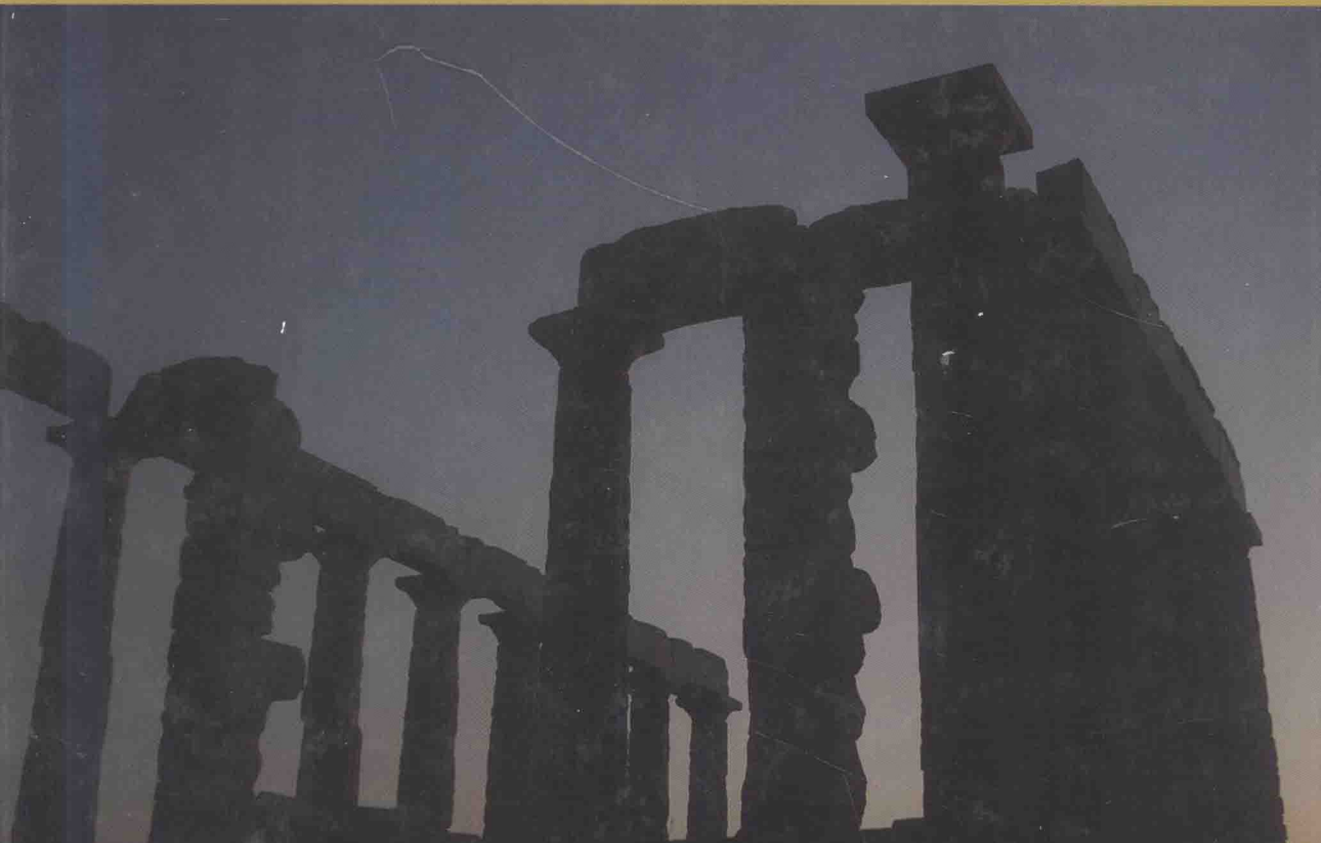
Introduction to Psychometric Theory



Tenko Raykov and George A. Marcoulides

Introduction to Psychometric Theory

Tenko Raykov
Michigan State University

George A. Marcoulides
University of California at Riverside

常州大学图书馆
藏书章

Routledge
Taylor & Francis Group
New York London

Routledge
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Routledge
Taylor & Francis Group
27 Church Road
Hove, East Sussex BN3 2FA

© 2011 by Taylor and Francis Group, LLC
Routledge is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-0-415-87822-7 (Hardback)

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Raykov, Tenko.
Introduction to psychometric theory / Tenko Raykov, George Marcoulides.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-415-87822-7 (hardcover : alk. paper)
1. Psychometrics. I. Marcoulides, George A. II. Title.

BF39.R34 2011
150.1'5195--dc22 2010032526

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the Psychology Press Web site at
http://www.psypress.com

Introduction to Psychometric Theory

Preface

This book arose from our experiences accumulated over a number of years in teaching measurement, testing, and psychometric theory courses. Throughout this time, we have not been able to identify a single textbook that would consistently serve all our instructional purposes. To meet these objectives, we developed a series of lecture notes whose synthesis this text represents. In the meantime, measurement and test theory evolved further into a very extensive field, which is challenging to cover in a single course. With this in mind, our book is meant to introduce behavioral, social, educational, marketing, business and biomedical students and researchers to this methodological area. To this end, we have identified a set of topics in the field, which are in our view of fundamental importance for its understanding and permit readers to proceed subsequently to other more advanced subjects.

Our pragmatic goal with this text is to provide a coherent introduction to psychometric theory, which would adequately cover the basics of the subject that we consider essential in many empirical settings in the behavioral, social, and educational sciences. We aim at a relatively non-technical introduction to the subject, and use mathematical formulas mainly in their definitional meaning. The audience for which this book is most suitable consists primarily of advanced undergraduate students, graduate students, and researchers in the behavioral, social, educational, marketing, business and biomedical disciplines, who have limited or no familiarity with the mathematical and statistical procedures that are involved in measurement and testing. As prerequisites for the book, an introductory statistics course with exposure to regression analysis and analysis of variance (ANOVA) is recommended, as is initial familiarity with some of the most widely circulated statistical analysis software in these sciences (e.g., IBM SPSS, SAS, STATA, or R; a brief introduction to R is provided in Section 2.8 of Chapter 2; see also Appendix to Chapter 11).

We believe that it is not possible to introduce a reader to psychometric theory, as available to us currently, without repeated use of statistical software. After the first couple of chapters, we utilize nearly routinely the latent variable modeling program M*plus* that is becoming increasingly popular and widely used across the social and behavioral sciences, in addition to our use of R, IBM SPSS, and SAS on a number of occasions in several of the chapters. On the webpage for this book (www.psypress.com/psychometric-theory), we provide essentially all data used in the text formatted in such a way that it can be used with IBM SPSS, SAS, M*plus*, or R. To enhance comprehension, the software codes and associated outputs are also included and discussed at appropriate places in the book. In addition, instructors will find helpful PowerPoint lecture slides and questions and problems for each chapter in the book, which are also available at the above webpage.

Our hope is that readers will find this text to present a useful introduction to and a basic treatment of psychometric theory, as well as prepare them for studying more advanced topics within this rich and exciting scientific field. There are two features that seem to set apart our book from others in this field: (i) the extensive and consistent use of the comprehensive framework of latent variable modeling (LVM); and (ii) the concern with interval estimation throughout, in addition to point estimation of parameters of main interest and model fit evaluation. We believe that in the coming years the applied statistical methodology of LVM will gain even further in popularity and utilization across the social and behavioral sciences. For this reason, we anticipate that LVM will be particularly helpful to

此为试读,需要完整PDF请访问: www.ertongbook.com

students and researchers who embark on their journey into the theoretically and empirically important subject of measurement and test theory in these sciences.

This book has been substantially influenced by the prior work of a number of scholars. We are especially grateful to J. Algina and R. P. McDonald for valuable discussions in various forms on measurement and related topics. Their contributions to psychometrics and books (Crocker & Algina, 1986; McDonald, 1999) have markedly impacted our understanding of the field and this text that can in many respects be seen as positioned closely to theirs. We are also thankful to P. M. Bentler, K. A. Bollen, M. W. Browne, K. G. Jöreksog, C. Lewis, S. Penev, M. D. Reckase, and R. Steyer for many valuable discussions on measurement related topics, as well as to T. Asparouhov, D. M. Dimitrov, G. Mels, S. Penev, P. E. Shrout, and R. E. Zinbarg for their important contributions to our joint research on reliability and related issues in behavioral and social measurement. We are similarly indebted to the M*plus* support team (L. K. Muthén, B. O. Muthén, T. Asparouhov, and T. Nguyen) for instrumental assistance with its applications. Tenko Raykov is also thankful to L. K. Muthén and B. O. Muthén for valuable instructions and discussions on LVM. George A. Marcoulides is particularly grateful to R. J. Shavelson and N. M. Webb for their years of guidance and discussions on a variety of measurement related topics. A number of our students provided very useful and at times critical feedback on the lecture notes we first developed for our courses in psychometric theory, from which this book emerged. We are also grateful to several reviewers for their critical comments on an earlier draft of the manuscript, which contributed substantially to its improvement: M. Meghan Davidson (University of Nebraska–Lincoln); Robert Henson (The University of North Carolina at Greensboro); Joseph J. Palladino (University of Southern Indiana); Scott L. Thomas (Claremont Graduate University); Andre A. Rupp (University of Maryland); and Jennifer Rose (Wesleyan University). Thanks are also due to Debra Riegert and Erin Flaherty from Taylor & Francis for their essential assistance during advanced stages of our work on this project, and to Suzanne Lassandro for valuable typesetting assistance. Last but not least, we are more than indebted to our families for their continued support in lots of ways. The first author thanks Albena and Anna; the second author thanks Laura and Katerina.

Tenko Raykov,
East Lansing, Michigan

George A. Marcoulides,
Riverside, California

Contents

此为试读，需要完整PDF请访问：www.ertongbook.com

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.

此为试读，需要完整PDF请访问：www.ertongbook.com