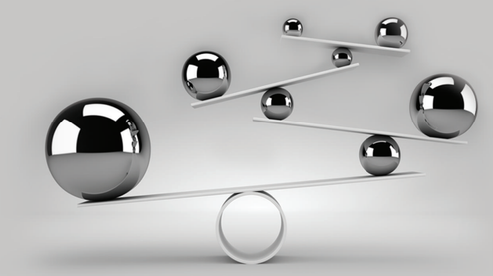


Third Edition

PSYCHOMETRICS

Using a meaning-based approach that emphasizes the "why" over the "how to," **Psychometrics: An Introduction** provides thorough coverage of fundamental issues in psychological measurement. Author R. Michael Furr discusses traditional psychometric perspectives and issues including reliability, validity, dimensionality, test bias, and response bias, as well as advanced procedures and perspectives such as item response theory and generalizability theory. The updated **Third Edition** includes broader and more in-depth coverage with new references, a glossary summarizing more than 200 key terms, and expanded suggested readings consisting of highly relevant papers to enhance the book's overall accessibility, scope, and usability for both instructors and students.

New and Key Features

- **Expanded depth and breadth of coverage** of key issues in psychometrics, including summaries of relevant statistical packages, introduces readers to a wide range of important concepts, principles, and procedures.
- **Updated and expanded references** allow readers to review original sources underlying psychometrics and access the latest developments in the literature.
- **Accompanying PowerPoint® slides** are available to instructors for support in the classroom.
- **Integration of statistics** with a discussion of their use as tools to solve particular psychometric problems encourages a more complete understanding of both.

Cover image: ©Shutterstock.com/Leigh Prather

ISBN 978-1-5063-3986-3

SAGE www.sagepublishing.com
Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

Furr PSYCHOMETRICS Third Edition

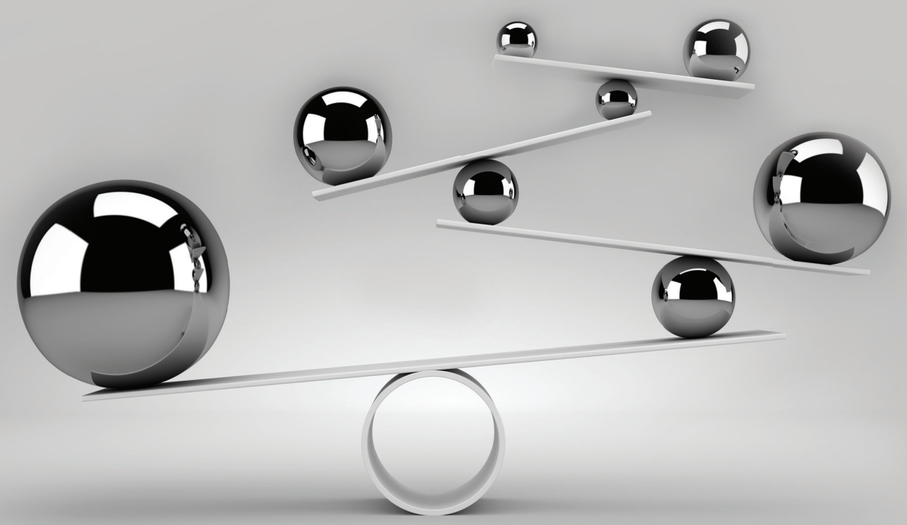
SAGE

Third Edition

PSYCHOMETRICS

An Introduction

R. Michael Furr



Psychometrics

Third 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 800 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

Third Edition

R. Michael Furr
Wake Forest University

SAGE

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

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.
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Printed in the United States of America

Names: Furr, R. Michael, author.

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

Description: Third edition. | Thousand Oaks, California : SAGE, [2018] | Includes bibliographical references and index.

Identifiers: LCCN 2017039902 | ISBN 9781506339863 (hardcover : alk. paper)

Subjects: LCSH: Psychometrics.

Classification: LCC BF39 .F87 2018 | DDC 150.1/5195—dc23 LC record available at https://lccn.loc.gov/2017039902

Acquisitions Editor: Abbie Rickard
Editorial Assistant: Jennifer Cline
Production Editor: Veronica Stapleton Hooper
Copy Editor: Gillian Dickens
Typesetter: Hurix Systems Pvt. Ltd.
Proofreader: Scott Oney
Indexer: Michael Ferreira
Cover Designer: Anupama Krishnan
Marketing Manager: Katherine Hepburn

This book is printed on acid-free paper.

17 18 19 20 21 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 attempt to understand or predict 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 rests on successful measurement at a level that is no less important than it is in research. 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 measures, 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. More specifically, our goal is to make these important issues as accessible and as clear as possible, to as many readers as possible—including people who might initially shy away from something that often might be seen as "obscure, esoteric, and cerebral."

With these goals in mind, our coverage of psychometrics is intended to be deep but intuitive and relatively nontechnical. We believe that this is a novel approach. On one hand, our 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. We believe that anyone familiar with basic algebra and something equivalent to an undergraduate course in statistics will be comfortable with most of the material in this book. In general, our hope is that readers will 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 of value to those who need a solid foundation in the basic concepts and logic of psychometrics or measurement more generally. Although it 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.

In our attempt to make the topics of psychometrics accessible to our target audience, we constructed 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, we have departed from the usual practice of having a separate chapter devoted to statistics. Instead, we introduce statistical concepts throughout the text as needed, and we present them as tools to help solve particular psychometric problems. For example, we discuss factor analysis initially in the context of exploring the dimensionality of a test. Thus, we tie the statistical procedures to a set of important and intuitive conceptual issues. Our experience as classroom instructors has taught us 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. In the first chapter, we address the basic importance of psychological measurement and psychometrics. In addition, we examine a few 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.

In Chapters 2 through 4, we 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, our intention is to focus these chapters at a level that is relatively intuitive and conceptual.

In Chapters 5 through 7, we examine the psychometric concept of reliability. In these chapters, we differentiate three fundamental aspects of reliability. In Chapter 5, we introduce the conceptual basis of reliability, focusing on the perspective of classical test theory. In Chapter 6, we discuss and evaluate the common methods of estimating and evaluating the reliability of test scores. In Chapter 7, we explore the importance of reliability in terms of applied testing, scientific research, and test development. We believe that differentiating these three aspects of reliability provides readers with an understanding of reliability that is clearer and deeper than what might be obtained from many existing treatments of the topic. In all these chapters, we emphasize the psychological meaning of the concepts and procedures. We hope that this maximizes readers' ability to interpret reliability information meaningfully.

In Chapters 8 and 9, we examine the psychometric concept of validity. In these chapters, we 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. In these chapters, we adopt the most 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 we discuss the traditional "tripartite" model of validity (i.e., content validity, criterion validity, and construct validity), which is emphasized in most existing measurement-oriented texts, our core discussion represents a more modern view of test validity and the evidence relevant to evaluating test validity.

In Chapters 10 and 11, we discuss two important threats to the psychometric quality of tests. We believe that it is vital to acknowledge and understand the challenges faced by those who develop, administer, and interpret psychological tests. Furthermore, we believe that 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. In Chapter 10, we explore response biases, which obscure the true differences among individuals taking psychological tests. In this chapter (which is unique to this book), we describe several different types of biases, we demonstrate their deleterious effects on psychological measurement, and we examine some methods of preventing or minimizing these effects. In Chapter 11, we examine test bias, which obscures the true differences between groups of people. In

this chapter, we describe 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, in Chapters 12 to 14, we present advanced contemporary approaches to psychometrics. Much of the book reflects the most common psychometric approach in behavioral research and application—classical test theory. In the final three chapters, we provide overviews of approaches that move beyond this traditional approach. In Chapter 12, we present 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. In Chapter 13, we discuss the basic concepts and purpose of generalizability theory, which can be seen as an expansion of the more traditional approaches to psychometric theory. In Chapter 14, we discuss item response theory (IRT) (aka 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. In all three chapters, we 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, our goal is to present these approaches at a level that emphasizes their conceptual basis more than their statistical foundations.

New to This Edition

The third edition of this book benefits from a variety of revisions. These revisions reflect, in part, suggestions made by reviewers of the second edition. They also reflect my views of the important issues that needed new coverage, greater attention, or better clarity. Beyond the book itself, this edition is accompanied by additional resources for instructors, including a set of PowerPoint slides and a test bank. 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. I have thought extensively about all the material in the book, searching for opportunities to make several types of general changes.

1. Many changes were made to increase the clarity and accessibility of the material. I identified sections, paragraphs, sentences, and words that, I felt, could be improved for clarity, and I rewrote and/or reorganized this material throughout the entire book.
2. The breadth and depth of coverage was enhanced significantly. Sometimes this breadth and depth was provided by adding a sentence or two, sometimes

it was a new paragraph, and in many cases it was an entirely new section. Through the 14 core chapters of the book, coverage has increased by more than 20% (in terms of word count). The result of these additions is a more thorough treatment of psychometrics—both in terms of the topics covered and in terms of the depth with which many important topics are covered.

3. A great deal of time was spent identifying and integrating recent literature, in order to present the most recent important and illustrative work in the field. Due to this, the average publication date of the references in the third edition is 11 years more recent than for the second edition. Moreover, approximately 60% of the references in this edition are from 2000 and later, whereas 65% of the references in the previous edition were from *before* 2000. Finally, fully 85 of the new references in this edition were published *after* the very latest reference in the previous edition (2012). Thus, this new edition better represents—by far—many of the most recent developments in the field.
4. Related to these first few changes, the references were expanded significantly with relevant literature. Specifically, this edition has expanded to approximately 400 references, from only about 220 references in the previous edition. This expansion by nearly 80% provides readers with dramatically more original sources that they can turn to for greater depth, more technical discussions, and useful illustrations. Importantly, as just noted, these additions generally reflect very recent developments or applications of the concepts discussed in the book.
5. The "Suggested Readings" at the end of each chapter have been expanded as well, with the goal of helping students and teachers identify some of the most important, interesting, or illustrative sources related to key concepts in each chapter. Overall, the number of suggested readings has approximately doubled.
6. The clarity of connections among the chapters was enhanced throughout the book. Many chapters now include a greater number of explicit references to other chapters, when discussing concepts that led to (or built on) principles or concepts that appeared in those other chapters. These changes were intended to help readers move back and forth more easily in the book, allowing them to remind themselves of important points when building on those points.
7. A glossary has been added to assist readers with identifying and understanding key terms. The new glossary provides useful definitions of nearly 250 such terms.

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 has benefited from three key revisions. First, it now includes a

discussion of the difference between scores based upon effect (reflective) indicators and causal (formative) indicators—a distinction that was completely omitted from previous editions of the book. Second, a new section called "A Brief History of Psychometrics" describes the conceptual roots of the field in a way that is more inclusive and conceptually informative than what was available in the previous editions (which focused almost exclusively on Francis Galton's contributions). Third, it expands and clarifies the important point that psychometrics is—or should be—a concern for all areas of behavioral science, not just for "differential psychology."

Chapter 2 (Scaling): The main revision to this chapter is a deepened discussion/illustration of the implication that scaling has for the meaningfulness of particular types of descriptive statistics. Other changes to this chapter are mostly minor revision to enhance clarity.

Chapter 3 (Individual Differences and Correlations): The previous edition benefited from significant revisions to this chapter, but revisions for the current edition were relatively light—mostly focusing on enhancing clarity.

Chapter 4 (Test Dimensionality and Factor Analysis): This chapter includes an entirely new section, "A Deeper Perspective on Factors, Factor Loadings, and Rotation." As a teacher, I find that students are often confused and dubious of the idea of rotation. But I find that clarifying this issue provides significant improvements in students' understanding of factor analysis more generally. Thus, the purpose of this significant new section is to provide such clarification through illustration and metaphor. A much more minor (but useful) revision was to point readers to the notion of bifactor models, which have received expanded attention in recent years.

Chapter 5 (Reliability: Conceptual Basis): This is one of the most heavily revised chapters in the new edition. It now includes an entirely new and lengthy section, "From Theory to Practice: Measurement Models and Their Implications for Estimating Reliability." This section covers the core measurement models (parallel tests, tau-equivalent tests, etc.) that have direct implications for the appropriateness of various methods of estimating reliability. Thus, this new section is crucial for bridging the gap between reliability theory and the actual practice of estimating reliability. Previous editions of the book were almost completely missing this important information. The new section includes an integrative table that summarizes and differentiates the models, extensive discussion of the models and their differences, and illustrations with simple data that exactly conform to each model.

Chapter 6 (Empirical Estimates of Reliability): There are several key additions to this chapter. First and most broadly, the chapter now integrates the discussion of the various methods of estimating reliability with the measurement models newly added to the previous chapter. The goal is to clarify when (and why) particular methods provide legitimate estimates of reliability. Thus, the revised chapter goes even further in bridging the gap between the conceptual and practical basis of reliability. Second, the chapter now discusses confidence intervals around alpha, as well as their meaning and methods for obtaining them (via SPSS). Third, it now discusses the omega coefficient and related estimates of reliability. These are relatively recent and important developments in psychometric theory, but they were completely omitted from previous editions of the book. Fourth, the chapter goes

much further in outlining the limitations of alpha as a "go-to" index of reliability, again with reference to the newly added content on measurement models. Psychometricians have raised serious doubts about the applicability of alpha in many circumstances that are probably common in psychological testing, and the revised edition outlines these issues in much greater depth, noting that omega (and related estimates) are more widely applicable.

Chapter 7 (The Importance of Reliability): Revisions to this chapter were relatively minor compared with the preceding Chapters 4 to 6. Aside from a deepened and more thorough discussion of confidence intervals around individual scores, revisions focused on clarity, readability, and updating with more recent sources.

Chapter 8 (Validity: Conceptual Basis): The new edition of this chapter reflects some of the most fundamental revisions in the entire book. Upon reflection, I realized that the previous editions lacked much depth with regard to three of the five types of validity that are covered—content validity, response process validity, and consequential validity. The other two facets of validity (internal structure and associations with other variables) received more attention and included information about the types of evidence relevant for each. However, coverage of content, response process, and consequential validity was light by comparison and included no systematic description of the types of evidence that test users should consider when evaluating those forms of validity. Thus, a significant amount of coverage has been added to resolve what were important (and obvious in retrospect) omissions in the previous editions of the book. Moreover, the chapter now includes an integrative table that defines each facet of validity, notes the relevant evidence, and provides highly relevant citations for additional reading.

Chapter 9 (Estimating and Evaluating Convergent and Discriminant Validity Evidence): There are three main revisions to this chapter. First, it includes a new introductory section that should be more engaging for readers and that should set the stage more firmly for the conceptual and practical issues to be discussed. This new section describes validity work conducted in the development and evaluation of the Need to Belong Scale (NTBS). It does so in a way that highlights the idea of a nomological network, which lies at the heart of the methods outlined later in the chapter. The second major revision is deepened coverage of restricted range and its effects on validity coefficients. The revised chapter provides a more robust example of that problem and a description of a relatively common correction for the problem. Third, the chapter provides a much deeper presentation of sensitivity and specificity, which are central to the evaluation of validity in some areas of behavioral science (and other domains of science as well, such as medical research).

Chapter 10 (Response Biases): This chapter benefited from a relatively large degree of revision, mainly focusing on integrating recent literature related to meaning, existence, and implications of each type of bias. Thus, the entire chapter provides a much more up-to-date discussion of response biases. Although each type of bias received significant attention and updating, the most extensive attention was given to random responding, which received relatively little attention in previous editions. The chapter also includes an integrative table that defines each type of bias and cites relevant literature. Finally, the discussion of balanced scales has been

expanded and updated to reflect current debate about the pros and cons of including negatively keyed items.

Chapter 11 (Test Bias): There are several substantial revisions to this chapter. First, it now includes a discussion of the connection between reliability and construct bias. Second, it has expanded coverage of the way that factor analysis—both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA)—can be used to evaluate construct bias. This is important, as EFA and CFA are likely the most common sophisticated way of conceptualizing and evaluating construct bias (i.e., measurement invariance). Third, the chapter has significantly improved coverage of the way that multiple regression is actually used to detect predictive bias—in terms of both intercept bias and slope bias. This improved coverage also includes discussion of recent literature addressing the challenges of detecting predictive bias.

Chapter 12 (Confirmatory Factor Analysis): This chapter was new to the second edition, and it received the greatest expansion for the third edition. The previous version of the chapter primarily focused on describing CFA, in terms of its purpose, logic, application, and interpretation. It provided a relatively minimal description of how CFA is used to address a variety of psychometric issues. The new version of the chapter does a much better job of describing and illustrating the way that CFA is used to conceptualize and evaluate a variety of fundamental psychometric issues. These issues include (a) using CFA to evaluate specific types of measurement models (e.g., parallel tests, tau-equivalent, etc.) as newly added to Chapters 5 and 6, (b) using CFA to estimate reliability (a deepened version of what was in the previous edition), and (c) using CFA to evaluate measurement invariance, resonating what has been added to Chapter 11's discussion of construct bias. Given the expanding use of CFA in psychometric theory and analysis, it was important to expand and strengthen this chapter.

Chapter 13 (Generalizability Theory): Revisions to this chapter were relatively minimal. They primarily focused on clarification throughout the chapter. In addition, however, the new chapter includes a large number of references to recent applications of generalizability theory to actual psychometric examinations across a range of disciplines, from sport psychology, to auditory perception, to medical research.

Chapter 14 (Item Response Theory and Rasch Models): Several key revisions to this chapter have been made. First, it provides better discussion of what item response theory (IRT) measurement models are and their relevance to IRT. Second and more substantially, it provides a completely new discussion of the three-parameter logistic model, which is widely used in some areas of testing. Third, it includes an expanded discussion/illustration of the process of estimating parameters, hopefully providing readers with a more complete and coherent understanding of the process. Finally and importantly, it includes a completely new discussion of model fit—what it is and why it is important.

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 third edition, I relied on the invaluable assistance of several people. Reid Hester, 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 Reid's consistent interest and support for this book over the years. In addition, Abbie Rickard and Jennifer Cline were wonderfully helpful with the logistics of preparing and submitting the revised manuscript. Earlier editions benefited greatly from other people at SAGE, including Jim Brace-Thompson, Cheri Dellelo, and Anna Mesick (first edition), along with Chris Cardone, Astrid Virding, Lisa Shaw, and Sarita Sarak (second edition). Earlier editions of this book also benefited from support from both Wake Forest University and Appalachian State University.

I realized that, over the first two editions of this book, I failed to express my deep appreciation 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 this edition 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 Publications 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

About the Author

R. Michael Furr is Professor of Psychology 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. 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.

CHAPTER 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. This suggestion might seem overly sensational, far-fetched, and perhaps even simply wrong, but 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 person with an intellectual disability shall be sentenced to death" (N.C. Gen. Stat. § 15A-2005); 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 he or she might not be sentenced to death by the state of North Carolina; however, if a person has an IQ score above 70, then he or

she 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, it is imperative that such tests have extremely high quality. If testing has such robust implications, then it should be conducted 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, we believe that everyone needs 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.

If you are reading this book, then 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 who make decisions about people 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 must take 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.

We address 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?

We hope 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.

We hope that 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 the observable properties of the physical world. For example, if a person wants to measure the length of a piece of lumber, then he or she 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 make this identification, we must measure each of their 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 develop a task involving observable behavior that would allow us to measure working memory. For example, we might ask the students to repeat a string of digits presented to them one at a time and in rapid succession. If our two students differ in their performance on this task, then we might assume that they differ in their working memory. That is, we observe the difference in their task performance, and we interpret it as reflecting a difference in their

working memory. If Sam could repeat more of the digits than William, then we might conclude that Sam's working memory is in some way superior to William's. This conclusion requires that we make an inference—that an observable behavior, the number of recalled digits, is systematically related to an unobservable mental attribute, working memory.

There are three things that you should 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 working memory. If our inference was 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, we say that our interpretation of scores on the measure is valid.

Second, for our interpretation of digit 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 the 100-meter dash. 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 psychological characteristics, processes, or states to an observable behavior that is thought to reflect differences in the 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 turn to 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 digits 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). The definition includes three important components: (1) tests involve behavioral samples of some kind, (2) the behavioral samples must be collected in some systematic way, and (3) the purpose of the tests is to compare the behaviors of two or more people. We would modify the third component to include a comparison of performance by the same individuals at different points in time, but otherwise we find the definition 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, Steer, & Brown, 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 he or she 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 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 Education 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 the 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). Note that we are not making any claims about the quality of the information obtained from the tests that we are using as examples. In Chapter 2, we will discuss the 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. 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: There are individually administered tests, and there are tests designed to be administered to groups of people.

Another common distinction concerns the intended purpose of test scores. Psychological tests are often categorized as either *criterion referenced* (also called domain referenced) or *norm referenced.* 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 performance criterion 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 the 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 well-defined population. 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 are of little value if the reference sample is not representative of some population, if the relevant population is not well defined, or if there is doubt that 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 take the test 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, we will see that there are slightly different methods used to assess the quality of criterion-referenced and norm-referenced tests.

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.

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 response on an intelligence test is 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

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. We 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 & Lennox, 1991; Diamantopoulos & Winklhofer, 2001; Edwards, 2010; Edwards & Bagozzi, 2000; Howell, Breivik, & Wilcox, 2007). Our goal here is 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.* We will sometimes refer to tests as instruments and sometimes as measures. The word *battery* will be restricted in use to references to bundled tests; bundled tests are instruments intended to be administered together but are not *necessarily* designed to measure a single psychological attribute. The word *measure* is one of the most confusing words in the psychology testing literature. In Chapter 2, we are going to discuss in detail the use of this word as a verb, as in "The BDI was designed *to measure* depression." The word *measure* also is often used in its noun form, as in "The BDI is a good *measure* of depression." We will use both forms of the term and rely on the context to clarify its meaning.

Psychometrics

What Is Psychometrics?

We previously defined a test as a procedure for systematically sampling behavior. These behavioral samples are attempts to measure, at least in some sense, psychological attributes of people. The act of giving psychological tests to people is referred to as testing. In this book, we will not be concerned with the process of testing; rather, our concern will focus on psychological tests themselves. We will not, however, be concerned with particular psychological tests, except as a test might illustrate an important principle. In sum, we focus on the *attributes* of tests.

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.

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. The important analogy is that just as psychological tests are about theoretical attributes of people, psychometrics is about theoretical attributes of psychological tests. Just as psychological attributes of people must be measured, so also psychometric attributes of tests 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 upon two key foundations. One foundation is the practice of psychological testing and measurement. As most textbooks in psychological testing point out (e.g., 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 quickly adopted 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 subsequent sections of this book.

Based upon 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, we focus 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 the sciences, measurement in the behavioral sciences has special challenges that do not exist or are greatly reduced in the physical sciences. These challenges affect our confidence in our understanding and interpretation of behavioral observations. We will find that one of these challenges is related to the complexity of psychological phenomena; notions such as intelligence, self-esteem, anxiety, depression, and so on have many different aspects to them. Thus, one of our challenges is to try to identify and capture the important aspects of these types of human psychological attributes in a single number.

Participant reactivity is another such 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 determine whether you are a racist. Your responses to the questionnaire might be influenced by your desire not to be thought of as a racist rather than by your true attitudes toward people who belong to ethnic or racial groups other than your own. Therefore, people's knowledge that they are being observed can cause them to react in ways that obscure the interpretation of the behavior that is being observed. This is usually not a problem when measuring features of nonsentient physical objects; the weight of a bunch of grapes is not influenced by the act of weighing them.

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 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 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.

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 biases and expectations to their task. Measurement quality is compromised when observers allow these influences to distort their observations. *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, he or she 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 third important respect. 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 some aspect of a particular 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 total 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 a later chapter), 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 fourth challenge to psychological measurement is *score sensitivity.* Sensitivity refers to the ability of a measure to discriminate adequately between meaningful amounts or units of the dimension that is being measured. As an example from the physical world, 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 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," 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, we hope to demonstrate 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. In this book, we discuss 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

Our ability to identify and characterize psychological differences is at the heart of all psychological measurement and 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 the person's score on a psychological test, 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 about the samples can be obtained only if people differ with respect to the behavior that we are sampling. If a behavioral sampling procedure produces scores that differ between people (or that differ across time or condition), then the psychometric properties of the scores obtained from the sampling procedure can be assessed along a wide variety of dimensions. In this book, we will present the logic and analytic procedures associated with these psychometric properties.

If we think that a particular behavioral sampling procedure is a measure of an unobservable psychological attribute, then we must be able to argue that differences in the scores derived from that procedure are indeed related to differences on 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 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.

In the following chapters, we will show how individual differences are quantified and how their quantification is the first step in solving many of the challenges to measurement in psychology to which we have already alluded. Individual differences represent the currency of psychometric analysis. In effect, individual differences 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 upon the existence and detection of psychological differences among people, we want to avoid a common misinterpretation. The misinterpretation 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, contemporary authors sometimes view psychometrics as an issue that concerns only those who study "individual differences" topics such as intelligence, ability/aptitude, or personality. They sometimes seem to believe that

psychometrics is not a concern for those who take a more experimental approach to human behavior. We absolutely disagree with this view.

Our view is that psychometrics is not limited to issues in differential psychology. Rather, our view is that 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. We will return to this issue later in the book, 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.* Boston, MA: 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 & S. Sinharay (Eds.), *Handbook of statistics, 26: Psychometrics* (pp. 1–27). Amsterdam: 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,* 515–534.

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.). Thousand Oaks, CA: Sage.

PART I

Basic Concepts in Measurement

CHAPTER 2

Scaling

If something exists, it must exist in some amount (E. L. 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 a full appreciation of scaling and its implications depends on a variety of abstract issues. In this chapter, we discuss 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, we emphasize psychological tests that are intended to measure unobservable psychological characteristics, such as attitudes, personality traits, and intelligence. Such characteristics present special problems for measurement, and we will discuss several possible solutions for these problems.

We acknowledge that 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.

Thus, we encourage you to devote attention to the concepts in this chapter. We believe that your attention will be rewarded with new insights into the foundations of psychological measurement and even into the nature of numbers. Indeed, in preparing this chapter, our own understanding of such issues has grown and evolved.

Fundamental Issues With Numbers

In psychological measurement, numerals are used to represent an individual's level of a psychological attribute. For example, we use your numerical score on an IQ test to represent your level of intelligence, we might use your numerical score on the Rosenberg Self-Esteem Inventory to represent your level of self-esteem, and we might even use a 0 or a 1 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.

Given this heavily numerical orientation, it is important to understand that numerals, however, can represent psychological attributes in different ways, depending on the nature of the numeral that is used to represent the attribute. In this section, we describe important properties of numerals, and we show how these properties influence the ways in which numerals represent psychological attributes.

We must understand three important numerical properties, and we must understand the meaning of zero. In essence, the numerical properties of 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 measurements are those that differentiate between categories 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, the children with behavior problems 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 or events 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 he or she 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.

At this level, numerals serve simply as labels of categories. The categories could be labeled with letters, names, or numerals. We could label the category of children with behavior problems as "Behavior Problem Children," we could refer to the category as "Category B," or we could assign a numeral to the category. For example, we could label the group as "0," "1," or "100." At this level, 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 merits some additional depth. When making categorical differentiations between people, the distinctions between members of different 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 quantity than in a simple presence/absence categorization. This leads to additional properties of numerals.

The Property of Order

Although the property of identity reflects the most fundamental form of measurement, the property of order conveys greater 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 he or she possesses 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 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, the numerals 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 he or she has. 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 expand on 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 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, you must understand which meaning is relevant.

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, we 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. The mean of a distribution of *z* scores will always be 0. Zero in this case represents an arbitrary or relative zero.

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

We will use an example from E. L. Thorndike (2005) to illustrate this problem. Thorndike describes a scenario in which sixth-grade children are given a spelling test. He asks us to imagine that one of the children fails to spell correctly any of the words on the test. That is, the child 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. That is, a test score of 0 means that the child failed to answer any of the spelling questions correctly. 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?

Interpretation of psychological test scores will be influenced by the type of zero associated with a test. As a technical matter, 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, we would probably want to restrict arithmetical 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, 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. In the next two sections, we will discuss two fundamental issues regarding the meaning and use of such quantitative test scores. Specifically, we will 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 we will discuss in the next section, quantitative measurement depends on our ability to count these units. Before we discuss the process and implications of counting the units of measurement, we must clarify what is meant by a unit of measurement.

In many familiar cases of physical measurement, the units of measurement are readily apparent. If people want to measure the length of a piece of lumber, then they will probably use some type of tape marked off in units of inches or centimeters. The length of the piece of lumber is determined by counting the number of these units from one end of the board to the other end.

In contrast, in many cases of psychological measurement, units of measurement are often less obvious. When we measure 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? We will return to these questions at a later time, as they represent the most vexing problems in psychometrics. At this point, we simply want to concentrate on the notion of a unit of measurement. Because this notion can be most easily illustrated in the context of the measurement of the length of physical objects (Michell, 1990), we will introduce it in that way.

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?

When push comes to shove, you could 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 an "xrod." Because your pieces of bookshelf wood are longer than your xrod, you will need a number of 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, place xrods end to end in a straight line until you reach the opposite end of the piece of wood. Now count the number of xrods, and you might find that the shelf is "8 xrods long."

You have just measured length in "units of xrods." You can use your set of xrods to measure the length of each and every piece of wood that you need. In fact, you could use your xrods to measure the length of many things, not just pieces of wood. In many ways, your measure is as good as any measure of length (except that you are the only one who knows what an xrod represents!).

Arbitrariness is an important concept in understanding units of measurement, and it distinguishes between different kinds of measurement units. There are three ways in which a measurement unit might be arbitrary. First, the unit size can be arbitrary. That is, the specific size of a unit might be arbitrary. Consider your xrod—the size of your original xrod could have been any length. When you cut that first xrod, your decision about its length could be completely arbitrary—there was no "true" xrod length that you were trying to obtain. You simply chose a length to cut, and that length became the "official" length of an xrod. In this sense, the actual length of our unit of measurement, the xrod, was completely arbitrary. Similarly, the amount of weight that is represented by a "pound" is an arbitrary amount. Although there is now clear consensus regarding the exact amount of weight represented by a pound, we can ask why a pound should reflect that *specific* amount. The choice was likely quite arbitrary.

A second form of arbitrariness is that some units of measurement are not tied to any one type of object. That is, there might be no inherent restriction on the objects to which a unit of measurement might be applied. Our xrods can be used to measure the spatial extent of anything that has spatial extent. For example, they could be used to measure the length of a piece of wood, the length of a table, the distance between two objects, or the depth of water in a swimming pool. Similarly, a pound can be used to measure the weight of many different kinds of objects.

A third form of arbitrariness is that, when they take a physical form, some units of measurement can be used to measure different features of objects. For example, the xrods that we used to measure the length of a piece of lumber could also be used as units of weight. Imagine that you needed to measure the weight of a bag of fruit. If you had a balance scale, you could put the bag in one of the balance's baskets, and you could gradually stack xrods in the other basket. When the two sides of the scale "balance," you would know that the bag of fruit weighs, say, 4 xrods.

Units of measurement, called *standard measures,* are based on arbitrary units of measurement in all three ways when they take a physical form. In physical measurement, standard units include units such as pounds, liters, and milliseconds. The fact that they are expressed in arbitrary units gives them flexibility and generality. For example, you can use milliseconds to measure anything from a person's reaction time to the presentation of a stimulus to the amount of time it takes a car to travel down the street.

In contrast to many physical measures, most psychological units of measurement (e.g., scores on tests such as mechanical aptitude tests or on intelligence tests) are generally arbitrary only in the first sense of the term *arbitrary* mentioned above. That is, most psychological units of measurement are arbitrary in size, but they are typically tied to specific objects or dimensions. For example, a "unit" of measurement on an IQ test is linked in a nonarbitrary way to intelligence, and it is not applicable to any other dimension. Because of this feature of IQ test scores, we refer

to IQ score units as "IQ points"; the points have no referent beyond the test used to measure intelligence. There is one important exception to this observation; standard measures are sometimes used to measure psychological attributes. For example, reaction times are often used to measure various cognitive processes.

Additivity and Counting

The need for counting is central to all attempts at measurement. Whether we are trying to measure a feature of the physical world or of the psychological world, all measurement involves counting. For example, when you used xrods to measure the length of a piece of wood, you placed the xrods end to end, starting from one end of the piece of wood and continuing until you reached the other end. You then counted the xrods to determine the length of the object. The resulting count was a measure of length. Similarly, when you use a behavioral sampling procedure (i.e., a test) to measure a person's self-esteem, you count responses of some kind. For example, you might count the number of test statements that a test respondent marks as "true," and you might interpret the number of "true" marks as indicating the level of the respondent's self-esteem. That is, you count units to obtain a score for your measurement.

Additivity

Importantly, the process of counting as a facet of measurement involves a key assumption that might not be valid in many applications of psychological measurement. The assumption is that the unit size does not change—that all units being counted are identical. In other words, additivity requires unit size to remain constant; a *unit* increase at one point in the measurement process must be the same as a unit increase at any other point.

Recall the xrod example, where you used the original xrod as a guide to cut additional xrods—we encouraged you to make "sure that each xrod is the same exact length as your original xrod." By doing so, you ensured that anytime you laid xrods side by side and counted them, you could trust that your count accurately reflected a length. Say that you had cut 10 xrods; if they are all identical, then it does not matter which xrods you used when measuring the length of any piece of wood. That is, a piece of wood that you measured as 5 xrods would be measured as 5 xrods no matter which particular 5 xrods you used to measure the piece of wood.

Now imagine that instead of having a collection of equal-length xrods, your xrods had various lengths. In that case, if you measured the same piece of wood on two occasions, you might get two different counts, indicating different lengths! That is, if some xrods were longer than the others, then your piece of wood might be 5 xrods when you use the shorter xrods, but it would be only 3 xrods if you happened to use the longer xrods. Because your units are not constant in magnitude, your entire measurement system is flawed—there is no single unit of length that is represented by an xrod. This would prevent you from determining the real length of the lumber.

In addition, the size of a measurement unit should not change as the conditions of measurement change. For example, the size of an xrod should remain constant regardless of the time of day that the xrod is used to measure a piece of wood. In effect, we want our measure to be affected by only one attribute of the thing we are measuring, regardless of the conditions that exist at the time or place of measurement. This condition is referred to as conjoint measurement (Luce & Tukey, 1964) and is a complex issue beyond the scope of this book (but see Green & Rao, 1971, for a clear, nontechnical discussion).

Although these issues might be initially clearest in terms of physical measurements (e.g., xrods), we are most concerned about psychological measurement. So imagine that you are a history teacher who wants to measure a psychological attribute such as "knowledge of American history." Generally, this would be done by asking students a series of questions that you believed were diagnostic of their knowledge, recording their responses to the questions. Let us temporarily differentiate between measurement units and psychological units. That is, each test item represents a measurement unit, and again you count the correctly answered items to obtain a score that you interpret as a student's knowledge of American history. In contrast, we will use the crude and informal idea of psychological units to mean "true" levels of knowledge. Ideally, the measurement units will correspond closely with psychological units. That is, we use test scores to represent levels of psychological attributes. With this in mind, you combine each student's test responses in some way (e.g., by counting the number of questions that each student answered correctly) to create a total score that is interpreted as a measure of true knowledge of American history.

Suppose that one of the questions on your test was "Who was the first president of the United States?" and another was "Who was the first European to sail into Puget Sound?" It should be clear that the amount of knowledge of American history you need to answer the first question correctly is considerably less than the amount you need to answer the second question correctly. In terms of psychological units, let's say that you needed only 1 psychological unit of American history knowledge to answer the first question correctly but you needed three times as much knowledge (i.e., 3 psychological units of knowledge) to answer the second question correctly.

Consider a student who answered both questions correctly. In terms of amount of true knowledge, that student would have 4 psychological units of history knowledge. However, in terms of measurement, that student would have a score of only 2. That is, if you simply summed the number of correct responses to the questions to get a total score, the student would get a score of 2. This would suggest that the person had 2 units of American history knowledge when in fact he or she had 4 units of knowledge.

This discrepancy occurs because the measurement units are not constant in terms of the underlying attribute that they are intended to reflect. That is, the answers to the questions are not a function of equal-sized units of knowledge—it takes less knowledge to answer the first question than it does to answer the second. Thus, the additive count of correct answers is not a good measure of amount of knowledge.

From a psychological perspective, the assumption is often made that a psychological attribute such as knowledge of American history actually exists in some amount. However, unlike a piece of wood, whose "length" can be directly observed,