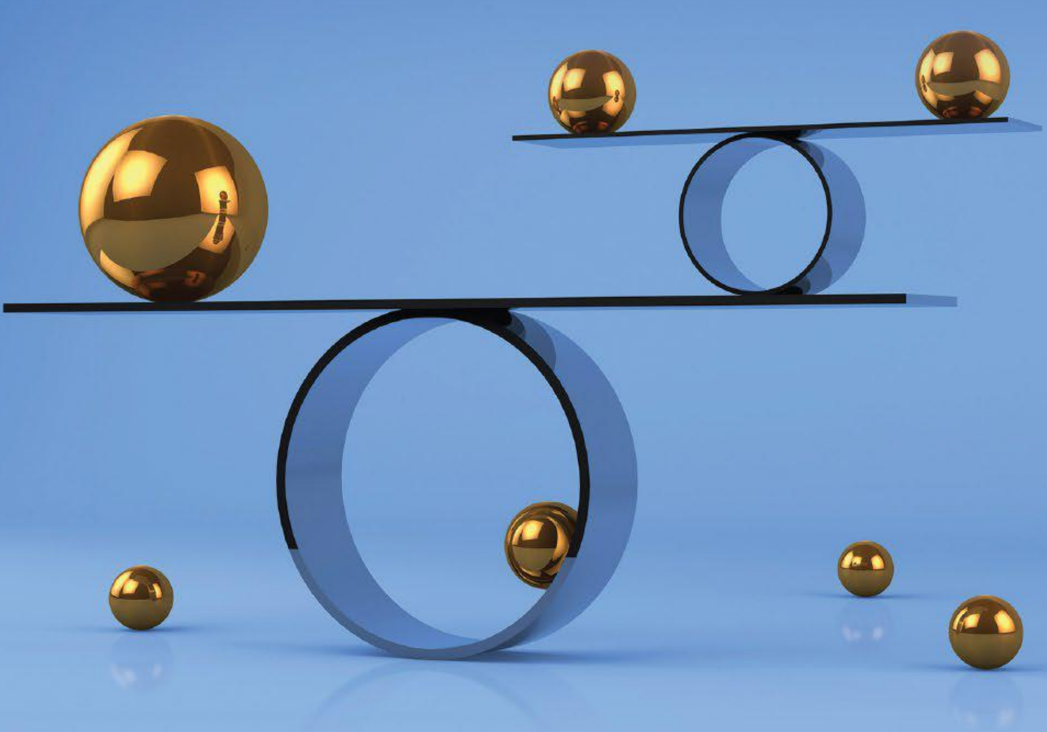
R. Michael Furr

Psychometrics

AN INTRODUCTION • **FOURTH EDITION**



Psychometrics

Fourth Edition

Mike Furr dedicates this book to his wife, Sarah, and to his sons, Sebastian and Abraham.

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 600 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

Psychometrics

An Introduction

Fourth Edition

R. Michael Furr

Wake Forest University

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE

FOR INFORMATION:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
18 Cross Street #10-10/11/12
China Square Central
Singapore 048423

Acquisitions Editor: Leah Fargotstein
Product Associate: Yumna Samie
Production Editor: Rebecca Lee
Copy Editor: Christina West
Typesetter: C&M Digitals (P) Ltd.
Indexer: Integra
Cover Designer: Candice Harman
Marketing Manager: Victoria Velasquez

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Names: Furr, R. Michael, author.

Title: Psychometrics : an introduction / R. Michael Furr, Wake Forest University.

Description: Fourth edition. | Thousand Oaks, California : SAGE, [2022] | Includes bibliographical references and index.

Identifiers: LCCN 2021012675 | ISBN 978-1-0718-2407-8 (paperback ; alk. paper) | ISBN 978-1-0718-2409-2 (epub) | ISBN 978-1-0718-2408-5 (epub) | ISBN 978-1-0718- 2410-8 (pdf)

Subjects: LCSH: Psychometrics.

Classification: LCC BF39 .F87 2022 |
DDC 150.1/5195— dc23

LC record available at https://lccn.loc.gov/2021012675

This book is printed on acid-free paper.

21 22 23 24 25 10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

Measurement is at the heart of all science and of all applications of science. This is true for all areas of science, including the scientific study of human behavior. Behavioral research, whether done by educators, psychologists, or other social scientists, depends on successful measurement of human behavior or of psychological attributes that are thought to affect that behavior. Likewise, the application of psychological or educational science often depends heavily on successful measurement. Indeed, scientifically sound clinical or educational programs and interventions require measurement of the behaviors or psychological attributes of the individuals enrolled in these programs.

This book is concerned with methods used to evaluate the quality of measurement tools, such as psychological tests, that are used in research and applied settings by psychologists and others interested in human behavior. The scientific study of the quality of psychological measures is called psychometrics. Psychometrics is an extremely important field of study, and it can be highly technical. In fact, an article published in the *New York Times* (Herszenhorn, 2006) stated that "psychometrics, one of the most obscure, esoteric and cerebral professions in America, is also one of the hottest."

THE CONCEPTUAL ORIENTATION OF THIS BOOK, ITS PURPOSE, AND THE INTENDED AUDIENCE

Despite the potential "esoteric and cerebral" nature of the field, psychometrics does not need to be presented in a highly technical manner. The purpose of this book is to introduce the *fundamentals* of psychometrics to people who need to understand the properties of measures used in psychology and other behavioral sciences. This book is intended to make these important issues as accessible and as clear as possible, to as many readers as possible—including those who might initially shy away from something that could be "obscure, esoteric, and cerebral."

With this general purpose in mind, this book is intended to be deep but intuitive and relatively nontechnical. This is a surprisingly novel approach to introducing psychometrics. On one hand, this book's treatment is much broader and deeper than the cursory treatment of psychometrics in undergraduate "Tests and Measurement" texts. On the other hand, it is more intuitive and conceptual than the highly technical treatment in books and journal articles intended for use by professionals in the field of psychometrics. Anyone who has taken something equivalent to an undergraduate course in statistics will be comfortable with most of the material in this book. In general, this book is intended to help readers attain a solid and intuitive understanding of the importance, meaning, and evaluation of a variety of fundamental psychometric concepts and issues.

This book is highly relevant for a variety of courses, including Psychological Testing, Psychometrics, Educational Measurement, Personality Assessment, Cognitive Assessment, Clinical Assessment, and, frankly, any type of Assessment course. Moreover, it could be an important part of courses with an emphasis on measurement in many areas of basic and applied science—for example, in medical training, sociology, exercise science, and public health.

Thus, this book is intended for use by advanced undergraduates, graduate students, and professionals across a variety of behavioral sciences and related disciplines. It will be useful to those who need a solid foundation in the basic concepts and logic of psychometrics or measurement more generally. Although this book was not primarily written for people who are intending to become or already are psychometricians, it can serve as a very useful complement to the more technical texts.

To make the topics of psychometrics accessible to the target audience, the book includes illustrative testing situations along with small artificial data sets to demonstrate important features of psychometric concepts. The data sets are used alongside algebraic proofs as a way of underscoring the conceptual meaning of fundamental psychometric concepts. In addition, the book departs from the usual practice of having a separate chapter devoted to statistics. Instead, it introduces statistical concepts throughout the text as needed, and it presents them as tools to help solve particular psychometric problems. For example, the book presents factor analysis initially in the context of exploring the dimensionality of a test. Thus, it ties the statistical procedures to a set of important and intuitive conceptual issues. Experience in the classroom reveals that students benefit when quantitative concepts are linked to problems in this way, as the links seem to reinforce students' understanding of both the statistical procedures and the psychometric concepts.

ORGANIZATIONAL OVERVIEW

The organization of this book is intended to facilitate the readers' insight into core psychometric concepts and perspectives. The first chapter addresses the basic importance of psychological measurement and psychometrics. In addition, it presents important issues and themes that cut across all remaining chapters. This explicit treatment of these issues and themes should help solidify the concepts that are addressed in the later chapters.

Chapters 2–4 address important issues in measurement theory and in the statistical basis of psychometric theory. These chapters are fundamental to a full appreciation and understanding of the later chapters that examine psychometric theory in depth. Specifically, these chapters examine issues of scaling in psychological measurement, concepts in the quantification of psychological differences and the quantification of associations among psychological variables, issues in the interpretation of test scores, and concepts in the meaning and evaluation of test dimensionality. Although these topics can be technical, these chapters cover them in a way that is relatively intuitive and conceptual. That said, Chapter 4 does include an introduction to exploratory factor analysis, the statistical tool frequently used to evaluate dimensionality. Some readers might prefer to avoid an early discussion of factor analysis, and that material certainly could be reserved for a later discussion. That said, there is compelling reason to understand factor analysis early in a discussion of psychometrics, as it has implications for a variety of issues throughout the book.

Chapters 5–7 examine the psychometric concept of reliability, and they differentiate three fundamental aspects of reliability. Chapter 5 introduces the conceptual basis of reliability, focusing on the perspective of classical test theory. Chapter 6 discusses and evaluates common methods of estimating and evaluating the reliability of test scores. Chapter 7 explores the importance of reliability in terms of applied testing, scientific research, and test development. Differentiating these three aspects of reliability hopefully provides readers with an understanding of reliability that is clearer and deeper than what might be obtained from many existing treatments of the topic. All three of these chapters emphasize the psychological meaning of the concepts and procedures, with the purpose of helping readers interpret reliability information meaningfully.

Chapters 8 and 9 examine the psychometric concept of validity. These chapters examine the conceptual foundations of this important psychometric issue, discuss many methods that are used to evaluate validity, and emphasize the important issues to consider in the evaluation process. These chapters adopt a contemporary perspective on validity, as articulated by three national organizations involved in psychological testing—the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). Although it discusses the traditional "tripartite" model of validity (i.e., content validity, criterion validity, and construct validity), which is emphasized in most existing measurement-oriented texts, the chapters represent a more modern view of test validity and the evidence relevant to evaluating test validity.

Chapters 10 and 11 discuss two important threats to the psychometric quality of tests. It is vital to acknowledge and understand the challenges faced by those who develop, administer, and interpret psychological tests. Furthermore, it is crucial to grasp the creative and effective methods that have been developed as ways of coping with many of these challenges to psychometric quality. Chapter 10 explores response biases, which obscure the true differences among individuals taking psychological tests. This chapter describes several different types of biases, demonstrates their deleterious effects on psychological measurement, and examines some methods of preventing or minimizing these effects. Chapter 11 examines test bias, which obscures the true differences between groups of people. This chapter describes the importance of test bias, the methods of detecting different forms of test bias, and the important difference between test bias and test fairness.

Finally, Chapters 12–14 present advanced contemporary approaches to psychometrics. Much of the book reflects the most common psychometric approach in behavioral research and application—classical test theory. These three chapters provide overviews of approaches that move beyond this traditional approach. Chapter 12 presents confirmatory factor analysis (CFA), which is a powerful tool that allows test developers and test users to examine important psychometric issues with flexibility and rigor. Chapter 13 discusses the basic concepts and purpose of generalizability theory, which can be seen as an expansion of the more traditional approaches to psychometric theory. Chapter 14 discusses item response theory (IRT; also known as latent trait theory or modern test theory), which is a very different way of conceptualizing the psychometric quality of tests, although it does have some similarities to classical test theory. All three chapters provide in-depth examples of the applications and interpretations, so that readers can have a deeper understanding of these important advanced approaches. Although a full understanding of these advanced approaches requires greater statistical knowledge than is required for most of the book, these approaches are presented here at a level that emphasizes their conceptual basis more than their statistical foundations.

NEW TO THIS EDITION

The fourth edition of this book benefits from a variety of revisions. These revisions reflect, in part, suggestions made by reviewers of the third edition. They also reflect insights into important issues that needed new coverage, greater attention, or better clarity. All revisions and additions are intended to increase the accessibility, scope, and usability of the book, for both students and instructors.

General Changes

Some changes are consistent throughout the book, not being limited to particular chapters. These include the following:

1. Changes were made to increase the clarity and accessibility of the material. A close reading revealed sections, paragraphs, sentences, and words that could be improved for clarity. Thus, material was rewritten and/or reorganized throughout the entire book.

2. In a major revision, this fourth edition includes many new tables and figures intended to help readers crystalize, synthesize, compare, contrast, and assimilate many facets of the book (e.g., key concepts, processes, examples, etc). The third edition of the book included approximately 45 figures, but this new edition more than doubles that amount with 50 *new* figures. In addition, 15 new integrative tables have been added. These figures and tables are mainly pedagogically oriented, hopefully facilitating understanding and insight.

3. In another major change from previous editions, nearly every chapter now includes a technical appendix that will help readers move from psychometric theory (the focus of the chapters) to psychometric action (the focus of the appendices). The appendices demonstrate how to use the R programming language (R Core Team, 2020) to conduct many of the analyses described in the chapters. The main text in this book focuses on the conceptual basis of psychometrics and on providing/illustrating the statistical terms related to those concepts. This material is crucial for truly understanding psychometric theory. However, it leaves readers on their own to figure out how to actually carry out psychometric analyses in practice. The new appendices help readers put theory into practice.

   R was highlighted in these appendices for three important reasons. First, R software is freely available at https://cran.r-project.org/. Thus, while many other statistical software packages are extremely expensive, R is completely free and thus available to anyone who has access to a computer. Second, in both academic work and industry, R has enjoyed expanding popularity while statistical packages such as SPSS and SAS have declined. As this fourth edition is being prepared and published, the field of data science seems to be focused mainly on R and Python (e.g., Bajuk, 2019; Muenchen, 2019; Srivastava, 2020). Programs such as SPSS and SAS barely seem to be in the conversation. For readers interested in data science more generally, an understanding of R's psychometric capabilities may have great value. Finally and crucially, R has many psychometric capabilities that are simply absent in other software packages. For almost every one of the psychometric indexes and procedures that are covered in this book, an R function is available to compute the index or carry out the procedure. Moreover, R's capabilities are constantly expanding. While other statistical software packages have much more limited psychometric capability, R greatly facilitates psychometric work. On the downside, users may initially find R less user-friendly than other statistical software packages such as SPSS and SAS. Nevertheless, the great benefits of R will hopefully outweigh this issue.

   It is worth noting that the new R appendices do assume some familiarity with R. They are not intended to provide a general introduction to R syntax and usage. They focus on getting test data into R, highlighting and loading relevant R packages, using functions relevant to psychometric analysis, and interpreting the output from those analyses. For readers who need a more general introduction to R, there are many online resources available, including tutorial online classes and videos.

The data that are used in the technical appendices are available as part of R, within the syntax itself, or via the SAGE website at **edge.sagepub.com/furr4e**.

4. The references have been expanded and updated. This edition now includes approximately 480 references, up from approximately 400 in the third edition. This provides readers with more original sources that they can turn to for greater depth, more technical discussions, and useful illustrations. Importantly, these additions generally reflect very recent developments or applications of the concepts discussed in the book. In fact, approximately 50 new references reflect work that has appeared since the third edition was published

Chapter-Specific Changes

Of course, there are changes to each individual chapter in the book. Although some chapters were revised more than others, all chapters went through changes that improve their content and style.

Chapter 1 (*Psychometrics and the Importance of Psychological Measurement*): This chapter includes two new figures and one new table. In addition, it now addresses head-on the relatively frequent objection to psychological measurement—that "you cannot reduce someone to a number." Many of us who study behavioral science in general or who specialize in assessment in particular have heard some version of this objection. The revised chapter acknowledges what is valid about this objection, while explaining that psychological measurement does not "reduce someone to a number." Hopefully a direct approach to addressing this criticism will help some readers develop greater comfort with, and understanding of, psychological measurement. This chapter also benefits from polishing and revision for clarity.

Chapter 2 (*Scaling*): This chapter benefits from polishing and from two new figures. In addition, its technical appendix introduces some basic points that will apply to all such appendices in the book. The appendix also demonstrates how to access an R data frame, how to view the data, how to get a sense of the types of variables within a data frame, and how to deal with "factors" (or nominal variables) in R. It introduces and uses a data set (available from SAGE at **edge.sagepub.com/furr4e** that will be used in several other chapters throughout the book.

Chapter 3 (*Differences, Consistency, and the Meaning of Test Scores*): This chapter has benefited from several revisions. First, it now presents and interprets methods for quantifying skew, which is often seen in basic descriptive statistics. Second, it explains scatterplots as a method of visually representing the association between two variables and as a way of introducing covariance and correlation. Third, it notes different types of correlations that are appropriate for different types of data (e.g., polychoric, tetrachoric, Spearman's rho, etc.). Fourth, it includes three new tables and two new figures. Fifth, it has undergone revision for clarity. Sixth, its technical appendix demonstrates how R can be used to compute the statistic discussed in the chapter—mean, variance, standard deviation, skew, covariance, correlation, standard scores, and converted standard scores. In addition, the appendix demonstrates how to plot distributions and scatterplots.

Chapter 4 (*Test Dimensionality and Factor Analysis*): The previous edition included a substantial new section in this chapter, but revisions for this fourth edition were relatively light. Revisions were made for clarity, and a technical appendix was provided. This appendix demonstrates how to use R to conduct an exploratory factors analysis, walking readers through the key steps.

Chapter 5 (*Reliability: Conceptual Basis*): This revised chapter now explicitly notes that reliability is a property of test scores, rather than a property of the test itself. Revisions were made to be consistent with this point throughout the chapter. In addition, several new figures were added to help readers develop a more intuitive and deeper understanding of issues such as the effect of measurement error and the meaning of the four CTT measurement models. Of course, revisions were made throughout for clarity as well. Because this chapter is primarily focused on a conceptual issue (the meaning of reliability), it does not include a technical appendix with R syntax.

Chapter 6 (*Empirical Estimates of Reliability*): Like all other chapters, this revised chapter benefited from polishing for clarity and from the inclusion of several new figures and tables. In addition, its new technical appendix addresses several important practical issues in the estimating of reliability—computing alpha, computing confidence intervals around alpha, and computing split-half estimates of reliability.

Chapter 7 (*The Importance of Reliability*): This chapter includes seven new figures that hopefully help readers develop more intuitive understanding of key concepts and synthesize key points. In addition, the chapter benefits from revisions for clarity. Its new technical appendix demonstrates how to (a) estimate true scores, (b) calculate the standard error of measurement, (c) obtain confidence intervals around scores, (d) correct for attenuation due to measurement error, and (e) obtain item-level information that is extremely useful for enhancing reliability.

Chapter 8 (*Validity*: *Conceptual Basis*): This updated chapter includes a published example of how test developers might enhance the content validity of a test. It also devotes new attention to the importance of discriminant validity, and it describes methods for evaluating discriminant validity (including two quite new and interesting methods). It also includes five new figures and one new table, and it has been polished for clarity throughout. This chapter is primarily conceptually focused, and it does not have an appendix with R syntax.

Chapter 9 (*Estimating and Evaluating Convergent and Discriminant Validity Evidence*): This chapter includes ten new figures and three new tables that should help readers in diverse ways. In addition, it includes an expanded discussion and illustration of Taylor-Russell tables, and it describes a recent new perspective on the interpretation of effect sizes. Its technical appendix demonstrates how to use R to (a) implement the quantifying construct validity procedure, (b) correct a correlation for range restriction, (c) produce a binomial effect size display, (d) conduct Taylor-Russell calculations, (e) conduct sensitivity/specificity analyses, and (f) test the statistical significance of a correlation coefficient.

Chapter 10 (*Response Biases*): This chapter includes revisions for clarity, along with three new figures and two new tables. It has a revised illustration of acquiesce bias, bringing the chapter more in line with the way that it illustrates other biases (and thus hopefully avoiding confusion that might arise otherwise). It also includes a new section describing the randomized response method as a way of minimizing the existence of social desirability bias. This chapter does not include a technical appendix.

Chapter 11 (*Test Bias*): This revised chapter does a better job of (briefly) explaining item characteristic curves as a method for conceptualizing and detecting differential item functioning. In addition, it has an improved presentation of intercept and slope bias. The chapter has been revised throughout for clarity, and it includes two new figures. Its technical appendix illustrates how R can (a) test for group differences in alpha reliability estimates, (b) evaluate group differences in rank-ordering of item difficulties, (c) compare factor loadings across groups (in an exploratory factor analytic context), and (d) use regression to detect predictive test bias.

Chapter 12 (*Confirmatory Factor Analysis*): This chapter includes three new figures to help readers with several issues (e.g., understanding some of the more widely used fit indices), and it has been revised for clarity. It has a substantial technical appendix that addresses several key uses of CFA. First, it demonstrates how to use R to conduct a basic CFA of a test's items. Second, it demonstrates the use of CFA to evaluate the key measurement models in classical test theory (e.g., as a precursor to estimating reliability). Third, it uses CFA to estimate the omega reliability index. Fourth, it uses CFA to test for measurement invariance.

Chapter 13 (*Generalizability Theory*): This chapter benefits from five new figures as well as a new section that presents "a practical, consistency-oriented interpretation of variance components." The previous edition's chapter, as with essentially all discussions of generalizability theory (G theory), provided only a highly theoretical and abstract definition of variance components. In addition, the earlier chapter's discussion of calculating variance components, again as with essentially all discussions of generalizability theory, probably left many readers unclear on the practical meaning and implications of variance components. This new section addresses this issue in a way that hopefully provides readers with a deeper understanding of variance components as the basic building block of G theory. In addition, this chapter's new technical appendix illustrates the use of R to conduct G theory analyses, including a G study phase and a D study phase. It does so for both a one-facet example and a two-facet example, it estimates both relative and absolute generalizability coefficients, and it demonstrates how to plot those coefficients.

Chapter 14 (*Item Response Theory and Rasch Models*): This chapter includes two new integrative tables and two new illustrative figures (e.g., outlining the process through which IRT parameters are estimated). It is revised for clarity throughout. Its technical appendix walks readers through the use of R to conduct an analysis of a Rasch model (or 1PL model), including parameter estimation, model fit, unidimensionality, and plotting of variance item and test curves.

AUTHOR'S ACKNOWLEDGMENTS

I deeply appreciate the help and guidance that have shaped this book over the years. Most important, I am grateful for the love, support, and encouragement of my wife, Sarah. From one of our first dates, during which she told me that she took an online test "to see what the Big Five was all about," she has been a constant source of validation, love, and support.

In moving toward the fourth edition, I relied on the invaluable assistance of several people. Abbie Rickard and Leah Fargotstein, at SAGE, encouraged me to prepare a new edition, oversaw feedback from reviewers who made recommendations for a revision, and helped shape the priorities of the new edition. I appreciate their interest and support for this book and their willingness to help with the logistics of preparing and submitting the revised manuscript. Earlier editions benefited greatly from other people at SAGE, including Reid Hester, Jim Brace-Thompson, Cheri Dellelo, Anna Mesick, Chris Cardone, Astrid Virding, Lisa Shaw, and Sarita Sarak. Earlier editions of this book also benefited from support from both Wake Forest University and Appalachian State University.

I feel enormous gratitude for the mentors, teachers, and colleagues who have had a huge impact on my understanding and interest in psychological measurement,

psychometrics, and psychological research in general. Dr. David Funder, my dissertation adviser and friend, is certainly one of those people, and I'm deeply indebted to him. Other major influences were Dr. Robert Rosenthal, Dr. Douglas Klieger, Dr. Deborah Kendzierski, and Dr. John Nezlek. However, the people who most directly influenced the material in this book are Dr. Dan Ozer, Dr. Steve Reise, and Dr. Keith Widaman. It was their classes in Measurement Theory, Personality Assessment, Regression, Multivariate Statistics, Factor Analysis, and Structural Equation Modeling that really laid the foundation for this book. I'm deeply appreciative of the opportunity to learn from them all. I'm particularly appreciative of Dan for his consistent willingness to share his time and insights with me during my days in graduate school. That said, any errors—egregious or otherwise—in this book are entirely my fault.

Finally, I would like to express my appreciation, respect, and deep affection for Verne Bacharach—my coauthor on the first and second editions of this book. Although Verne is not an author on the recent editions of the book, this book likely would not have happened without his enthusiasm, energy, and initiative. I've truly been fortunate to have Verne as a wonderful coauthor, colleague, and friend.

PUBLISHER'S ACKNOWLEDGMENTS

SAGE gratefully acknowledges the contributions of the following reviewers:

First Edition Reviewers

Rainer Banse, Sozial- und Rechtspsychologie, Institut für Psychologie, Universität Bonn

Patricia L. Busk, University of San Francisco

Kevin D. Crehan, University of Nevada, Las Vegas

Dennis Doverspike, University of Akron

Barbara A. Fritzsche, University of Central Florida

Jeffrey H. Kahn, Illinois State University

Howard B. Lee, California State University, Northridge

Craig Parks, Washington State University

Steven Pulos, University of Northern Colorado

David Routh, Exeter University and University of Bristol

Aloen L. Townsend, Mandel School of Applied Social Sciences, Case Western Reserve University

Vish C. Viswesvaran, Florida International University

Alfred W. Ward, Pace University

Second Edition Reviewers

Barbara Fritzsche, University of Central Florida

Michael R. Kotowski, University of Tennessee, Knoxville

Keith Kraseski, Touro College

Sunny Lu Liu, California State University, Long Beach

Joel T. Nadler, Southern Illinois University, Edwardsville

Patricia Newcomb, University of Texas at Arlington

Third Edition Reviewers

Ismael Diaz, California State University, San Bernardino

W. Holmes Finch, Ball State University

Jemeen W. Horton, Carleton University, Royal Ottawa Mental Health Centre

Karen Machleit, University of Cincinnati

Shlomo Sawilowsky, Wayne State University

Kenneth B. Solberg, Saint Mary's University of Minnesota

Fourth Edition Reviewers

Jacqueline S. Craven, Delta State University

Sarah Flynn, University of the Cumberlands

W. Grant Willis, University of Rhode Island

Sarah Wood, University of Wisconsin, Stout

Kyle Maurice Woosnam, University of Georgia

ABOUT THE AUTHOR

Michael Furr is Professor of Psychology and Wright Faculty Fellow at Wake Forest University, where he teaches and conducts research in personality psychology, psychological measurement, and quantitative methods. He earned a BA from the College of William and Mary, an MS from Villanova University, and a PhD from the University of California at Riverside. He is an editor of the "Statistical Developments and Applications" section of the *Journal of Personality Assessment*, a former associate editor of the *Journal of Research in Personality*, a former executive editor of the *Journal of Social Psychology*, and a consulting editor for several other scholarly journals. He received Wake Forest University's 2012 Award for Excellence in Research, and he won the Society for Personality Assessment's 2017 Bruno Klopfer Award for Distinguished Contributions to the Literature in Personality Assessment. He is a fellow of Divisions 5 (Quantitative and Qualitative Methods) and 8 (Social and Personality Psychology) of the American Psychological Association, a fellow of the Association for Psychological Science, and a fellow of the Society for Personality and Social Psychology.

1

PSYCHOMETRICS AND THE IMPORTANCE OF PSYCHOLOGICAL MEASUREMENT

Your life has probably been shaped, in part, by psychological measurement. Whether you are a student, a teacher, a parent, a psychologist, a physician, a nurse, a patient, a lawyer, a police officer, or a businessperson, you have taken psychological tests, your family members have taken psychological tests, or you have been affected by people who have taken psychological tests. These tests can affect our education, our careers, our family life, our safety, our health, our wealth, and, potentially, our happiness. Indeed, almost every member of an industrialized society is affected by psychological measurement at some point in his or her life—both directly and indirectly.

It is even fair to say that, in extreme situations, psychological measurement can have life or death consequences. Although this might seem overly sensational, far-fetched, and perhaps even simply wrong, it is true. The fact is that in some states and nations, prisoners who have severe cognitive disabilities cannot receive a death penalty. For example, in the state of North Carolina, the General Assembly states that "no defendant with an intellectual disability shall be sentenced to death" (N.C. Gen. Stat. § 15A-2005, 2019); it defines intellectual disability, in part, as general intellectual functioning that is "significantly subaverage." But what is "significantly subaverage" intellectual functioning, and how could we know whether a person's intelligence is indeed significantly subaverage?

These difficult questions are answered in terms of psychological tests. Specifically, the General Assembly states that significantly subaverage intellectual functioning is indicated by a score of 70 or below "on an individually administered, scientifically recognized standardized intelligence quotient test administered by a licensed psychiatrist or psychologist." Put simply, if a person has an intelligence quotient (IQ) score below 70, then they might not be sentenced to death by the state of North Carolina; however, if a person has an IQ score above 70, then they can legally be put to death. Thus, although it might seem hard to believe, intelligence testing can affect whether men and women might live or die, quite literally. Of course, few consequences of psychological measurement are so dramatic, but they can indeed be real, long- lasting, and important.

Given the important role of psychological tests in our lives and in society more generally, those tests must have extremely high quality. If testing has such robust implications, then it should be done with the strongest possible tools and procedures.

This book is about understanding whether such tools and procedures are indeed strong—how to determine whether a test produces scores that are psychologically meaningful and trustworthy. In addition, the principles and concepts discussed in this book are important for creating tests that are psychologically meaningful and trustworthy. These principles and concepts are known as psychometrics.

WHY PSYCHOLOGICAL TESTING MATTERS TO YOU

Considering the potential real-life impact of psychological testing, you need to understand the basic principles of psychological measurement. Whether you wish to be a practitioner of behavioral science, a behavioral researcher, or a sophisticated member of modern society, your life is likely to be affected by psychological measurement.

You might be considering a career involving psychological measurement. Some of you might be considering careers in the practice or application of a behavioral science. Whether you are a clinical psychologist, a school psychologist, a human resources director, a university admissions officer, or a teacher, your work might require you to make decisions on the basis of scores obtained from some kind of psychological test. When a patient responds to a psychopathology assessment, when a student completes a test of cognitive ability or academic aptitude, or when a job applicant fills out a personality inventory, there is an attempt to measure some type of psychological characteristic.

In such cases, test users have a responsibility to examine and interpret important information about the meaning and quality of the tests they use. Without a solid understanding of the basic principles of psychological measurement, test users risk misinterpreting or misusing the information derived from psychological tests. Such misinterpretation or misuse might harm patients, students, clients, employees, and applicants, and it can lead to lawsuits for the test user. Proper test interpretation and use can be extremely valuable for test users and beneficial for test takers.

Some of you might be considering careers in behavioral research. Whether your area is psychology, education, or any other behavioral science, measurement is at the heart of your research process. Whether you conduct experimental research, survey research, or any other kind of quantitative research, measurement is at the heart of your research process. Whether you are interested in differences between individuals, changes in people across time, differences between genders, differences between classrooms, differences between treatment conditions, differences between teachers, or differences between cultures, measurement is at the heart of your research process. If something is not measured or is not measured well, then it cannot be studied with any scientific validity. If your goal is meaningful and accurate interpretation of your research findings, then you must evaluate critically the measurements that you have collected in your research.

As mentioned earlier, even if you do not pursue a career involving psychological measurement, you will almost surely face the consequences of psychological measurement, either directly or indirectly. Applicants to graduate school and various professional schools might be accepted (or not) partially on the basis of tests of knowledge and achievement. Job applicants might be hired (or not) partially on the basis of scores on personality tests. Employees might be promoted (or passed over for promotion) partially on the basis of supervisor ratings of psychological characteristics such as attitude, competence, or collegiality. Parents must cope with the consequences of their children's educational testing. People seeking psychological services might be diagnosed and treated partially on the basis of their responses to various psychological measures.

Even more broadly, our society receives information and recommendations based on research findings. Whether you are (or will be) an applicant, an employee, a parent, a psychological client, or an informed member of society, the more knowledge you have about psychological measurement, the more discriminating a consumer you will be. You will have a better sense of when to accept or believe test scores, when to question the use and interpretation of test scores, and what you need to know to make such important judgments.

Given the widespread use and importance of psychological measurement, it is crucial to understand the properties affecting the quality of such measurements. This book is about the important *attributes of the instruments* that psychologists use to measure psychological attributes and processes.

This book addresses several fundamental questions related to the logic, development, evaluation, and use of psychological measures.

- What does it mean to attribute scores to characteristics such as intelligence, memory, self-esteem, shyness, happiness, or executive functioning?

- How do you know if a particular psychological measure is trustworthy and interpretable?
- How confident should you be when interpreting an individual's score on a particular psychological test?
- What kinds of questions should you ask to evaluate the quality of a psychological test?
- What are some of the different kinds of psychological measures?
- What are some of the challenges to psychological measurement?
- How is the measurement of psychological characteristics similar to and different from the measurement of physical characteristics of objects?
- How should you interpret some of the technical information regarding psychological measurement?

The goal of this book is to address these kinds of questions in a way that provides a deep and intuitive understanding of psychometrics. This book is intended to help you develop the knowledge and skills needed to evaluate psychological tests intelligently. Psychological testing plays an important role in psychological science and in psychological practice, and it plays an increasingly important role in our society.

Hopefully, this book helps you become a more informed consumer and, possibly, producer of psychological information.

OBSERVABLE BEHAVIOR AND UNOBSERVABLE PSYCHOLOGICAL ATTRIBUTES

People use many kinds of instruments to measure observable properties of the physical world. For example, if you want to measure the length of a piece of lumber, then you might use a tape measure. People also use various instruments to measure the properties of the physical world that are not directly observable. For example, clocks are used to measure time, and voltmeters are used to measure the change in voltage between two points in an electric circuit.

Similarly, psychologists, educators, and others use psychological tests as instruments to measure observable events in the physical world. In the behavioral sciences, these observable events are typically some kind of behavior, and behavioral measurement is usually conducted for two purposes. Sometimes, psychologists measure a behavior because they are interested in that specific behavior in its own right. For example, some psychologists have studied the way facial expressions affect the perception of emotions. The Facial Action Coding System (FACS; Ekman & Friesen,

1978) was developed to allow researchers to pinpoint movements of very specific facial muscles. Researchers using the FACS can measure precise "facial behavior" to examine which of a person's facial movements affect other people's perceptions of emotions. In such cases, researchers are interested in the specific facial behaviors themselves; they do not interpret them as signals of some underlying psychological process or characteristics.

Much more commonly, however, behavioral scientists observe human behavior as a way of assessing unobservable psychological attributes such as intelligence, depression, knowledge, aptitude, extroversion, or ability. In such cases, they identify some type of observable behavior that they think represents the particular unobservable psychological attribute, state, or process. They then measure the behavior and try to interpret those measurements in terms of the unobservable psychological characteristics that they think are reflected in the behavior. In most but not all cases, psychologists develop psychological tests as a way to sample the behavior that they think reflects the underlying psychological attribute.

For example, suppose that we wish to identify which of two students, Sam and William, had greater working memory. To do this, we must measure both students' working memories. Unfortunately, there is no known way to observe directly working memory—we cannot directly "see" memory inside a person's head. Therefore, we must look for something that we can see (e.g., some type of behavior) and that could indicate how much working memory someone has. For example, we might ask the students to repeat a series of numbers presented to them rapidly. If the two students differ in their performance on this task, then we might assume that they differ in their working memory. That is, if we observe a difference in their behavior, then we interpret it as revealing a difference in their working memory. If Sam repeats more of the numbers than William, then we might conclude that Sam's working memory is greater than William's. This conclusion requires that we make an inference—that an observable behavior, the number of recalled numbers, is systematically related to an unobservable mental attribute, working memory.

There are several things to notice about this attempt to measure working memory. First, we make an inference from an observable behavior to an unobservable psychological attribute. That is, we assume that the particular behavior that we observe reflects or reveals working memory. If this inference is reasonable, then we would say that our interpretation of the behavior has a degree of *validity.* Although validity is a matter of degree, if the scores from a measure seem to be actually measuring the mental state or mental process that we think they are measuring, then we say that our interpretation of scores on the measure is valid.

Second, for our interpretation of "number recall" scores to be considered valid, the recall task must be theoretically linked to working memory. It would not have made theoretical sense, for example, to measure working memory by timing William's and Sam's running speed in a footrace. In the behavioral sciences, we often make an inference from an observable behavior to an unobservable psychological attribute. Therefore, measurement in psychology often, but not always, involves some type of theory linking a psychological characteristic, process, or state to an observable behavior that is thought to reflect differences in that psychological attribute.

There is a third important feature of our attempt to measure working memory. Working memory is itself a theoretical concept. When measuring working memory, we assume that working memory is more than a figment of our imagination. Psychologists, educators, and other social scientists often use theoretical concepts such as working memory to explain differences in people's behavior. Psychologists refer to these theoretical concepts as *hypothetical* **constructs** or **latent variables**. They are theoretical psychological characteristics, attributes, processes, or states that cannot be directly observed, and they include things such as knowledge, intelligence, self-esteem, attitudes, hunger, memory, personality traits, depression, and attention. The operations or procedures that we use to measure these hypothetical constructs, or for that matter to measure anything, are called **operational definitions**. In our example, the number of recalled numbers was used as an operational definition of some aspect of working memory, which itself is an unobservable hypothetical construct.

You should not be dismayed by the fact that psychologists, educators, and other social scientists rely on unobservable hypothetical constructs to explain human behavior. This reliance is true of many branches of science. Measurement in the physical sciences, as well as the behavioral sciences, often involves making inferences about unobservable events, things, and processes based on observable events. As an example, physicists write about four types of "forces" that exist in the universe: (1) the strong force, (2) the electromagnetic force, (3) the weak force, and (4) gravity. Each of these forces is invisible, but their effects on the behavior of visible events can be seen. For example, objects do not float into space off the surface of our planet. Theoretically, the force of gravity is preventing this from happening. Physicists have built equipment to create opportunities to observe the effects of some of these forces on observable phenomena. In effect, the equipment is used to create scenarios in which to measure observable phenomena that are believed to be caused by the unseen forces.

To be sure, the sciences differ in the number and nature of unobservable characteristics, events, or processes that are of concern to them. Some sciences might rely on relatively few, while others might rely on many. Some sciences might have strong empirical bases for their unobservable constructs (e.g., gravity), while others might have weak empirical bases (e.g., penis envy). Nevertheless, all sciences rely on unobservable constructs to some degree, and they all measure those constructs by measuring some observable events or behaviors.

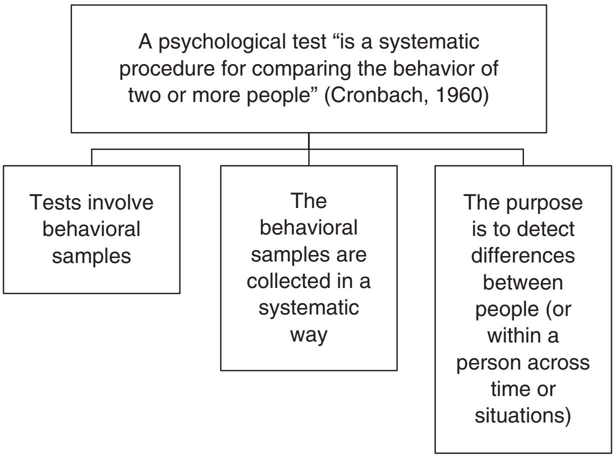
PSYCHOLOGICAL TESTS: DEFINITION AND TYPES

What Is a Psychological Test?

According to Cronbach (1960), a psychological test "is a systematic procedure for comparing the behavior of two or more people" (p. 21). As shown in Figure 1.1, this definition includes three important components: (1) tests involve behavioral samples of some kind, (2) the behavioral samples must be collected in some systematic (i.e., clear and standardized) way, and (3) the purpose of the tests is to detect differences between people. The third component could be modified to include a comparison of performance by the same individuals at different points in time or in different situations, but otherwise the definition is appealing. This appeal is based on several important features.

One appealing feature of the definition is its generality. The idea of a test is sometimes limited to paper-and-pencil tests, but psychological tests can come in many forms. For example, the Beck Depression Inventory–II (BDI-II; Beck et al., 1996) is a fairly traditional 21-item paper-and-pencil test designed to measure depression. People who take the test read each question and then choose an answer from one of several supplied answers. A person's degree of depression is evaluated by counting the number of answers of a certain type that they gave to the questions. The BDI is clearly a test, but other methods of systematically sampling behavior are also tests. For example, in laboratory situations, researchers ask participants to respond in various ways to well-defined stimulus events; participants might be asked to watch for a particular

FIGURE 1.1 ⬢ Cronbach's Definition of a Psychological Test, With Three of Its Key Components Emphasized


A psychological test "is a systematic procedure for comparing the behavior of two or more people" (Cronbach, 1960)
Tests involve behavioral samples
The behavioral samples are collected in a systematic way
The purpose is to detect differences between people (or within a person across time or situations)

visual event and respond by pressing, as quickly as possible, a response key. In other laboratory situations, participants might be asked to make judgments regarding the intensity of stimuli such as sounds. By Cronbach's definition, these are also tests.

The generality of Cronbach's definition also extends to the type of information produced by tests. Some tests produce numbers that represent the amount of some psychological attribute possessed by a person. For example, the U.S. National Assessment of Educational Progress (NAEP; http://nces.ed.gov/nationsreportcard/reading/whatmeasure.aspx) uses statistical procedures to select test items that, at least in theory, produce data that can be interpreted as reflecting the amount of knowledge or skill possessed by children in various academic areas, such as reading. Other tests produce categorical data—people who take the test can be sorted into groups based on their responses to test items. The House-Tree-Person Test (Burns, 1987) is an example of such a test. Children who take this test are asked to draw a house, a tree, and a person. The drawings are evaluated for certain characteristics, and on the basis of these evaluations, children can be sorted into groups (however, this procedure might not be "systematic" in Cronbach's terms). Chapter 2 discusses more about the types data produced by psychological tests.

Another extremely important feature of Cronbach's definition concerns the general purpose of psychological tests. Specifically, tests must be capable of comparing the behavior of different people (*interindividual differences*) or the behavior of the same individuals at different points in time or under different circumstances (*intraindividual differences*). The purpose of measurement in psychology is to identify and, if possible, quantify such interindividual or intraindividual differences. This purpose is a fundamental theme throughout this book, and we will return to it in every chapter. Inter- and intraindividual differences on test performance contribute to test score variability, a necessary component of any attempt to measure any psychological attribute.

Types of Tests

There are tens of thousands of psychological tests in the public domain (Educational Testing Service, 2016). These tests vary from each other along dozens of different dimensions, some of which are reflected in Table 1.1.

TABLE 1.1 ⬢ Some Key Ways in Which Psychological Tests Differ

Differences	Examples
Content	Aptitude, achievement, intelligence, personality, etc.
Response required	Open ended vs. closed ended
Method of administration	Individual vs. group
Use	Criterion refenced vs. norm referenced
Timing	Speeded vs. power
The meaning of "indicators"	Reflective/effect vs. formative/causal

For example, tests can vary in content: There are achievement tests, aptitude tests, intelligence tests, personality tests, attitude surveys, and so on. Tests also vary with regard to the type of response required: There are **open-ended tests**, in which people can answer test questions by saying anything they want in response to the questions on the test, and there are **closed-ended tests**, which require people to answer questions by choosing among alternative answers provided in the test. Tests also vary according to the methods used to administer them. Some are individually administered, in which one person administers the test to one test taker at a time. Other tests can be administered to multiple people all at the same time.

Another major distinction concerns the intended purpose of test scores. Psychological tests are often categorized as either *criterion referenced* (also called domain referenced) or *norm referenced* (Glaser, 1963). **Criterion-referenced tests** are most often seen in settings in which a decision must be made about a person's skill level. In those settings, a cutoff test score is established as a criterion, and it is used to sort people into two groups: (1) those whose performance exceeds the criterion score and (2) those whose performance does not. In contrast, **norm-referenced tests** are usually used to understand how a person compares with other people. This is done by comparing a person's test score with scores from a **reference sample** or **normative sample**. A reference sample is typically a sample of people who complete a test, and the sample is thought to be representative of some broader population of people. Thus, a person's test score can be compared with the scores obtained from the people in the reference sample, telling us, for example, whether the individual has a higher or lower score than the "average person" (and how much higher or lower) in the relevant population. Scores on norm-referenced tests can be valuable when the reference sample is representative of some population, when the relevant population is well defined, and when the person being tested is a member of the relevant population. In principle, none of these issues arise when evaluating a score on a criterion-referenced test.

In practice, the distinction between norm-referenced tests and criterion-referenced tests is often blurred. Criterion-referenced tests are always "normed" in some sense. That is, criterion cutoff scores are not determined at random. The cutoff score will be associated with a decision criterion based on some standard or expected level of performance of people who might take the test. Most of us have taken written driver's license tests. These are criterion-referenced tests because a person taking the test must obtain a score that exceeds some predetermined cutoff. The questions on these tests were selected to ensure that the average person who is qualified to drive has a good chance of answering enough of the questions to pass the test. The distinction between criterion- and norm-referenced tests is further blurred when scores from norm-referenced tests are used as cutoff scores. Institutions of higher education might have minimum SAT or American College Testing (ACT) score requirements for admission or for various types of scholarships. Public schools use cutoff scores from intelligence tests to sort children into groups. In some cases, the use of scores from norm-referenced tests can have life or death consequences, as noted at the beginning of this chapter. Despite the problems with the distinction between criterion-referenced tests and norm-referenced tests, there are slightly different methods used to assess the quality of criterion-referenced and norm-referenced tests (Kane, 1986; Popham & Husek, 1969).

Yet another common distinction is between *speeded tests* and *power tests.* **Speeded tests** are time-limited tests. In general, people who take a speeded test are not expected to complete the entire test in the allotted time. Speeded tests are scored by counting the number of questions answered in the allotted time period. It is assumed that there is a high probability that each question will be answered correctly; each of the questions on a speeded test should be of comparable difficulty. In contrast, **power tests** are not time limited, and test takers are expected to answer all the test questions. Often, power tests are scored also by counting the number of correct answers made on the test. Test items must range in difficulty if scores on these tests are to be used to discriminate among people with regard to the psychological attribute of interest. As is the case with the distinction between criterion-referenced tests and norm-referenced tests, slightly different methods are used to assess the quality of speeded and power tests (Angoff, 1953; Cronbach & Warrington, 1951).

It is worth noting that most of the procedures outlined in this book are relevant mainly for scores based on what are called **reflective (or effect) indicators** (Bollen & Lennox, 1991). For example, scores on intelligence or personality tests are of this kind. A person's responses on an intelligence test are typically seen as being caused by his or her actual level of intelligence. That is, the hypothetical construct (i.e., intelligence) determines, in part, a person's responses to the items on the intelligence test, and these responses are seen as "indicators" of the construct. Such tests are very common in psychology. There are, however, different types of scores that are based on what are called **formative (or causal) indicators**. Socioeconomic status (SES) is the classic example. You could quantify a person's SES by quantitatively combining "indicators" such as her income, education level, and occupational status. In this case, the indicators are not viewed as being "caused" by the person's SES. Instead, the indicators of SES are, in part, exactly what define SES. A full discussion of the distinction between formative/effect and reflective/causal scores—or of the usefulness of the supposed distinction—is beyond the scope of this section (interested readers are directed to Bollen & Diamantopoulos, 2017a, 2017b; Bollen & Lennox, 1991; Diamantopoulos & Winklhofer, 2001; Edwards, 2011; Edwards & Bagozzi, 2000; Hardin, 2017; Howell et al., 2007; MacKenzie et al., 2005; Markus, 2018; Myszkowski et al., 2019; Rhemtulla et al., 2020). The goal here is simply to note the existence of this important distinction and to acknowledge that this book focuses on test scores derived from reflective/effect indicators—as is typical for most tests and measures used in psychology.

A brief note concerning terminology: Several different terms are often used as synonyms for the word *test.* The words *measure*, *instrument*, *scale*, *inventory*, *battery*, *schedule*, and *assessment* have all been used in different contexts and by different authors as synonyms for the word *test*. This book will sometimes refer to tests as instruments and sometimes as measures. The word *battery* will refer to bundled tests, which are tests that are intended to be administered together but are not *necessarily* designed to measure a single psychological attribute. The word *measure* can be used as a verb, as in "The BDI was designed *to measure* depression." It is also often used as a noun, as in "The BDI is a good *measure* of depression." This book will use both forms of the term and rely on the context to clarify its meaning.

WHAT IS PSYCHOMETRICS?

Psychometrics

Just as psychological tests are designed to measure psychological attributes of people (e.g., anxiety, intelligence), **psychometrics** is the science concerned with evaluating the attributes of psychological tests. Three of these attributes will be of particular interest: (1) the type of information (in most cases, scores) generated by the use of psychological tests, (2) the reliability of data from psychological tests, and (3) issues concerning the validity of data obtained from psychological tests. The remaining chapters in this book describe the procedures that psychometricians use to evaluate these attributes of tests. This book addresses the process of testing to a much lesser extent, and it describes particular tests only when illustrating important principles and concepts.

Note that just as psychological attributes of people (e.g., anxiety) are most often conceptualized as hypothetical constructs (i.e., abstract theoretical attributes of the mind), psychological tests also have attributes that are represented by theoretical concepts such as validity or reliability. Just as psychological tests are about theoretical attributes of people, psychometrics is about theoretical attributes of psychological tests. Just as psychological attributes of people are unobservable and must be measured, psychometric attributes of tests are also unobservable and must be estimated. Psychometrics is about the procedures used to estimate and evaluate the attributes of tests.

A Brief History of Psychometrics

The field of psychometrics has been built on two key foundations. One foundation is the practice of psychological testing and measurement. As most textbooks in psychological testing point out (e.g., Dubois, 1970; Miller & Lovler, 2016), the practice of using formal tests (of some kind) to assess individuals' abilities goes back 2,000 or perhaps even 4,000 years in China, as applicants for governmental positions completed various exams. Psychological measurement increased in the 19th century as psychological science emerged and as researchers began systematically measuring various qualities and responses of individuals in experimental studies. The practice of psychological measurement increased even more dramatically in the

20th century, with the development of early intelligence tests and early personality inventories. Over the course of the past 100+ years, the number, kinds, and applications of psychological tests have exploded. With such development comes the desire to create high-quality tests and to evaluate and improve tests. This desire inspired the development of psychometrics as the body of concepts and tools to do this.

A second and related historical foundation is the development of particular statistical concepts and procedures. Starting in the 19th century, scholars began to develop ways of understanding and working with the types of quantitative information that are produced by psychological tests. Among the early pioneers of this work are scholars such as Charles Spearman, Karl Pearson, and Francis Galton, all making key contributions in the late 1800s and early 1900s. Galton in particular is sometimes considered the founding father of modern psychometrics. He had diverse scholarly interests, including—it should be acknowledged—an advocacy for the now-rejected theory of eugenics. However, it is Galton's, Spearman's, and Pearson's important conceptual and technical innovations that are relevant for our discussion. In fact, you might already be familiar with some of these—the standard deviation and the correlation coefficient (see Chapter 3), factor analysis (see Chapters 4 and 12), the use of the normal distribution (or "bell curve"; see Chapter 3) to represent many human characteristics, and the use of sampling for the purpose of identifying and treating measurement error. These crucial statistical concepts and tools were adopted quickly and sometimes developed explicitly in order to make sense out of the numerical information gathered through the use of psychological tests. We will examine such concepts and tools in detail in this book.

Based on the application of these new statistical tools to the evaluation of psychological tests, the field of psychometrics truly came into its own by the 1930s and 1940s. During this period, the journal *Psychometrika* began publication, the Psychometric Society was formed, the American Psychological Association created its "Division of Evaluation and Measurement," and scholars such as J. P. Guilford and L. L. Thurstone published field-defining texts (Jones & Thissen, 2007). By this time, many tenets of what is now known as classical test theory (CTT) had been articulated (see Chapters 5–7)—providing the foundation for the most widely known perspective on test scores and test attributes. Somewhat later (1970s), CTT was expanded into generalizability theory by Lee Cronbach and his colleagues (see Chapter 13). At approximately the same time (or a bit earlier, in the 1950s and 1960s), an alternative to CTT was emerging, leading to what's now known as item response theory (IRT; see Chapter 14). Also in the 1950s, the crucial concept of test validity was undergoing robust development and articulation, with additional important reconceptualizations in the 1990s—leading to the framework addressed in Chapters 8 and 9 (Angoff, 1988).

Over the past few decades, the field of psychometrics has expanded in all of these directions. CTT itself has evolved, as, for example, researchers recognize the limits of commonly used indices of reliability. IRT has enjoyed increased attention as well, with the development of various models and applications. Moreover, as statistical tools such as structural equation modeling have evolved, researchers have discovered ways of using those tools to conceptualize and examine key psychometric concepts.

In sum, psychometrics, as a scientific discipline, is relatively young but has enjoyed a quick evolution and widespread application. From this point on, this book focuses very little on history, devoting attention instead to contemporary concepts, tools, and practices that have grown out of the pioneering work of Galton, Spearman, Pearson, Thurstone, Cronbach, and many others.

CHALLENGES TO MEASUREMENT IN PSYCHOLOGY

We can never be sure that a measurement is perfect. Is your bathroom scale completely accurate? Is the odometer in your car a flawless measure of distance? Is your new tape measure 100% correct? When you visit your physician, is it possible that the nurse's measure of your blood pressure is off a bit? Even the use of highly precise scientific instruments is potentially affected by various errors, not the least of which is human error in reading the instruments. All measurements, and therefore all sciences, are affected by various challenges that can reduce measurement accuracy.

Despite the many similarities among all sciences, measurement in the behavioral sciences has special challenges that do not exist or are greatly reduced in the physical sciences (see Figure 1.2). These challenges affect our confidence in our understanding and interpretation of behavioral observations.

One of these challenges is related to the complexity of psychological phenomena; notions such as intelligence, self-esteem, anxiety, depression, and so on may have many different aspects to them. Thus, one key challenge is to identify and capture the important aspects of these types of human psychological attributes in a single number or score.

FIGURE 1.2 ⬢ Difficult Challenges in Psychological Measurement

Complexity of Concepts	Participant Reactivity	Observer Expectancy/Bias
Composite Scores	Score Sensitivity	(Lack of) Awareness of Psychometrics

You may hear people object to the very idea of psychological assessment on the grounds that, for example, "you can't reduce people to a number" or "you just can't quantify creativity." Indeed, no reasonable psychologist would try to use a single number to represent an individual's unique totality. Given the richness of human psychology and the extraordinary variety of ways in which people differ from each other, no single number or set of numbers would fully represent any individual in some general or holistic sense. We cannot reduce someone's "total psychology" to a single number any more than we can reduce their "total physicality" to a single number.

However, it might indeed be possible to quantify something like creativity, or at least specific aspects or dimensions of creativity. Again, no one seriously attempts to quantify an individual's "total physicality"; however, we do quantify specific physical dimensions such as height, weight, and blood pressure. In a similar way, psychologists and others attempt to quantify specific psychological dimensions such as verbal intelligence, self-esteem (or specific forms of self-esteem), achievement motivation, attentional control, and so on. A key challenge is to make sure that the way in which we quantify such specific psychological dimensions does indeed reflect the complexity of those dimensions adequately. If psychologists can identify specific, coherent dimensions along which people differ, then they may be able to quantify those differences quite precisely. Chapters 4 and 12 address this crucial issue of dimensionality.

Participant reactivity is a second difficult challenge. Because, in most cases, psychologists are measuring psychological characteristics of people who are conscious and generally know that they are being measured, the act of measurement can itself influence the psychological state or process being measured. For example, suppose we design a questionnaire to assess racism. People's responses to the questionnaire might be influenced by their desire not to be thought of as a racist rather than by their true attitudes toward particular ethnic or racial groups. Therefore, people's knowledge that they are being observed or assessed can cause them to react in ways that obscure the meaning of their behavior. This is usually not a problem when measuring features of inanimate objects that do not know they are being measured; the weight of a bunch of grapes is not influenced by the act of weighing them, and black holes do not mind when astrophysicists attempt to measure their size.

Participant reactivity can take many forms. In research situations, some participants may try to figure out the researcher's purpose for a study, changing their behavior to accommodate the researcher (**demand characteristics**). In contrast, in both research and applied measurement situations, some people might become apprehensive, others might change their behavior to try to impress the person doing the measurement (**social desirability**), and still others might even change their behavior to convey a poor impression to the person doing the measurement (*malingering*). In each case, the validity and meaning of the measure is compromised—the person's "true" psychological characteristic is obscured by a temporary motivation or state that is a reaction to the very act of being measured. Chapter 10 discusses this important issue in detail.

Yet another challenge to psychological measurement is that, in the behavioral sciences, the people collecting the behavioral data (observing the behavior, scoring a test, interpreting a verbal response, etc.) can bring their own biases and expectations to their task. Measurement quality is compromised when these factors distort the observations that are made. *Expectation* and *bias* effects can be difficult to detect. In most cases, we can trust that people who collect behavioral data are not consciously cheating; however, even subtle, unintended biases can have effects. For example, a researcher might give intelligence tests to young children as part of a study of a program to improve the cognitive development of the children. The researcher might have a vested interest in certain intelligence test score outcomes, and as a result, they might allow a bias, perhaps even an unconscious one, to influence the testing procedures. **Observer (or scorer) bias** of this type can occur in the physical sciences, but it is less likely to occur because physical scientists rely more heavily than do social scientists on mechanical devices as data collection agents.

The measures used in the behavioral sciences tend to differ from those used by physical scientists in a fourth important respect as well. Psychologists tend to rely on **composite scores** when measuring psychological attributes. Many of the tests used by psychologists involve a series of questions, all of which are intended to measure a specific psychological attribute or process. For example, a personality test might have 10 questions designed to measure extroversion. Similarly, class examinations that are used to measure learning or knowledge generally include many questions.

It is common practice to score each question and then to sum or otherwise combine the items' scores to create a total or composite score. The composite score represents the final measure of the relevant construct—for example, an extroversion score or a "knowledge of algebra" score. Although composite scores do have their benefits (as we will discuss in later chapters, including Chapter 6), several issues complicate their use and evaluation. In contrast, the physical sciences are less likely to rely on composite scores in their measurement procedures (although there are exceptions to this). When measuring a physical feature of the world, such as the length of a piece of lumber, the weight of a molecule, or the speed of a moving object, scientists can usually rely on a single value obtained from a single type of measurement.

A fifth challenge to psychological measurement is **score sensitivity**. **Sensitivity** refers to a measure's ability to discriminate between meaningful amounts of the dimension being measured. For a physical example, consider someone trying to measure the width of a hair with a standard yardstick. Yardstick units are simply too large to be of any use in this situation. Similarly, a psychologist may find that a particular procedure for measuring a psychological attribute or process may not be sensitive enough to discriminate between the real differences that exist in the attribute or process.

For example, imagine a clinical psychologist who wishes to track her clients' emotional changes from one therapeutic session to another. If she chooses a measure that is not sufficiently sensitive to pick up small differences, then she might miss small but important differences in mood. For example, she might ask her clients to complete this very straightforward "measure" after each session:

Check the box below that best describes your general emotional state over the past week:

□ Good

□ Bad

The psychologist might become disheartened by her clients' apparent lack of progress because her clients might rarely, if ever, feel sufficiently happy to checkmark the "Good" box. The key measurement point is that her measure might be masking real improvement by her clients. That is, her clients might be making meaningful improvements—originally feeling extremely anxious and depressed and eventually feeling much less anxious and depressed. However, they might not actually feel good enough to checkmark "good," even though they feel much better than they did at the beginning of therapy. Unfortunately, her scale is too crude or insensitive, in that it allows only two responses and does not distinguish among important levels of "badness" or among levels of "goodness." A more precise and sensitive scale might look like this:

Choose the number that best describes your general emotional state over the past week:

1 2 3	4 5	6 7	8 9
Extremely Good	Somewhat Good	Somewhat Bad	Extremely Bad

A scale of this kind might allow more fine-grained differentiation along the "good versus bad" dimension as compared with the original scale.

For psychologists, the sensitivity problem is exacerbated because we might not anticipate the magnitude of meaningful differences associated with the mental attributes being measured. Although this problem can emerge in the physical sciences, physical scientists are usually aware of it before they do their research. In contrast, social scientists may be unaware of the scale sensitivity issue even after they have collected their measurements.

A final challenge to mention at this point is an apparent lack of awareness of important psychometric information. In the behavioral sciences, particularly in the application of behavioral science, psychological measurement is often a social or cultural activity. Whether it provides information from a client to a therapist regarding psychiatric symptoms, from a student to a teacher regarding the student's level of knowledge, or from a job applicant to a potential employer regarding the applicant's personality traits and skill, applied psychological measurement often is used to facilitate the flow of information among people. Unfortunately, such measurement often seems to be conducted with little or no regard for the psychometric quality of the tests.

For example, most classroom instructors give class examinations. Only on very rare occasions do instructors have any information about the psychometric properties of their examinations. In fact, instructors might not even be able to clearly define the reason for giving the examination. Is the instructor trying to measure knowledge (a latent variable or hypothetical construct), determine which students can answer the most questions, or motivate students to learn relevant information? Some classroom tests might have questionable quality as indicators of differences among students in their knowledge of a particular subject. Even so, the tests might serve the very useful purpose of motivating students to acquire the relevant knowledge.

Although a poorly constructed test might serve a meaningful purpose in some community of people (e.g., motivating students to learn important information), psychometrically well-formed information is better than information that is not well formed. Furthermore, if a test or measure is intended to reflect the psychological differences among people, then the test must have strong psychometric properties. Knowledge of these properties should inform the development or selection of a test—all else being equal, test users should use psychometrically sound instruments.

In sum, this survey of challenges should indicate that although measurement in the behavioral sciences and measurement in the physical sciences have much in common, there are important differences. These differences should always inform our understanding of data collected from psychological measures. For example, we should be aware that participant reactivity can affect responses to psychological tests.

At the same time, it is important to emphasize that behavioral scientists have significant understanding of these challenges and that they have generated effective methods of minimizing, detecting, and accounting for various problems. Similarly, behavioral scientists have developed methods that reduce the potential impact of experimenter bias in the measurement process. This book covers many of the extensive methods that psychometricians have developed to handle the challenges associated with the development, evaluation, and process of measurement of psychological attributes and behavioral characteristics.

THE IMPORTANCE OF INDIVIDUAL DIFFERENCES

The ability to identify and characterize psychological differences is at the heart of all psychological measurement, and it is the foundation of all methods used to evaluate tests. Indeed, the purpose of measurement in psychology is to identify and quantify the psychological differences that exist between people, over time, or across conditions. These psychological differences contribute to differences in test scores and are the basis of all psychometric information. Even when a practicing psychologist, educator, or consultant makes a decision about a single person based on that person's test score, the meaning and quality of the person's score can be understood only in the context of the test's ability to detect differences among people.

All measures in psychology require that we obtain behavioral samples of some kind. Behavioral samples might include scores on a paper-and-pencil test, written or oral responses to questions, or records based on behavioral observations. Useful psychometric information can be obtained only if people differ with respect to the behavior that is sampled. If a behavioral sampling procedure produces scores that differ between people (or that differ across time or condition), then the psychometric properties of those scores can be assessed. This book presents the logic and analytic procedures associated with these psychometric properties.

If we think that a particular test is a measure of a particular psychological attribute, then we must be able to argue that differences in the test scores are related to differences in the relevant underlying psychological attribute. For example, a psychologist might be interested in measuring visual attention. Because visual attention is an unobservable hypothetical construct, the psychologist must create a behavioral sampling procedure or test that reflects individual differences in visual attention. However, before firmly concluding that the procedure is indeed interpretable as a measure of visual attention, the psychologist must accumulate evidence that there is an association between individuals' scores on the test and their "true" levels of visual attention. The process by which the psychologist accumulates this evidence is called the validation process; it will be examined in later chapters.

The following chapters show how individual differences are quantified and how their quantification is the first step in solving many of the challenges to measurement. Individual differences represent the currency of psychometric analysis—they provide the data for psychometric analyses of tests.

BUT PSYCHOMETRICS GOES WELL BEYOND "DIFFERENTIAL" PSYCHOLOGY

Although the previous section highlights the fact that measurement is based on the existence and detection of psychological differences among people, it is important to avoid a common misunderstanding. The misunderstanding is that psychometrics, or even a general concern about psychological measurement, is relevant only to those psychologists who study a certain set of phenomena that are sometimes called "individual difference" variables.

It may be true that psychometrics evolved largely in the context of certain areas of research, such as intelligence testing, that would be considered part of "differential" psychology. Indeed, while many early pioneers in psychology pursued general laws or principles of mental phenomena that apply to all people, Galton, Spearman, and others focused on the variability of human characteristics. For example, Galton was primarily interested in the ways in which people differ from each other—some people are taller than others, some are smarter than others, some are more attractive than others, and some are more aggressive than others. He was interested in understanding the magnitude of those types of differences, the causes of such differences, and the consequences of such differences.

Thus, the approach to psychology that was taken by Galton, Spearman, and others became known as **differential psychology**, the study of individual differences. There is no hard-and-fast definition or classification of what constitutes differential psychology, but it is often seen to include intelligence, aptitude, and personality. This is usually seen as contrasting with experimental psychology, which focused mainly on the average person instead of the differences among people.

Perhaps because Galton is closely associated with both psychometrics and differential psychology, people sometimes view psychometrics as an issue that concerns only those who study "individual differences" topics such as intelligence, ability/aptitude, or personality. Some seem to believe that psychometrics is not a concern for those who take a more experimental approach to human behavior. This belief is incorrect.

Psychometric issues are by no means limited to so-called differential psychology. Rather, all psychologists, whatever their specific area of research or practice, must be concerned with measuring behavior and psychological attributes. Therefore, they should all understand the problems associated with measuring behavior and psychological attributes, and these problems are the subject matter of psychometrics.

Regardless of one's specific interest, all behavioral sciences and all applications of the behavioral sciences depend on the ability to identify and quantify variability in human behavior. The book will revisit this issue later in depth, with specific examples and principles underscoring the wide relevance of psychometric concepts. Psychometrics is the study of the operations and procedures used to measure variability in behavior and to connect those measurements to psychological phenomena.

Suggested Readings

For a history of early developments in psychological testing:

DuBois, P. H. (1970). *A history of psychological testing*. Allyn & Bacon.

For a history more focused on psychometrics specifically:

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & Sinharay (Eds.), *Handbook of statistics, 26: Psychometrics* (pp. 1–27). North Holland.

For a modern historical and philosophical treatment of the history of measurement in psychology:

Michell, J. (2003). Epistemology of measurement: The relevance of its history for quantification in the social sciences. *Social Science Information*, *42*(4), 515–534. https://doi.org/10.1177/0539018403424004

For an overview of contemporary tests and issues in psychological testing:

Miller, L. A., & Lovler, R. L. (2016). *Foundations of psychological testing: A practical approach* (5th ed.). SAGE.

BASIC CONCEPTS IN MEASUREMENT

PART I

2

SCALING

If something exists, it must exist in some amount (Thorndike, 1918). Psychologists generally believe that people have psychological attributes, such as thoughts, feelings, emotions, personality characteristics, intelligence, learning styles, and so on. If we believe this, then we must assume that each psychological attribute exists in some quantity. With this in mind, psychological measurement can be seen as a process through which numbers are assigned to represent the quantities of psychological attributes. The measurement process succeeds if the numbers assigned to an attribute reflect the actual amounts of that attribute.

The standard definition of measurement (borrowed from Stevens, 1946) found in most introductory test and measurement texts goes something like this: "Measurement is the assignment of numerals to objects or events according to rules." In the case of psychology, education, and other behavioral sciences, the "events" of interest are generally samples of individuals' behaviors. The "rules" mentioned in this definition usually refer to the scales of measurement proposed by Stevens (1946).

This chapter is about **scaling**, which concerns the way numerical values are assigned to psychological attributes. Scaling is a fundamental issue in measurement, and it involves a variety of considerations. This chapter discusses the meaning of numerals, the way in which numerals can be used to represent psychological attributes, and the problems associated with trying to connect psychological attributes with numerals. As discussed in the previous chapter, psychological tests are intended to measure unobservable psychological characteristics such as attitudes, personality traits, and intelligence. Such characteristics present special problems for measurement, and this chapter discusses several possible solutions for these problems.

These issues might not elicit cheers of excitement and enthusiasm among some readers or perhaps among most readers (or perhaps in any reader?); however, these issues are fundamental to psychological measurement, to measurement in general, and to the pursuit and application of science. More specifically, they are important because they help define scales of measurement. That is, they help differentiate the ways in which psychologists apply numerical values in psychological measurement. In turn, these differences have important implications for the use and interpretation of scores from psychological tests. The way scientists and practitioners use and make sense out of tests depends heavily on the scales of measurement being used. Your attention to the material in this chapter should be rewarded with new insights into the foundations of psychological measurement and even into the nature of numbers.

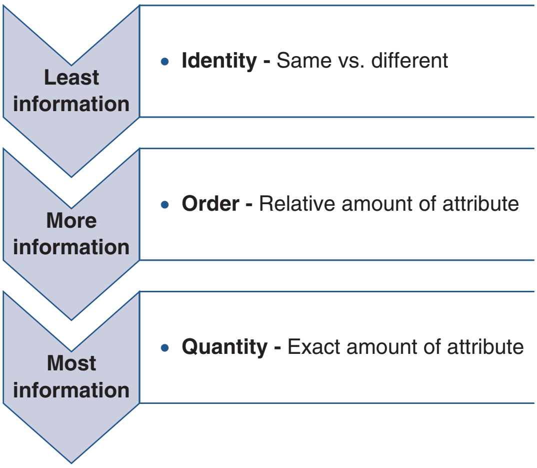
FUNDAMENTAL ISSUES WITH NUMBERS

In psychological measurement, numerals are used to represent an individual's level of a psychological attribute. For example, your numerical score on an IQ test is used to represent your level of intelligence, your numerical score on the Rosenberg Self-Esteem Inventory is used to represent your level of self-esteem, and a numerical value can even be used to represent your biological sex (e.g., males might be referred to as "Group 0" and females as "Group 1"). Thus, psychological measurement is heavily oriented toward numbers and quantification.

Importantly, numerals can represent psychological attributes in different ways, depending on the nature of the numeral that is used to represent an attribute. This section describes important properties of numerals, and it shows how these properties influence the ways in which numerals represent psychological attributes.

As shown in Figure 2.1. this section outlines three important numerical properties, and it discusses the meaning of zero. In essence, the numerical properties of

FIGURE 2.1 ⬢ Properties of Numbers


Least information
• Identity - Same vs. different
More information
• Order - Relative amount of attribute
Most information
• Quantity - Exact amount of attribute

identity, order, and quantity reflect the ways in which numerals represent potential differences in psychological attributes. Furthermore, zero is an interestingly complex number, and this complexity has implications for the meaning of different kinds of test scores. A "score" of zero can have extremely different meanings in different measurement contexts.

The Property of Identity

The most fundamental form of measurement is the ability to reflect "sameness versus differentness." Indeed, the simplest psychological measurements are those that differentiate between categories or groups of people.

For example, you might ask first-grade teachers to identify those children in their classrooms who have behavior problems. The children who are classified as having behavior problems should be *similar to* each other with respect to their behavior. In addition, those children should be *different from* the children who are classified as not having behavioral problems. That is, the individuals within a category should be the same as each other in terms of sharing a psychological feature, but they should be different from the individuals in another category. In psychology, this requires that we sort people into at least two categories. The idea is that objects, events, or people can be sorted into categories that are based on similarity of features. In many cases, these features are behavioral characteristics reflecting psychological attributes, such as happy or sad, introverted or extroverted, and so on.

Certain rules must be followed when sorting people into categories. The first and most straightforward rule is that, to establish a category, the people within a category must satisfy the property of **identity**. That is, all people within a particular category must be "identical" with respect to the feature reflected by the category. For example, everyone in the "behavioral problem" group must, in fact, have behavioral problems, and everyone in the "no behavioral problem" group must not have behavioral problems. Second, the categories must be *mutually exclusive.* If a person is classified as having a behavioral problem, then they cannot simultaneously be classified as not having a behavioral problem. Third, the categories must be *exhaustive.* If you think that all first-graders can be classified as either having behavioral problems or not having behavioral problems, then these categories would be exhaustive. If, on the other hand, you can imagine someone who cannot be so easily classified, then you would need another category to capture that person's behavior. To summarize the second and third rules, each person should fall into one and only one category.

When numerals have only the property of identity, they represent sameness vs. differentness, and they serve simply as labels of categories. The categories could be labeled with letters, names, or numerals. You could label the category of children with behavior problems as "Behavior Problem Children," you could refer to the category as "Category B," or you could assign a numeral to the category. For example, you could label the group as "0," "1," or "100."

When having only the property of identity, numerals are generally not thought of as having true mathematical value. For example, if "1" is used to reflect the category of children with behavioral problems and "2" is used to represent the category of children without behavioral problems, then we would not interpret the apparent 1-point difference between the numerical labels as having any form of quantitative significance.

The latter point deserves some additional comment. When making categorical differentiations between people, the distinctions between categories represent differences in kind or quality rather than differences in amount. Again returning to the teachers' classifications of children, the difference between the two groups is a difference between *types* of children—those children who have behavioral problems and those who do not. In this example, the classification is not intended to represent the amount of problems (e.g., a lot vs. a little) but rather the presence or absence of problems. In this way, the classification is intended to represent two qualitatively distinct groups of children.

Of course, you might object that this is a rather crude and imprecise way of measuring or representing behavioral problems. You might suggest that such an attribute is more accurately reflected in some degree, level, or amount than in a simple presence/absence categorization. This leads to additional properties of numerals.

The Property of Order

Although identity is the most fundamental property of a numeral, the property of order conveys more information. As discussed above, when numerals have only the property of identity, they convey information about whether two individuals are similar or different but nothing more. In contrast, when numerals have the property of **order**, they convey information about the relative amount of an attribute that people possess.

When numerals have the property of order, they indicate the rank order of people relative to each other along some dimension. In this case, the numeral 1 might be assigned to a person because they possess more of an attribute than anyone else in the group. The numeral 2 might be assigned to the person with the next greatest amount of the attribute, and so on.

For example, teachers might be asked to rank children in their classrooms according to the children's interest in learning. Teachers might be instructed to assign the numeral 1 to the child who shows the most interest in learning and 2 to the child whose interest in learning is greater than all the other children except the first child, continuing in this way until all the children have been ranked according to their interest in learning.

When numerals are used to indicate order, they again serve essentially as labels. For example, the numeral 1 indicated a person who had more of an attribute than anyone else in the group. The child with the greatest interest in learning was assigned the numeral 1 as a label indicating the child's rank. In fact, we could just as easily assign letters as numerals to indicate the children's ranks. The child with the most (least) interest in learning might have been assigned the letter A to indicate his or her rank. Each person in a group of people receives a numeral (or letter) indicating that person's relative standing within the group with respect to some attribute. For communication purposes, it is essential that the meaning of the symbol used to indicate rank be clearly defined. We simply need to know what 1, or A, means in each context.

Although the property of order conveys more information than the property of identity, it is still quite limited. While it tells us the relative amount of differences between people, it does not tell us about the actual degree of differences in that attribute. For example, based on ordinal information, we might know that the child ranked 1 has more interest in learning than the child ranked 2, but we do not know *how much* more interest they have. The two children could differ only slightly in their amount of interest in learning, or they could differ dramatically. In this way, when numerals have the property of order, they are still a rather imprecise way of representing psychological differences.

The Property of Quantity

Although the property of order conveys more information than the property of identity, the property of quantity conveys even greater information. As noted above, numerals that have the property of order convey information about which of two individuals has a higher level of a psychological attribute, but they convey no information about the exact amounts of that attribute. In contrast, when numerals have the property of **quantity**, they provide information about the magnitude of differences between people.

At this level, numerals reflect *real numbers* or, for our purposes, numbers. The number 1 is used to define the size of the basic *unit* on any particular scale. All other values on the scale are multiples of 1 or fractions of 1. Each numeral (e.g., the numeral 4) represents a count of basic units.

Think about a thermometer that you might use to measure temperature. To describe how warm the weather is, your thermometer reflects temperature in terms of "number of degrees" (above or below 0). The degree is the unit of measurement, and temperature is represented in terms of this unit.

Units of measurement are standardized quantities; the size of a unit will be determined by some convention. For example, 1 degree Celsius (1°C) is defined (originally) in terms of 1/100th of the difference between the temperature at which ice melts and the temperature at which water boils. We will revisit this important point shortly.

Real numbers are also said to be continuous. In principle, any real number can be divided into infinitely small parts. In the context of measurement, real numbers are often referred to as *scalar*, *metric*, or *cardinal*, or sometimes simply as *quantitative* values.

The power of real numbers derives from the fact that they can be used to measure the amount or quantity of an attribute of a thing, person, or event. When applied to an attribute in an appropriate way, a real number indicates the amount of something. For example, a day that has a temperature of 50°C is not simply warmer than a day that has a temperature of 40°C; it is precisely 10 units (i.e., degrees) warmer.

When psychologists use psychological tests to measure psychological attributes, they often assume that the test scores have the property of quantity. As we will see later, this often might not be a reasonable assumption.

The Number 0

The number 0 is a strange number (see Seife, 2000), with at least two potential meanings. To properly interpret a score of 0 in any particular situation, you must understand which meaning is relevant in that situation.

In one possible meaning, zero reflects a state in which an attribute of an object or event has no existence. If you said that an object was 0.0 cm long, you would be claiming that the object has no length, at least in any ordinary sense of the term *length.* Zero in this context is referred to as **absolute zero**. In psychology, the best example of a behavioral measure with an absolute 0 point might be reaction time.

The second possible meaning of zero is to view it as an arbitrary quantity of an attribute. A zero of this type is called a relative or **arbitrary zero**. In the physical world, attributes such as time (e.g., calendar, clock) and temperature measured by standard thermometers are examples. In these examples, 0 is simply an arbitrary point on a scale used to measure that feature. For example, a temperature of 0 on the Celsius scale represents the melting point of ice, but it does not represent the "absence" of anything (i.e., it does not represent the absence of temperature or of warmth).

The psychological world is filled, at least potentially, with attributes having a relative 0 point. For example, it is difficult to think that conscious people could truly have no (zero) intelligence, self-esteem, introversion, social skills, attitudes, and so on. Although we might informally say that someone "has no social skill," psychologists would not suggest this formally—indeed, we actually believe that everyone has some level of social skill (and self-esteem, etc.), although some people might have much lower levels than other people.

Despite the fact that most psychological attributes do not have an absolute 0 point, psychological tests of such attributes could produce a score of 0. In such cases, the zero would be considered arbitrary, not truly reflecting an absence of the attribute. Furthermore, you will see that many if not most psychological test scores can be expressed as a type of score called a *z* score, which will be discussed in Chapter 3.

A *z* score of 0 indicates an average score within the set of score. In this case, zero represents an arbitrary or relative zero.

In psychology, there can be a problem in determining whether a test score of zero should be thought of as relative or absolute. The problem concerns the distinction between the test being used to measure a psychological attribute and that psychological attribute itself.

Consider an example that Thorndike (2005) used to illustrate this problem. Thorndike describes a scenario in which a sixth-grade child takes a spelling test and fails to spell any of the words correctly. The child thus receives a score of 0 on the test. In this case, the spelling test is the instrument used to measure an attribute of the child—the child's spelling ability. The test itself has an absolute 0 point, indicating that the child failed to spell any words correctly. That is, the test score of 0 indicates an absence of correctly spelled words. It is difficult, however, to imagine that a sixth-grade child is incapable of spelling; the child's *spelling ability* is probably not zero. The question then becomes how we are going to treat the child's test score. Should we consider it an absolute zero or a relative zero?

This is important because the type of zero associated with a test affects how we interpret and use the test scores. For example, we might plan to conduct statistical analyses on test scores for a research study. Importantly, the types of analyses that we can legitimately conduct are determined, in part, by the type of zero that is reflected in the test scores. On one hand, if we can assume that a test has an absolute zero, then we can feel comfortable performing the arithmetic operations of multiplication and division on the test scores. On the other hand, if a test has a relative 0 point, then we should restrict arithmetic operations on the scores to addition and subtraction. As a matter of evaluation, it is important to know what zero means—does it mean that a person who scored 0 on a test had none of the attribute that was being measured, or does it mean that the person might not have had a measurable amount of the attribute, at least not measurable with respect to the particular test you used to measure the attribute?

In sum, the three properties of numerals and the meaning of zero are fundamental issues that shape our understanding of psychological test scores. If two people share a psychological feature, then we have established the property of identity. If two people share a common attribute but one person has more of that attribute than the other, then we can establish order. If order can be established and if we can determine *how much* more of the attribute one person has compared with others, then we have established the property of quantity. Put another way, identity is the most fundamental level of measurement. To measure anything, the identity of the thing must be established. Once the identity of an attribute is known, it might be possible to establish order. Furthermore, order is a fundamental characteristic of quantity. As we will see, numbers play a different role in representing psychological attributes depending on their level of measurement.

Most psychological tests are treated as if they provide numerical scores that possess the property of quantity. The next two sections discuss key issues regarding the meaning and use of such quantitative test scores. Specifically, they discuss the meaning of a "unit of measurement," the issues involved with counting those units, and the implications of those counts.

UNITS OF MEASUREMENT

The property of quantity requires that units of measurement be clearly defined. As discussed in the next section, quantitative measurement depends on our ability to count these units. Before discussing the process and implications of counting the units of measurement, we must clarify what is meant by a unit of measurement.

In many everyday cases of physical measurement, the units of measurement are familiar. When measuring the length of a piece of lumber, the width of a couch, or the height of their children, people typically use a tape or ruler marked off in units of inches or centimeters. Length, width, and height are measured by counting the number of these units from one end of the lumber, couch, or child to the other end.

In contrast, in many cases of psychological measurement, units of measurement are often less obvious. When measuring a psychological characteristic such as shyness, working memory, attention, or intelligence, what are the units of measurement? Presumably, they are responses of some kind, perhaps to a series of questions or items. But how do we know whether, or to what extent, those responses are related to the psychological attributes themselves? This book returns to these questions at a later time, as they represent the most vexing problems in psychometrics. At this point, let's focus on the notion of a unit of measurement. This can be illustrated in the context of the measurement of the length of physical objects (Michell, 1990).

Imagine that you are building a bookshelf and you need to measure the length of pieces of wood. Unfortunately, you cannot find a tape measure, a yardstick, or a ruler of any kind—how can you precisely quantify the lengths of your various pieces of wood?

One solution is to create your own unique measurement system. First, imagine that you happen to find a long wooden curtain rod left over from a previous project. You cut a small piece of the curtain rod; let us call this piece an "xrod" (see Figure 2.2). Because your pieces of bookshelf wood are longer than the xrod, you will need several xrods. Therefore, you can use this original xrod as a template to produce a collection of identical xrods. That is, you can cut additional xrods from the curtain rod, making sure that each xrod is the same exact length as your original xrod. You can now use your xrods to measure the length of all your pieces of wood. For example, to measure the length of one of your shelves, place one of the xrods at one end of the piece of wood that you will use as a shelf. Next, as shown in