, we would not have means that would allow us to introduce order into an unmanageable variety of observed behavioral phenomena. The consequence of this would be a situation in which scientists would need to deal with a chaotic set of observed phenomena. This chaos and ensuing confusion would not allow them to deduce any principles that may underlie or govern these behavioral phenomena.

These problems could be resolved to a substantial degree if one adopts the use of constructs that are carefully conceptualized, developed, and measured through their manifestations in observed behavior. This is due to the fact that constructs help researchers to group or cluster instances of similar behaviors and communicate in compact terms what has in fact been observed. Moreover, constructs are also the building blocks of most theories about human behavior. They also account for the common features across similar types of behavior in different situations and circumstances. For these reasons, constructs can be seen as an indispensable tool in contemporary behavioral, social, and educational research.

This view, which is adopted throughout the present book, also allows us to consider a behavioral theory as a set of statements about (a) relationships between behavior-related constructs and (b) relationships between constructs on the one hand and observable phenomena of practical (empirical) consequence on the other hand. The value of such theories is that when correct, or at least plausible, they can be used to explain or predict and possibly control or even optimize certain patterns of behavior. The behavioral and social sciences reach such theory levels through empirical investigation and substantiation, which is a lengthy and involved process that includes testing, revision, modification, and improvement of initial theories about studied phenomena. Thereby, an essential element in accomplishing this goal is the quantification of observations of behaviors that are representative of constructs posited by theory. This quantification is the cornerstone of what measurement and in particular test theory in these sciences is about.

1.4 Main Challenges When Measuring Constructs

Given that the constructs we are interested in are such abstractions from observed interrelated behaviors, which can be measured only indirectly, the development of instruments assessing them represents a series of serious challenges for the social, behavioral, or educational researcher. In this section, we discuss several of these challenges, which developers of multiple-component measuring instruments—e.g., tests, scales, self-reports, subscales, inventories, testlets, questionnaires, or test batteries—have to deal with when attempting to measure constructs under consideration.

First, there is no single approach to construct measurement, which would be always applicable and yield a satisfactory measuring instrument. This is because construct measurement is based on behaviors deemed to be relevant for the latent dimension under study. Hence, it is possible that two theorists having in mind the same construct may select different types of behavior to operationally define that construct. As an example, consider the situation when one wishes to measure elementary school students' ability to carry out long division (e.g., Crocker & Algina, 1986). To this end, one could decide to use tests that are focused on (a) detecting errors made during this process, (b) describing steps involved in the process, or, alternatively, (c) solving a series of division problems. Either of these approaches could be viewed as aiming at evaluating the same construct under consideration, the ability to conduct long division.

A second challenge emerges from the fact that construct measurement is typically based only on limited samples of behavior. The reason is that in empirical research, it is impossible to confront individual subjects using a given instrument with all situations in which a certain construct is presumed to be indicated, demonstrated, or manifested. How many such situations must be included in a test, and of what kinds would they need to be, in order to provide an adequate sample of the pertinent behavioral domain? This is a major problem in developing a sound measurement procedure, specifically one in which domain sampling is involved. For example, an educational researcher cannot ask examined students to solve all possible division problems when he or she is interested in measuring, say, ability to carry out number division. Instead of that, the researcher will need to sample from the domain consisting of all these problems.

As a third challenge, behavioral and social measurements are typically subject to error, i.e., they are error prone rather than error free. Error of measurement results in part from inconsistencies in obtained individual scores, some of which are due to sampling of tasks and/or the actual process of measurement. Main contributors to this error are effects of factors that are unrelated to the studied construct but affect individual performance on a given measure (indicator) of it. Measurement error encompasses such factor effects and will be the subject of further and more detailed discussion in a later chapter of this book dealing with classical test theory (see Chapter 5), which represents a main approach to behavioral measurement. As we will see in that chapter, a persistent problem in measurement is how to evaluate the extent of error present in a given set of observations pertaining to a construct in question. The concepts of reliability and validity, which we will discuss in detail subsequently, are in particular concerned with this problem.

A fourth challenge results from the lack of (substantively) well-defined units as well as origins, or zero points, on used measurement scales in most behavioral and social measurements. For instance, is an individual score of 20 obtained on an aggression test indicative of twice as much aggression than that in a person with a score of 10? Furthermore, does an individual score of 0 on an ability test demonstrate no ability at all? None of these

and many other similar questions can necessarily be answered affirmatively because of this lack of well-defined units and origins of measurement in most of present-day behavioral and social research.

Last but not least, constructs cannot be defined only in terms of operational definitions but must also demonstrate relationships (or lack thereof) with other constructs and observable phenomena. Typically, a construct is considered for use in a subject-matter domain where there has been prior theoretical work as well as other constructs studied previously. Hence, this new construct should demonstrate relationships (or, if appropriate, lack thereof) to those already available constructs in that substantive area. With this in mind, there are two levels at which a construct should be defined (e.g., Lord & Novick, 1968). One is operational, which deals with how measurement of a considered construct is to be obtained, i.e., how that construct relates to observable behavior. Another equally important level of construct definition is nomothetic. Accordingly, the construct should fit well into an already existing theoretical 'net' of other, previously established constructs and their relationships in the substantive area of concern. That is, a new construct should demonstrate predictable relationships (whether strong or weak, none, positive or negative) with other constructs available in the subject-matter domain of concern. When this is indeed the case, these relationships provide a basis for interpreting measurements obtained on the proposed construct. If there are no such relationships between the new and earlier established constructs, however, i.e., if there is no empirical evidence for them, there is little value if any in a newly proposed construct.

As an example, consider a test of abstract thinking ability for high school students. That is, suppose this ability is the construct of interest. The nomothetic level of construct definition requires here that scores obtained with this test should exhibit (a) notable relationships with scores on existing tests of algebra, geometry, inductive reasoning, and figural relations and (b) not nearly as strong relationships with scores on tests of knowledge of a foreign language, or history knowledge, as well as other weakly or essentially unrelated on theoretical grounds constructs. In other words, scores on the test of abstract thinking ability should be related markedly with scores on the tests of algebra, geometry, inductive reasoning, and figural relations and at the same time should be only up to marginally related with scores on tests of foreign language or history knowledge.

We will return to this last challenge for behavioral measurement in a later chapter of this book (Chapter 8). We will see then that this challenge has fundamental relevance for most of what measurement is about, i.e., validity of measurement, which can be viewed as the bottom line of behavioral and social measurement. In particular, we will observe that this challenge is of special importance for the most encompassing form of validity, called construct validity.

1.5 Psychometric Theory as a Scientific Discipline

The preceding discussion in this chapter allows us to define now the subject of test theory and more generally that of behavioral measurement, which is often referred to as psychometric theory. (In this text, we will use frequently 'test theory' and 'psychometric theory' as synonyms.) Specifically, psychometric theory is a scientific discipline that is concerned with the study of the above and related pervasive problems and

challenges of human behavior measurement (see Section 1.4), using a particular set of methods developed to systematically manage or resolve them (cf. Crocker & Algina, 1986). That is, psychometric theory deals with (a) evaluating the degree to which these problems affect behavioral measurement in a given situation and (b) developing methods to overcome or minimize the adverse impact of these and related problems and challenges.

As a scientific discipline, psychometric theory is based on formal logic as well as mathematical and statistical methods and models (cf. McDonald, 1999). These also underlie standard practices in the process of construction, development, and revision of measuring instruments, as well as in their applications. Becoming aware of these models and underlying methods as well as their assumptions and limitations, in order to ensure improved practice in test construction and use of test information in decision making, is the primary goal of most measurement and test theory treatments, including the present book.

For the purposes of this text, we also need to draw a distinction between (a) psychometric theory and (b) behavioral assessment. Psychometric theory (test theory) is a theoretically oriented field of study, which is of general relevance regardless of the particular test, scale, or measuring instrument used in a given situation where evaluation of behavioral attributes is required. This book is concerned with psychometric theory. In contrast to it, behavioral assessment is primarily an applied, rather than theoretical, subject that is usually focused on administration and interpretation of particular tests used as measuring instruments under certain circumstances. Hence, the subject of behavioral assessment has relevance with regard to specific measuring instruments. This text will not be concerned with behavioral assessment, and we will not discuss it any further. Instead, we emphasize that psychometric theory provides a general framework for behavioral measuring instrument development, including instrument construction, revision, and modification. In order to be in a position to accomplish this goal, psychometric theory is based on general mathematical and statistical approaches, methods, and models that are valid irrespective of any behavioral theory that a researcher may be adopting. Being this general, psychometric theory is useful for measurement of any behavioral construct, such as an attribute, ability, trait, attitude, or aptitude. Although being based on generally valid mathematical and statistical principles, psychometric theory at the same time has been uniquely developed as a scientific discipline to meet the specific needs of behavioral, social, and educational measurement.

With these goals, psychometric theory provides essential input into research in the behavioral and social sciences, especially as far as development and selection of instruments and procedures for quantification of observations on studied variables are concerned. Psychometric theory has some of its major applications in the process of 'pretesting' and improvement of measuring instruments so as to minimize possible error and ensure highest validity with regard to variables involved in examined research hypotheses and questions. In this way, psychometric theory contributes to the provision of a set of high-quality, pretried measuring instruments with known characteristics, from which a scientist can select for use in a given research situation. Psychometric theory is a scientific discipline dealing with the study of how general epistemological problems related to measurement impact the process of quantification of aspects of behavioral phenomena under investigation, as well as with methods aimed at providing information of highest quality about associated constructs from their indirect observations (cf., e.g., Crocker & Algina, 1986).

1.6 Initial Steps in Instrument Construction

As has been indicated on a few occasions in this chapter, our concerns in this book will be with what is oftentimes referred to as subject-centered (person-centered) measurement. In this activity, the goal is to 'reveal' the location of individuals on a (presumed) quantitative continuum with respect to a particular behavioral construct, such as aggression, ability, intelligence, motivation, depression, aspiration, etc. Similarly, instruments of interest to us can also aim to provide information with regard to the level of mastery or proficiency that individuals or examinees possess in a particular subject-matter area.

The remainder of this text deals with a systematic approach to measuring instrument construction with wide applicability to various types of instruments that can be used across the behavioral, social, and educational disciplines. We will treat test construction as a principled approach to the development of behavioral measuring instruments. This process includes a number of stages that we will briefly review in this section. For a more detailed discussion, we refer readers to Allen and Yen (1979), Crocker and Algina (1986), Guilford and Fruchter (1978), McDonald (1999), and Suen (1990); these sources have also significantly influenced the following discussion in this section.

One of the earliest steps in this process is to decide for what purpose the resulting test (instrument) scores will be utilized. A main question a developer should ask initially is whether the instrument is supposed to differentiate among individuals with regard to a given construct or whether the test is to provide scores that describe in some sense absolute (as opposed to relative) levels of proficiency in a given content area. The former is the question raised when an instrument is supposed to evaluate an underlying trait—an activity that is often referred to as trait evaluation—whereas the latter is the query of relevance when one is interested in obtaining information about the level of achievement accomplished by studied subjects and is at times referred to as achievement testing. Within the achievement evaluation approach, the specific goal of the instrument needs to be clarified next. For instance, if an educational diagnostic test is to be designed, areas of specific weaknesses for low-ability students should be identified, and within them items that are relatively easy to solve for the general population should be developed. Alternatively, if a test is required to discriminate well among subjects over a broad range of ability, then items of medium difficulty level are needed, whereas a test measuring high levels of achievement should be composed of more difficult items.

At the next stage of instrument development, the researcher needs to identify behaviors representing the underlying (presumed) construct or define the subject-matter domain of relevance if involved in achievement evaluation (cf. Crocker & Algina, 1986). Either activity requires a thorough content analysis, which should include a critical review of available research. At times, types of behaviors most frequently studied by other scientists can be used when defining a construct of interest, or alternatively delineating extremely low or high performance levels can be of particular help when outlining the domain of knowledge and/or skills to be evaluated by the test under development. Input from substantive experts then will be especially helpful in narrowing down types of behavior pertaining to the construct of interest. Similarly, when developing tests of achievement, researchers should also take instructional objectives into account.

Whenever interest lies in measuring proficiency—as in achievement and aptitude tests—one becomes involved in what is referred to as criterion-referenced measurement, unlike the norm-referenced measurement that is usually conducted when one is concerned with trait evaluation. The underlying basis for the former type of measurement is the activity

of domain sampling, which we briefly discussed earlier. The meaning of resulting test scores is then obtained by reference to the criterion (acceptable level of mastery), not to the scores of other individuals. Conversely, the meaning of test scores resulting from norm-referenced measurement is derived from their comparison with other individuals' scores on the same test.

In order to engage in domain sampling, an item domain needs to be available. This is a well-defined population of presumably homogeneous items (measuring the same ability) from which one or more test forms may be constructed by selection of a sample(s) of items. The creation of the item domain is facilitated by producing a set of item-domain specifications so structured that items written according to them could be considered to be interchangeable. In some cases, an item domain could be defined by certain algorithms for item creation, as in mathematics ability testing. More generally, the generation of an item domain is a sophisticated process that requires professional expertise and special training. The specific activity of item writing that is involved thereby will not be a concern of this text. For extensive discussion on the topic we refer the reader to Thissen and Wainer (2001) and Haladyna (2004; see also McDonald, 1999).

In a next step of the process of test construction, an early form of the instrument is tried out (piloted) on a small sample of subjects from the intended population (cf. Suen, 1990). Thereby, one invites comments from participants on how they perceived each item in the test. Following that, the descriptive statistics for the response distributions on each item are examined. This preliminary tryout can help make decisions about revision of some items, upon which usually a 'field test' on a larger sample from the studied population is conducted. Statistical properties of the items are examined then, possibly using procedures known as item analysis—frequently referred to as classical item analysis—which we will be in part concerned with in Chapters 2 and 4. At this stage, it may be decided that some items should be removed from the test. A study of reliability and validity of the instrument is then to be carried out, based on results of this field test. Corresponding revisions are usually undertaken then, with the aim of producing an improved version of the instrument. Establishing guides for test score interpretation is the next step, which is an involved process requiring professional attention. This process is part of the subject behavioral assessment, which as mentioned before will not be of concern in this book.

This section aimed only to provide a brief discussion of some of the steps involved in the process of instrument (test, scale) construction. Subsequent chapters will be concerned with how one can study important properties of the resulting instrument and how to improve these. Before we turn to them, however, we take another look at measuring instruments used in the behavioral and social sciences, with the particular aim of making a few observations with regard to their building blocks.

1.7 Measuring Instrument Scores as Composites

The earlier discussion in this chapter indicated that because of a number of difficult problems and challenges in behavioral and social measurement, one typically aims at obtaining multiple sources of information about a studied construct. They are usually furnished by distinct measures of construct manifestations, and their interrelationships may permit evaluation of that latent dimension.

Such an evaluation is also often aimed at in empirical research by a composite score (overall score) obtained from the multiple measures, or construct indicators, frequently referred to as components of the composite. They are typically individual testlets, items, subscales, subtests, or questions. Hence, a 'composite' can be viewed as synonymous to a 'test' consisting of multiple components or, more generally, to a 'test battery'. An alternative reference to a composite is scale, multiple-component measuring instrument, or just measuring instrument. Each of its components yields a separate score, and usually their sum furnishes an overall score, often referred to as the sum score. For example, the score on a depression test is usually the sum of the scores a person obtains on each of its items. The score on an algebra knowledge test is often the number of correct answers on the set of problems to be solved by the examined students, which is the sum of formal indicators for a true versus false response on these tasks. Once obtained, the composite score may be considered an indirect measure of the underlying latent construct.

More generally, a composite is a sum of scores on subtests (items, components, testlets, subscales, questions), whereby this sum may be directly resulting from their simple addition or such after their multiplication with some component-specific weights. In the former case, the composite is also referred to as unit weighted or unweighted, whereas in the latter case it is typically called a weighted composite (or weighted scale). This book deals predominantly with unweighted composites and will use the term 'test', 'scale', or 'multiple-component measuring instrument' ('instrument') as synonymous to 'composite' or 'sum score'. The majority of the procedures discussed in this book are also applicable to weighted scales with known weights, and extensions for some of them are available in case of unknown weights (e.g., Li, 1997; Li, Rosenthal, & Rubin, 1996; Raykov, 2004, 2005; Raykov & Hancock, 2005; Raykov & Penev, 2006, 2009, and references therein).

When administered and scored in empirical research, behavior measuring instruments give rise to a table called 'person–item score matrix' that contains the raw data, i.e., the data collected from the examined individuals. In this table, the rows correspond to persons and the columns correspond to the items, with a possibly subsequently added last row and last column as totals for items and for persons, respectively. Oftentimes, the entries of this table are 0s and 1s, corresponding to correct or false response, respectively (or to present–absent, endorsed–not endorsed, yes–no, agree–disagree answers). To analyze the so-obtained raw data set, one needs to make use of certain statistical concepts, methods, and models. A brief review of some important notions in this respect is provided in the next chapter.