Robert F. DeVellis • Carolyn T. Thorpe

Scale Development

Theory and Applications

FIFTH EDITION

Scale Development

Fifth Edition

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

Scale Development

Theory and Applications

Fifth Edition

Robert F. DeVellis
The University of North Carolina at Chapel Hill
Carolyn T. Thorpe
The University of North Carolina at Chapel Hill

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.
18 Cross Street #10-10/11/12
China Square Central
Singapore 048423

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Names: DeVellis, Robert F., author. | Thorpe, Carolyn T., author.

Title: Scale development : theory and applications / Robert F. DeVellis, The University of North Carolina at Chapel Hill, Carolyn T. Thorpe, The University of North Carolina at Chapel Hill.

Description: Fifth edition. | Thousand Oaks, California : SAGE Publications, Inc., [2022] | Includes bibliographical references and index. | Identifiers: LCCN 2021025638 | ISBN 9781544379340 (paperback) | ISBN 9781544379357 (epub) | ISBN 9781544379333 (epub) | ISBN 9781544379326 (ebook)

Subjects: LCSH: Scaling (Social sciences)

Classification: LCC H61.27 .D48 2022 | DDC 300.72—dc23

LC record available at https://lccn.loc.gov/2021025638

This book is printed on acid-free paper.

Acquisitions Editor: Helen Salmon
Product Associate: Ivey Mellem
Production Editor: Gagan Mahindra
Copy Editor: Karin Rathert
Typesetter: Hurix Digital
Cover Designer: Gail Buschman
Marketing Manager: Victoria Velasquez

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

• Brief Contents •

Preface			**xiii**
Acknowledgments			**xv**
About the Authors			**xvii**
Chapter 1	•	Overview	1
Chapter 2	•	Understanding the Latent Variable	19
Chapter 3	•	Scale Reliability	33
Chapter 4	•	Scale Validity	71
Chapter 5	•	Guidelines in Scale Development	91
Chapter 6	•	Factor Analysis	137
Chapter 7	•	The Index	183
Chapter 8	•	An Overview of Item Response Theory	231
Chapter 9	•	Measurement in the Broader Research Context	257
References			**269**
Index			**281**

• Detailed Contents •

Preface xiii

Acknowledgments xv

About the Authors xvii

Chapter 1 • Overview 1

General Perspectives on Measurement 2

Historical Origins of Measurement in Social Science 3

Early Examples 3

Emergence of Statistical Methods and the Role of Mental Testing 5

The Role of Psychophysics 5

Later Developments in Measurement 6

Evolution of Basic Concepts 6

Evolution of Mental Testing 7

Assessment of Mental Illness 7

Broadening the Domain of Psychometrics 10

The Role of Measurement in the Social Sciences 11

The Relationship of Theory to Measurement 11

Theoretical and Atheoretical Measures 12

Composite Measurement Tools 13

All Scales Are Not Created Equal 15

Costs of Poor Measurement 16

Summary and Preview 17

Exercises 18

Chapter 2 • Understanding the Latent Variable 19

Constructs Versus Measures 19

Latent Variable as the Presumed Cause of Scale Item Values 21

Path Diagrams 22

Diagrammatic Conventions 22

Path Diagrams in Scale Development 22

Further Elaboration of the Measurement Model 25

Classical Measurement Assumptions 25

Parallel Tests 26

Alternative Models 29

Choosing a Causal Model 30
Exercises 32
Note 32

Chapter 3 • Scale Reliability 33

Methods Based on the Analysis of Variance 34
Continuous Versus Dichotomous Items 35
Internal Consistency 36
Coefficient Alpha 36
The Covariance Matrix 37
Covariance Matrices for Multi-Item Scales 38
Alpha and the Covariance Matrix 39
Alternative Formula for Alpha 43
Critique of Alpha 45
Remedies to Alpha's Limitations 49
Coefficient Omega (ω) 50
Reliability Based on Correlations Between Scale Scores 52
Alternate-Forms Reliability 52
Split-Half Reliability 53
Inter-Rater Agreement 56
Temporal Stability 58
Reliability of Change Scores 60
Reliability and Statistical Power 65
Generalizability Theory 66
Summary 68
Exercises 69
Notes 70

Chapter 4 • Scale Validity 71

Content Validity 72
Scope of the Variable and Implications for Content Validity 73
Criterion-Related Validity 78
Criterion-Related Validity Versus Accuracy 79
Construct Validity 82
Differentiating Construct From Criterion-Related Validity 83
Attenuation 84
How Strong Should Correlations Be to Demonstrate Construct Validity? 85
Multitrait-Multimethod Matrix 85
What About Face Validity? 87
Exercises 89

Chapter 5 • Guidelines in Scale Development **91**

Step 1: Determine Clearly What It Is You Want to Measure 91
Theory as an Aid to Clarity 91
Specificity as an Aid to Clarity 92
Being Clear About What to Include in a Measure 93
Step 2: Generate an Item Pool 94
Choose Items That Reflect the Scale's Purpose 94
Redundancy 95
Number of Items 97
Beginning the Process of Writing Items 98
Characteristics of Good and Bad Items 98
Positively and Negatively Worded Items 101
Conclusion 102
Step 3: Determine the Format for Measurement 102
Thurstone Scaling 103
Guttman Scaling 104
Scales With Equally Weighted Items 105
How Many Response Categories? 106
Specific Types of Response Formats 110
Likert Scale *110*
Semantic Differential *112*
Visual Analog *113*
Pictorial Response Options *114*
Numerical Response Formats and Basic Neural Processes *116*
Binary Options *117*
Item Time Frames *117*
Mode of Administration 118
Step 4: Have Initial Item Pool Reviewed by Experts 119
Step 5: Cognitive Interviewing 120
Step 6: Consider Inclusion of Validation Items 122
Step 7: Administer Items to a Development Sample 123
Step 8: Evaluate the Items 125
Initial Examination of Items' Performance 125
Reverse Scoring *126*
Item-Scale Correlations *127*
Item Variances *128*
Item Means *128*
Dimensionality 129
Reliability 129
Step 9: Optimize Scale Length 131
Effect of Scale Length on Reliability 131

Effects of Dropping "Bad" Items 131
Tinkering With Scale Length 132
Split Samples 133
Exercises 134
Note 135
Chapter 6 • Factor Analysis 137
Overview of Factor Analysis 138
Examples of Methods Analogous to Factor Analytic Concepts 139
Example 1 *139*
Example 2 *141*
Shortcomings of These Methods *142*
Conceptual Description of Factor Analysis 144
Extracting Factors 144
The First Factor *144*
Subsequent Factors *146*
Deciding How Many Factors to Extract *147*
Rotating Factors 152
Rotation Analogy 1 *153*
Rotation Analogy 2 *154*
Rotation Analogy 3 *155*
Orthogonal Versus Oblique Rotation *161*
Choosing Type of Rotation *164*
Bifactor and Hierarchical Factor Models 165
Interpreting Factors 170
Principal Components Versus Common Factors 172
Same or Different? 172
Confirmatory Factor Analysis 176
Using Factor Analysis in Scale Development 178
Sample Size 180
Conclusion 182
Exercises 182
Chapter 7 • The Index 183
How an Index Differs From a Scale 183
The Two Distinct Types of Index 185
Causal Formative Measures *185*
Composite Formative Measures *186*
Other Conceptual Differences Between Scales and Indices 188
Empirical Differences Between Scales and Indices 189
Relationships Among Indicators *189*
Impact of Adding Items *190*
Rules of Thumb for Differentiating an Index From a Scale 192
Limitations of Conceptual Criteria 193
Limitations of Empirical Criteria 195

Is It a Scale or an Index? Formal Methods for Distinguishing Effect and Causal Indicators 196
The Correlation Matrix 196
Factor Analysis 197
Vanishing Tetrads 198
Steps in Developing and Evaluating an Index 202
Index Item Development 202
The Role of Theory *202*
Index Item Creation *202*
Index Item Redundancy and Number *203*
Other Index Development Considerations *203*
Evaluating Items 203
Regression as a Heuristic for Index Development *204*
Regression as an Alternative to Index Development *205*
Regression as a Method for Index Development and Validation *207*
Regression-Based Examples *207*
Index Validity 212
Content Validity *212*
Construct Validity *212*
Criterion Validity *216*
Group Comparison Approach *217*
Index Reliability 218
Hybrid Measures 219
Hierarchical Hybrids 219
Hierarchical Hybrid Indices and Multidimensional Scales 220
Hybrids Involving Nonhierarchical Heterogeneous Indicators 221
Methods Based on Structural Equation Modeling 224
Heuristic Overview of Structural Equation Modeling 224
MIMIC Models 226
Criticisms of Index Composites 228
Exercises 229
Note 229
Chapter 8 • An Overview of Item Response Theory 231
Item Difficulty 234
Item Discrimination 236
Guessing, or False Positives 236
Item-Characteristic Curves 238
IRT Applied to Multiresponse Items 241
Theta and Computerized Adaptive Testing (CAT) 248
Complexities of IRT 250
Conclusions 252
Exercises 255

Chapter 9 • Measurement in the Broader Research Context 257

Before Scale Development 258
Look for Existing Tools 258
View the Construct in the Context of the Population of Interest 259
Consider the Scale in the Context of Other Measures or Procedures 260
After Scale Administration 261
Analytic Issues 261
Interpretation Issues 262
Generalizability 262
Final Thoughts 263
Small Measurement and Big Measurement 263
Canoes and Cruise Ships 263
Measurement "Canoes" and Measurement "Cruise Ships" 264
Practical Implications of Small Versus Big Measurement 266
Remember, Measurement Matters 267
Exercise 268

References 269

Index 281

• Preface •

Working on a new edition of this text is an opportunity to make valuable changes that benefit readers. The first change readers will notice is the addition of a second author, Dr. Carolyn Thorpe, to the writing team. Dr. Thorpe is an outstanding scholar whose qualifications and background are described more fully elsewhere in this volume. Suffice it to say here that she brings an extremely valuable perspective and a rich skill set, as both a teacher and researcher, to the task of communicating important topics in instrument development and evaluation. Her knowledge, clarity, and judgment have greatly improved this edition.

The goal of every revision of this text has been to meet the needs more effectively and thoroughly of researchers and scholars interested in measurement. The praise that users most often have voiced for past editions of this book is that technical topics are presented in a way that is more understandable than it is in most texts. Obviously, we have done our best to continue in that vein. The criticism most often heard is that the topic of indices, as opposed to scales, has not been given sufficient coverage. With this edition, we have addressed that shortcoming.

Although scales remain the primary focus of this book, we have long recognized that many measurement situations involve combining indicators into composites that are, in fact, indices rather than scales. Although both are measures comprising multiple separate indicators, scales and indices differ in important ways. As we talked about what we felt was most lacking in previous editions, we kept returning to the topic of indices. They received only brief mention in past editions, in part because it has taken time for a consensus to emerge among measurement experts. The differences between the two types of measures—and the distinct approaches that each require—are often misunderstood. In talking with researchers working in applied settings, we often have people present us with ideas for potential "scales" that actually seem much more like indices. But the researchers are often unaware of the distinction and the different methodologies it dictates. Once these people became aware of the basic distinction, they usually are eager for greater clarity. However, much of the relevant scholarship is technical and presents concepts that may not be familiar to non-experts. Now, having gone through much of the current work on the topic and wrestling with how to make it as clear as possible for the widest possible readership, we think we can provide some answers in ways that will be accessible to a broader audience. Accordingly, Chapter 7 in this new edition lays out the key concepts that distinguish indices from scales, contrasts various

types of indices, suggests approaches for developing them, reviews validity and reliability issues, and discusses in broad terms some analytic approaches. Throughout the chapter, we have tried to explain seemingly arcane concepts (for example, "vanishing tetrads") in simpler, more familiar terms than one typically finds in other sources. We also describe real-world examples of index development and usage. We believe that this information is a valuable addition to *Scale Development: Theory and Applications* that will help researchers develop more effective measurement tools, allow them to make more informed choices among existing tools, and use them appropriately.

Although we have paid new attention to indices, we still focus primarily on scales, as in previous editions. We have carefully reviewed all previously existing chapters. Each has been refreshed and updated while retaining the features that have served readers well in previous editions. Where warranted, we have added new information, revised illustrations, changed or added examples, added citations to more recent work, and edited for further clarity. In Chapter 4 (Scale Validity), for example, we have added information on receiver operating characteristic (ROC) curves. Similarly, in Chapter 5 (Guidelines in Scale Development), we have added a section on modes of item administration as well as new information concerning different types of item response options and expanded discussion of the use of cognitive interviewing in scale development. In our coverage of Factor Analysis (Chapter 6), we have added more about the distinction between principal components and common factors and expanded the discussion of factor rotation.

As with previous transitions between editions, the goal has been to provide readers with the background they need to understand measurement issues commonly encountered in behavioral and social research. Moreover, we have tried to present this information in a form that is clear and accessible. Our hope is to strike a balance between including relevant topics and highlighting recent developments in measurement while retaining an accessible, user-friendly approach to the material covered. We feel that, particularly with the addition of a full chapter on indices, this edition achieves that goal. We hope you will agree.

• Acknowledgments •

SAGE and the authors would like to thank the following reviewers for their comments in the development of this book:

Daniel Corts, Augustana College

Jill Lohmeier, University of Massachusetts Lowell

Kenneth L. Miller, Youngstown State University

Lydia Zablotska, University of California, San Francisco

• About the Authors •

Robert F. DeVellis is professor emeritus in the Department of Health Behavior (Gillings School of Global Public Health) at the University of North Carolina at Chapel Hill. Dr. DeVellis has more than 40 years of experience in the measurement of psychological and social variables. He served as the first domain chair for Social Outcomes of the Patient-Reported Outcomes Measurement Information System (PROMIS) consortium, a multisite National Institutes of Health (NIH) Roadmap initiative directed at identifying, modifying, testing, and disseminating outcome measures for use by NIH investigators. He has served on the board of directors for the American Psychological Association's Division of Health Psychology (38), on the Arthritis Foundation's Clinical/Outcomes/Therapeutics Research Study Section, and on the advisory board of the Veterans Affairs Measurement Excellence Initiative. He is the recipient of the 2005 Distinguished Scholar Award from the Association of Rheumatology Health Professionals and is an associate editor of *Arthritis Care and Research*. In addition, he has served as guest editor, guest associate editor, or reviewer for more than two dozen other journals. He has served as principal investigator or co-investigator since the early 1980s on a series of research projects funded by the federal government and private foundations.

Carolyn T. Thorpe, PhD, MPH, is a tenured associate professor in the Division of Pharmaceutical Outcomes and Policy at the University of North Carolina Eshelman School of Pharmacy. She holds a joint appointment as a research health scientist and core investigator in the Veterans Affairs (VA) Pittsburgh Healthcare System's Center for Health Equity Research and Promotion (CHERP), and serves as co-director of the Department of Veteran's Affairs Postdoctoral Pharmacy Fellowship in Medication Safety and Pharmacy Outcomes. She has served on study sections for the National Institutes of Health and Department of Veterans Affairs, as an associate editor for the journal *Research in Social and Administrative Pharmacy*, and as a reviewer for a diverse range of peer-reviewed journals at the intersection of public health, medicine, and social science. Over the past two decades, Dr. Thorpe has led a federally funded research program that aims to improve the health of older adults managing complex chronic conditions and has published over 100 peer-reviewed manuscripts. Her research interests lie in measuring and examining psychological, social, and behavioral factors related to the safe and appropriate use of medications in older adults, and she has specific expertise in developing and evaluating measures of patient illness self-management behavior and medication nonadherence. She enjoys teaching principles of study design and measurement to professional pharmacy and graduate students and serving as an active research mentor of trainees at all levels.

1 Overview

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

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

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

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

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

General Perspectives on Measurement

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

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

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

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

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

Historical Origins of Measurement in Social Science

Early Examples

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

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

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

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

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

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

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

Emergence of Statistical Methods and the Role of Mental Testing

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

The Role of Psychophysics

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

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

Later Developments in Measurement

Evolution of Basic Concepts

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

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

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

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

Evolution of Mental Testing

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

Assessment of Mental Illness

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

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

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

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

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

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

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

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

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

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

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

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

Broadening the Domain of Psychometrics

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

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

The Role of Measurement in the Social Sciences

The Relationship of Theory to Measurement

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

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

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

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

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

Theoretical and Atheoretical Measures

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

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

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

Composite Measurement Tools

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

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

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

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

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

All Scales Are Not Created Equal

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

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

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

Costs of Poor Measurement

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

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

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

Summary and Preview

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

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

Exercises

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

2

Understanding the Latent Variable

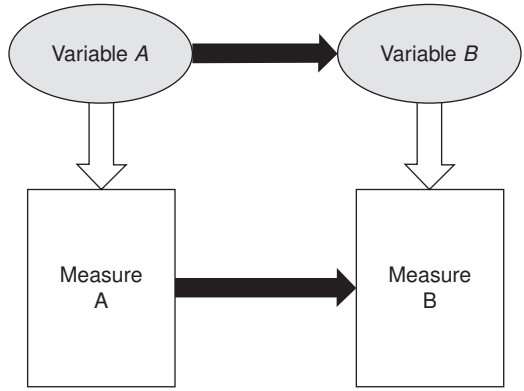
This chapter presents a conceptual schema for understanding the relationship between measures and the constructs they represent, though it is not the only framework available. Item response theory is an alternative measurement perspective that we will examine in Chapter 8. Because of its relative conceptual and computational accessibility and wide usage, we emphasize the classical measurement model, which assumes that individual items are comparable indicators of the underlying construct.

Constructs Versus Measures

Typically, researchers are interested in constructs rather than items or scales per se. For example, a market researcher measuring parents' aspirations for their children would be more interested in intangible parental sentiments and hopes about what their children will accomplish than in where those parents place marks on a questionnaire. However, recording responses to a questionnaire may, in many cases, be the best method of assessing those sentiments and hopes. Scale items are usually a means to the end of construct assessment. In other words, they are necessary because many constructs cannot be assessed directly. In a sense, measures are proxies for variables that we cannot directly observe. By assessing the relationships between measures, we indirectly infer the relationships between constructs. In Figure 2.1, for example, although our primary interest is the relationship between Variables *A* and *B*, we estimate that relationship on the basis of the connection between measures corresponding to those variables.

The underlying phenomenon or construct that a scale is intended to reflect is often called the *latent variable*. As we use the terms in this text, all scales (and

FIGURE 2.1 ● Relationships between instruments correspond with relationships between latent variables only when each measure corresponds to its latent variable



some indices) involve a latent variable. In this chapter, unless otherwise noted, our discussion is limited to scale items. Exactly what is a latent variable? Its name reveals two chief features. Consider the example of parents' aspirations for children's achievement. First, it is *latent* rather than manifest. Parents' aspirations for their children's achievement are not directly observable. In addition, the construct is *variable* rather than constant—that is, some aspect of it, such as its strength or magnitude, changes. Parents' aspirations for their children's achievement may vary according to time (e.g., during the child's infancy versus adolescence), place (e.g., on an athletic field versus a classroom), people (e.g., parents whose own backgrounds or careers differ), or any combination of these and other dimensions. The latent variable is the actual phenomenon that is of interest—in this case, child achievement aspirations.

Another noteworthy aspect of the latent variable in the case of a scale is that it is typically a characteristic of the individual who is the source of data. Thus, in our present example, parental aspirations are a characteristic of the parents and not of the children. Accordingly, we assess it by collecting data about the parents' beliefs from the parents themselves. While there may be circumstances in which some form of proxy reporting (e.g., asking parents to report some characteristic of their children) is appropriate, in general, we will ask respondents to self-report information pertaining to themselves. When this is not the case, as in a study involving parents describing the aspirations their children have for themselves, care must be taken in interpreting the resulting information. Arguably, in this hypothetical instance, the latent variable might more accurately be described as *parents' perceptions of their children's aspirations* than as *children's aspirations* per se. Likewise, if we ask a group of shoppers to evaluate characteristics of a particular store, we are assessing *shoppers' perceptions* rather than aspects of the store itself (which might be more easily assessed

by direct observation). How important the distinction is between assessing the perceptions of a respondent with regard to some external stimulus (e.g., perceptions of the store), as opposed to characteristics of the external stimulus (e.g., the store itself), will depend on the specific circumstances and goals of the assessment; however, in all cases, it is important to be mindful of the distinction and to make appropriate interpretations of the resultant data.

Although we cannot observe or quantify it directly, the latent variable presumably takes on a specific value under some specified set of conditions. A scale developed to measure a latent variable is intended to estimate its actual magnitude at the time and place of measurement for each thing measured. This unobservable actual magnitude is the *true score*.

Latent Variable as the Presumed Cause of Scale Item Values

The notion of a latent variable implies a certain relationship between it and the items that tap it. The latent variable is regarded as a *cause* of the scale item score—that is, the strength or quantity of the latent variable (i.e., the value of its true score) is presumed to cause an item (or set of items) to take on a certain value.

An example may reinforce this point: The following are hypothetical items for assessing parents' aspirations for children's achievement:

1. My child's achievements determine my own success.
2. I will do almost anything to ensure my child's success.
3. No sacrifice is too great if it helps my child achieve success.
4. My child's accomplishments are more important to me than just about anything else I can think of.

If parents were given an opportunity to express how strongly they agree with each of these items, their underlying aspirations for childhood achievement should influence their responses. In other words, each item should give an indication of how strong the latent variable (aspirations for children's achievement) is. The score obtained on the item is caused by the strength or quantity of the latent variable for that person at that particular time.

A causal relationship between a latent variable and a measure implies certain empirical relationships. For example, if an item value is caused by a latent variable, then there should be a correlation between that value and the true score of the latent variable. As a consequence of each of the indicators correlating with the latent variable, they should also correlate with each other. Because we cannot directly assess the true score, we cannot compute a correlation between it and the item. However, when we examine a set of items that are presumably caused by the same latent variable, we can examine their relationships to one

another. So if we had several items like the ones preceding measuring parental aspirations for child achievement, we could look directly at how they correlated with one another, invoke the latent variable as the basis for the correlations among items, and use that information to infer how highly each item was correlated with the latent variable. Shortly, we will explain how all this can be learned from correlations among items. First, however, we will introduce some diagrammatic procedures to help make this explanation more clear.

Path Diagrams

Coverage of this topic will be limited to a brief review of issues pertinent to scale development. For greater depth, consult Asher (1983) or Loehlin (1998).

Diagrammatic Conventions

Path diagrams are a method for depicting *causal* relationships among variables. Although they can be used in conjunction with path analysis, which is a data analytic method, path diagrams have more general utility as a means of specifying how a set of variables are interrelated. These diagrams adhere to certain conventions. A *straight arrow* drawn from one variable label to another indicates that the two are *causally related* and that the direction of causality is as indicated by the arrow. Thus $X \rightarrow Y$ indicates explicitly that X is the cause of Y. Often, associational paths are identified by labels, such as the letter *a* in Figure 2.2.

The *absence* of an arrow also has an explicit meaning—namely, that two variables are *unrelated*. Thus, $A \rightarrow B \rightarrow C\ D \rightarrow E$ specifies that A causes B, B causes C, C and D are *unrelated,* and D causes E.

Another convention of path diagrams is the method of representing *error,* which is usually depicted as an additional causal variable. This error term is a *residual,* representing all sources of variation not accounted for by other causes explicitly depicted in the diagram.

Because this error term is a residual, it represents the discrepancy between the actual value of Y and what we would predict Y to be based on knowledge of X and Z (in this case; see Figure 2.3). Sometimes, the error term is assumed and, thus, not included in the diagram.

Path Diagrams in Scale Development

Path diagrams can help us see how scale items are causally related to a latent variable. They can also help us understand how certain relationships among items imply certain relationships between items and the latent variable. We

FIGURE 2.2 ● The causal pathway from *X* to *Y*

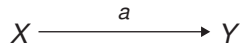


FIGURE 2.3 ● Two variables plus error determine *Y*

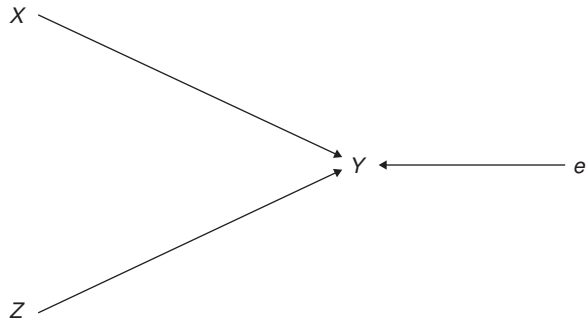
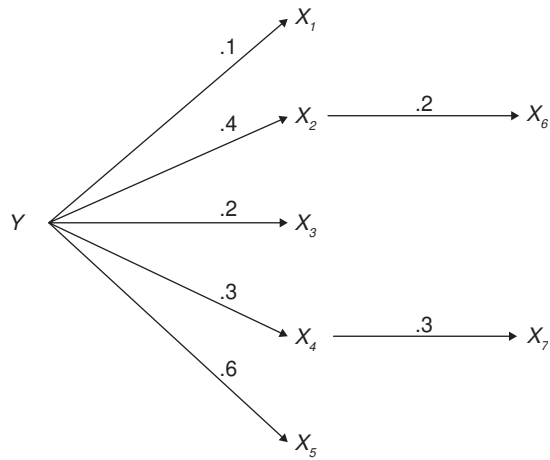


FIGURE 2.4 ● A path diagram with path coefficients, which can be used to compute correlations between variables



begin by examining a simple computational rule for path diagrams. Let us look at the simple path diagram in Figure 2.4.

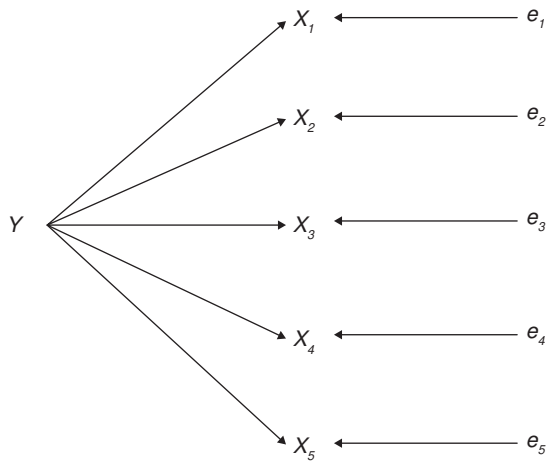
The numbers along the paths are *standardized path coefficients*. Each one expresses the strength of the causal relationship between the variables joined by the arrow. The fact that the coefficients are standardized means that they all use the same scale to quantify the causal relationships and that their values can range from –1.0 to +1.0. In this diagram, Y is a cause of X_1 through X_5. A useful relationship exists between the values of path coefficients and the correlations between the Xs (which would represent items in the case of a scale-development–type path diagram). For diagrams like this one having only one common origin (Y in this case), the correlation between any two Xs is equal

to the product of the coefficients for the arrows forming a route, through *Y*, between the *X* variables in question. For example, the correlation between X_1 and X_5 is calculated by multiplying the two standardized path coefficients that join them via *Y*. Thus, $r_{1,5} = .6 \times .1 = .06$. Variables X_6 and X_7 also share *Y* as a common source, but the route connecting them is longer. However, the rule still applies. Beginning at X_7, we can trace back to *Y* and then forward again to X_6 (or in the other direction, from X_6 to X_7). The result is $.3 \times .3 \times .4 \times .2 = .0072$. Thus, $r_{6,7} = .0072$.

This relationship between path coefficients and correlations provides a basis for estimating paths between a latent variable and the items that it influences. Even though the latent variable is hypothetical and unmeasurable, the items are real and the correlations among them can be directly computed. By using these correlations, the simple rule just discussed, and some assumptions about the relationships among items and the true score, we can come up with estimates for the paths between the items and the latent variable. We can begin with a set of correlations among variables. Then, working backward from the relationship among paths and correlations, we can determine what the values of certain paths must be if the assumptions are correct. Let us consider the example in Figure 2.5.

This diagram is similar to the example considered earlier in Figure 2.4, except that there are no path values, the variables X_6 and X_7 have been dropped, the remaining *X* variables represent scale items, and each item has a variable (error) other than *Y* influencing it. These *e* variables are unique in the case of each item and represent the residual variation in each item not explained by *Y*. This diagram indicates that all the items are influenced by *Y*. In addition, each is influenced by a unique set of variables other than *Y* that are collectively treated as error.

FIGURE 2.5 ● A path diagram with error terms



This revised diagram represents how five individual items are related to a single latent variable, *Y*. The numerical subscripts given to the *e*s and *X*s indicate that the five items are different and that the five sources of error, one for each item, are also different. The diagram has no arrows going directly from one *X* to another *X* or going from an *e* to another *e* or from an *e* to an *X* other than the one with which it is associated. These aspects of the diagram represent assumptions that will be discussed later.

If we had five actual items that a group of people had completed, we would have item scores that we could then correlate with one another. The rule examined earlier allowed the computations of correlations from path coefficients. With the addition of some assumptions, it also lets us compute path coefficients from correlations—that is, correlations computed from actual items can be used to determine how each item relates to the latent variable. If, for example, X_1 and X_4 have a correlation of .49, then we know that the product of the values for the path leading from *Y* to X_1 and the path leading from *Y* to X_4 is equal to .49. We know this because our rule established that the correlation of two variables equals the product of the path coefficients along the route that joins them. If we also assume that the *two path values are equal,* then they both must be .70.$_1$

Further Elaboration of the Measurement Model

Classical Measurement Assumptions

The classical measurement model—which asserts that an observed score, *X*, results from the summation of a true score, *T*, plus error, *e*—starts with common assumptions about items and their relationships to the latent variable and sources of error:

1. The amount of error associated with individual items varies randomly. The error associated with individual items has a mean of zero when aggregated across a large number of people. Thus, items' means tend to be unaffected by error when a large number of respondents complete the items.
2. One item's error term is *not* correlated with another item's error term; the only routes linking items always pass through the latent variable, never through any error term.
3. Error terms are *not* correlated with the true score of the latent variable. Note that the paths emanating from the latent variable do not extend outward to the error terms. The arrow between an item and its error term aims the other way.

The first two assumptions above are common statistical assumptions that underlie many analytic procedures. The third amounts to defining "error" as

the residual remaining after considering all the relationships between a set of predictors and an outcome or in this case, a set of items and their latent variable.

Parallel Tests

Classical measurement theory, in its most orthodox form, is based on the assumption of parallel tests. The term *parallel tests* stems from the fact that one can view each individual item as a "test" for the value of the latent variable. For our purposes, referring to parallel items would be more accurate. However, we will defer to convention and use the traditional name.

A virtue of the parallel tests model is that its assumptions make it quite easy to reach useful conclusions about how individual items relate to the latent variable based on our observations of how the items relate to one another. Earlier, we suggested that, with knowledge of the correlations among items and with certain assumptions, one could make inferences about the paths leading from a causal variable to an item. As will be shown in the next chapter, being able to assign a numerical value to the relationships between the latent variable and the items themselves is quite important. Thus, in this section, I will examine in some detail how the assumptions of parallel tests lead to certain conclusions that make this possible.

The rationale underlying the model of parallel tests is that each item of a scale is precisely as good a measure of the latent variable as any other of the scale items. The individual items are thus *strictly parallel,* which is to say that each item's relationship to the latent variable is presumed identical to every other item's relationship to that variable *and* the amount of error present in each item is also presumed to be identical. Diagrammatically, this model can be represented as shown in Figure 2.6.

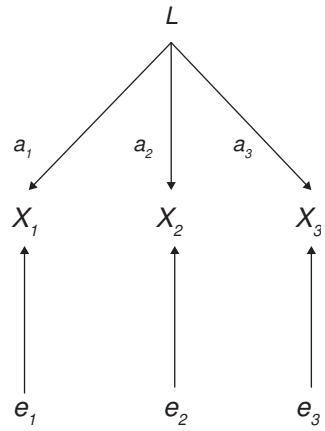
This model adds two assumptions to those listed earlier:

1. The amount of influence from the latent variable to each item is assumed to be the same for all items.
2. Each item is assumed to have the same amount of error as any other item, meaning that the influence of factors *other* than the latent variable is equal for all items.

These added assumptions mean that the correlations of each item with the true score are identical. Being able to assert that these correlations are *equal* is important because it leads to a means of determining the *value* for each of these identical correlations. This, in turn, leads to a means of quantifying reliability, which will be discussed in the next chapter.

Asserting that correlations between the true score and each item are equal requires *both* of the preceding assumptions. A squared correlation is the proportion of variance shared between two variables. So if correlations between the true score and each of two items are equal, the proportions of variance shared

FIGURE 2.6 ● A diagram of a parallel tests model, in which all pathways from the latent variable (*L*) to the items (X_1, X_2, X_3) are equal in value to one another, as are all pathways from the error terms to the items



between the true score and each item also must be equal. Assume that a true score contributes the same *amount* of variance to each of two items. This amount can be an equal *proportion* of total variance for each item only if the items have identical total variances. In order for the total variances to be equal for the two items, the amount of variance each item receives from sources other than the true score must also be equal. As all variation sources other than the true score are lumped together as error, this means that the two items must have equal error variances. For example, if X_1 got 9 arbitrary units of variation from its true score and 1 from error, the true score proportion would be 90% of total variation. If X_2 also got 9 units of variation from the true score, these 9 units could be 90% of the total only if the total variation were 10. The total could equal 10 only if error contributed 1 unit to X_2 as it did to X_1. The correlation between each item and the true score then would equal the square root of the proportion of each item's variance that is attributable to the true score or roughly .95 in this case.

Thus, because the parallel tests model assumes that the amount of influence from the latent variable is the same for each item *and* that the amount from other sources (error) is the same for each item, the proportions of item variance attributable to the latent variable and to error are equal for all items. This also means that, under the assumptions of parallel tests, standardized path coefficients from the latent variable to each item are equal for all items. It was assuming that standardized path coefficients were equal that made it possible, in an earlier example, to compute path coefficients from correlations between items. The path diagram rule relating path coefficients to correlations, discussed earlier, should help us understand why these equalities hold when one accepts the preceding assumptions.

The assumptions of this model also imply that correlations among items are identical (e.g., the correlation between X_1 and X_2 is identical to the correlation between X_1 and X_3 or X_2 and X_3). How do we arrive at this conclusion from the assumptions? The correlations are all the same because the only mechanism to account for the correlation between any two items is the route through the latent variable that links those items. For example, X_1 and X_2 are linked only by the route made up of paths a_1 and a_2. The correlation can be computed by tracing the route joining the two items in question and multiplying the path values. For any two items, this entails multiplying two paths that have identical values (i.e., $a_1 = a_2 = a_3$). Correlations computed by multiplying equal values will, of course, be equal.

The assumptions also imply that each of these correlations between items equals the square of any path from the latent variable to an individual item. How do we reach this conclusion? The product of two different paths (e.g., a_1 and a_2) is identical to the square of either path because both path coefficients are identical. If $a_1 = a_2 = a_3$ and $(a_1 \times a_2) = (a_1 \times a_3) = (a_2 \times a_3)$, then each of these latter products must also equal the value of any of the paths multiplied by itself. Looking back at Figure 2.6 may make these relationships and their implications clearer.

It also follows from the assumptions of this model that the proportion of error associated with each item is the complement of the proportion of variance that is related to the latent variable. In other words, any effect on a given item that is not explained by the latent variable must be explained by error. Together, these two effects explain 100% of the variation in any given item. This is so simply because the error term (e) is defined as encompassing all sources of variation in the item other than the latent variable.

These assumptions support at least one other conclusion: Because each item is influenced equally by the latent variable and each error term's influence on its corresponding item is also equal, the items all have equal means and equal variances. If the only two sources that can influence the mean are identical for all items, then clearly the means for the items also will be identical. This reasoning also holds for the item variances.

In conclusion, the parallel tests model assumes the following:

1. Error is random.
2. Errors are not correlated with one another.
3. Errors are not correlated with true score.
4. The latent variable affects all items equally.
5. The amount of error for each item is equal.

These assumptions allow us to reach a variety of interesting conclusions. Furthermore, the model enables us to make inferences about the latent variable based on the items' correlations with one another. However, the model accomplishes this feat by setting forth fairly stringent assumptions.

Alternative Models

As it happens, all the narrowly restrictive assumptions associated with strictly parallel tests are not necessary in order to make useful inferences about the relationship of true scores to observed scores. A model based on what are technically called *tau-equivalent tests* makes a more liberal assumption—namely, that the amount of error variance associated with a given item need not equal the error variance of the other items (e.g., Allen & Yen, 1979). Tau-equivalent tests still require identical true scores for items, although a slight loosening of that assumption defines *essentially tau-equivalent tests* (or occasionally, *randomly parallel tests*). Any pair of items adhering to essential tau equivalence may have true scores that differ by some constant. Of course, adding a constant to one item has no effect on any correlation involving that item because correlations are standardized expressions. Consequently, the correlation between any pair of items or between an item's true score and the item's obtained score is not affected by relaxing the assumptions of strict tau equivalence to those of essential tau equivalence. So what we have said thus far about tau equivalence also applies to essential tau equivalence. In either of these cases, the *standardized* values of the paths from the latent variable to each item may not be equal. However, the *unstandardized* values of the path from the latent variable to each item (i.e., the *amount* as opposed to *proportion* of influence that the latent variable has on each item) are still presumed to be identical for all items. This means that items are parallel with respect to how much they are influenced by the latent variable but are not necessarily influenced to exactly the same extent by extraneous factors that are lumped together as error. Under strictly parallel assumptions, not only do different items tap the true score to the same degree; their error components are also the same. Tau equivalency (tau is the Greek equivalent to *t*, as in true score) is much easier to live with because it does not impose the "equal errors" condition. Because errors may vary, item means and variances may also vary. The more liberal assumptions of this model are attractive because finding equivalent measures of equal variance are rare. This model allows us to reach many of the same conclusions as with strictly parallel tests but with less restrictive assumptions. Readers may wish to compare this model with Nunnally and Bernstein's (1994) discussion of the domain sampling model.

Some scale developers consider even the essentially tau-equivalent model too restrictive. After all, how often can we assume that each item is influenced by the latent variable to the same degree? Tests developed under what is called the *congeneric model* (Jöreskog, 1971) are subject to an even more relaxed set of assumptions (see Carmines & McIver, 1981, for a discussion of congeneric tests). This model assumes (beyond the basic measurement assumptions) merely that all the items share a common latent variable. They need not bear equally strong relationships to the latent variable, and their error variances need not be equal. One must assume only that each item reflects the true score to some degree. Of course, the more strongly each item correlates with the true score, the more reliable the scale will be.

An even less constrained approach is the *general factor model,* which allows multiple latent variables to underlie a given set of items. Carmines and McIver (1981), Loehlin (1998), and Long (1983) have discussed the merits of this type of very general model, chief among them being its improved correspondence to real-world data. Structural equation modeling approaches often incorporate factor analyses into their measurement models; situations in which multiple latent variables underlie a set of indicators exemplify the general factor model (Loehlin, 1998).

The congeneric model is a special case of the factor model (i.e., a single-factor case). Likewise, an essentially tau-equivalent measure is a special case of a congeneric measure—one for which the relationships of items to their latent variable are assumed to be equal. Finally, a strictly parallel test is a special case of an essentially tau-equivalent one, adding the assumption of equal relationships between each item and its associated sources of error.

Another measurement strategy should be mentioned. This strategy is item response theory (IRT). This approach has been used primarily but not exclusively with dichotomous-response (e.g., correct versus incorrect) items in developing ability tests. IRT assumes that each individual item has its own characteristic sensitivity to the latent variable, represented by an item-characteristic curve—a plot of the relationship between the value of the latent variable (e.g., ability) and the probability of a certain response to an item (e.g., answering it correctly). Thus, the curve reveals how much ability an item demands to be answered correctly. We will consider IRT further in Chapter 8.

In Chapters 6, 7, and 8, we will look at factor analysis, indices, and item response theory respectively. In those chapters, we will necessarily go beyond the models we have discussed so far. In Chapters 1 through 5, however, we will focus primarily on parallel and essentially tau-equivalent models for several reasons. First, they exemplify "classical" measurement theory. Second, discussing the mechanisms by which other models operate can quickly complicate topics unnecessarily if those models are not necessary to a basic understanding. Finally, classical models have proven very useful for social scientists with primary interests other than measurement who, nonetheless, take careful measurement seriously. This group is the audience for whom the present text has been written. For these individuals, the scale development procedures that follow from a classical model generally yield satisfactory scales. Indeed, to my knowledge although no tally is readily available, I suspect that (outside ability testing) a substantial majority of the well-known and highly regarded scales used in social science research were developed using such procedures.

Choosing a Causal Model

Choosing the causal model that underpins a variable, when feasible, can be an important aspect of measurement. The very conceptualization of a variable can sometimes be subtly adapted at the outset of a research project to

make its eventual measurement more manageable. As an example, consider a researcher who wants to assess how the physical work environment affects employee productivity. One approach might be to develop a long list of environmental factors that are thought to influence productivity—such as lighting, sense of privacy, or access to a computer—and develop an instrument that has workers rate the extent to which those factors are present in a given workplace. A problem with this approach is that the instrument may end up being an index rather than a scale or perhaps a hybrid of the two (topics we discuss in Chapter 7). That is, the indicators (e.g., good lighting, reasonable privacy, computer access) might not really share a common cause but rather a common effect, namely, an improvement in the work environment. If, instead, the investigator considered the eventual measurement problem early on in the research process, he or she may have decided to conceptualize the variable somewhat differently. For example, had the investigator defined the variable of interest as employees' perceptions of the work environment, that definition may have led to a more tractable set of items. For example, employees could be asked to endorse items such as, "My workplace environment provides the basic equipment I need to do my job effectively." Here, the latent variable is not a feature of the environment per se but the employees' perceptions. How the employees perceive the environment is the common cause driving their responses to individual items. It may be easier to assume that an employee has a sense of the work environment that will give rise to answers across a set of questions about its adequacy than to imagine the environment itself as a cause of employee responses. Moreover, the psychological nature of employee perceptions may actually be closer to what the investigator considered relevant to productivity than the mere presence or absence of specific environmental features. That is, whether a given worker perceives the environment as conducive to productivity may be a more relevant variable than someone else's judgment regarding the adequacy of the work environment. So conceptualizing the variable of interest in this way may serve the underlying research question well while also potentially facilitating the eventual measurement of the variable.

Of course, if the variable simply does not lend itself to a causal conceptualization consistent with a straightforward measurement strategy, the integrity of the variable of interest should not be compromised. Chapter 7 offers ways to proceed in those instances. Certain approaches may help the investigator work around the limitations inherent in the variable and the way in which it is operationalized. But if an acceptable alternative conceptualization of the variable and the model relating it to its indicators can be simplified, it well may be possible to develop a measurement tool that meets a simpler set of assumptions and thus can be explored using less complex analytic tools. Having the tools to handle the more complex situations is certainly a good thing, but avoiding those complexities and precluding the need for those more advanced tools may be even better, assuming that it does justice to the construct.

Exercises

1. How can we infer the relationship between the latent variable and two items related to it based on the correlations between the two items?
2. What is the chief difference in assumptions between the parallel tests and essentially tau-equivalent models?
3. Which measurement model assumes, beyond the basic assumptions common to all measurement approaches, only that the items share a common latent variable?
4. Assume an essentially tau-equivalent model with true score *T* and indicators A, B, and C. In such a model, any two indicators (e.g., A and B) that share a common true score must have a covariance identical to the covariance between any other two indicators (e.g., B and C) sharing that true score. However, the correlations between different pairs of indicators need not be equal. Explain why this is so.

Note

1. Although -.70 is also an allowable square root of .49, deciding between the positive or negative root is typically of less concern than one would think. As long as all the items can be made to correlate positively with one another (if necessary, by reverse scoring certain items, as discussed in Chapter 5), then the signs of the path coefficients from the latent variable to the individual items will be the same and are arbitrary. Note, however, that giving positive signs to these paths implies that the items indicate more of the construct, whereas negative coefficients would imply the opposite.

3

Scale Reliability

Reliability is a fundamental issue in psychological measurement. Its importance is clear once its meaning is fully understood. As the term implies, a reliable instrument is one that performs in consistent, predictable ways. For a scale to be reliable, the scores it yields must represent some true state of the variable being assessed. In practice, this implies that the score produced by the instrument should not change unless there has been an actual change in the variable the instrument is measuring and, thus, that any observed change in scores can be attributed to actual change in that variable. A perfectly reliable scale would be a reflection of the true score and nothing else. This will seldom be achievable; however, we can gauge how closely we approximate that ideal. The more the score we obtain from a scale represents the true score of the variable and the less it reflects other extraneous factors, the more reliable our scale is. Stated more formally, scale reliability is the proportion of variance attributable to the true score of the latent variable. There are several methods for computing reliability, but all share this fundamental definition.

Although alternative methods for computing scale reliability may appear to be different, the common underlying definition requires that they be computationally equivalent in some basic and important way. This is indeed the case. All these methods involve estimating the variable's true score and determining what proportion of the obtained scale score that true score represents. Our basic measurement model, described in Chapter 2, suggests that a scale's observed score represents the summation of a true score for the variable being assessed plus error arising from extraneous factors. It follows, then, that we can estimate the true score for the variable by subtracting variance arising from error from the total variance of the observed score obtained from a particular measure. We can then compute reliability as a ratio of the estimated true score to the observed score. Thus,

$$\text{true score} = \text{observed score} - \text{error}$$

$$reliability = \frac{true\ score}{observed\ score}.$$

Methods for estimating error are largely what differentiate alternative formulas for computing reliability. Different methods are tailored to specific types of data, although all share a common conceptual foundation: that reliability is the proportion of variance in an observed score that can be attributed to the true score of the variable being assessed.

Methods Based on the Analysis of Variance

One means of estimating error is based on the analysis of variance (ANOVA). This data analytic approach partitions the total variance observed into various sources, primarily those that are of substantive interest (i.e., *signal*) and those that arise from some error source (i.e., *noise*), such as imperfections in sampling participants from a population. Although this is not the approach on which we will focus for assessing the reliability of measurement scales, looking at it in condensed form underscores the continuity across definitions of and approaches to reliability.

Thus, by way of a cursory review, consider a very simple set of observations that involve the temperatures of eight identical objects, four of which are in direct sunlight and four of which are in the shade. (I have specified a small number of observed objects in this example for simplicity.) The objects are identical except for their exposure to the sun; however, the thermometer used to measure their temperatures is a bit suspect and, therefore, is a potential source of error in the observed temperatures. We could assess the extent of that error by recording the temperatures of all eight individual objects and arranging the information in several ways. First, we could summarize information about the objects as a single group by computing an overall sum of squared deviations in object temperatures from the overall mean for all the objects. This value would be the total sum of squares, or SS_T. By dividing the SS_T by the degrees of freedom associated with the entire sample (i.e., $N - 1 = 8 - 1 = 7$), we would obtain the total variance for the objects' temperatures. The following steps isolate subcomponents of that overall variance. We could proceed to estimate the extent to which error affected those scores and, thus, was a subcomponent of the total variance. In the ANOVA framework, this is accomplished by assessing how much variation occurs under identical conditions. In this case, all the objects in the sun are exposed to identical conditions, as are all the objects in the shade. Within each of these two subgroups, the objects themselves are presumed to be identical and the presence or absence of sunlight is identical. So the only basis for differences in *observed* temperatures should be some form of *error*. Thus, we can examine the variation in temperatures of objects within groups to compute an error sum of squares (SS_E). By subtracting

this SS_E from SS_T, we can compute a sum of squares for the effect of sunlight. This last sum of squares is essentially the sum of squares for the true score—that is, an indication of the amount of variation in object temperatures after removing the effect of measurement error. We can then compute the true score variance from this sum of squares. Finally, by computing the ratio between that true score variance and the total variance, we arrive at the proportion of total variance that can be attributed to the *true score* (i.e., the effect of the sun). We could interpret that proportion as the reliability of our measurement of the objects' temperatures.

Note that if all the objects in the sun had identical temperatures and all the objects in the shade had identical, presumably lower, temperatures, the error variance would be 0.0. Thus, nothing would get subtracted from the observed SS_T, the true score variance and total variance would be equal, and the ratio representing the reliability of measuring object temperatures would be 1.0.

I have referred to the ratio arising from the ANOVA example described in the preceding paragraphs as a reliability coefficient, which is correct. More generally, however, a ratio comparing the variance arising from some specific source in an ANOVA design with the total variance is known as an intraclass correlation coefficient, or ICC. Depending on the type and complexity of the ANOVA design, there can be several types of ICC that will have various interpretations, not all of which are equivalent to measurement reliability. Although readers may not be as familiar with the ICC as they are with other, more common expressions for reliability, we will see that the logic on which more specialized indicators of reliability are based is identical to the logic of the ICC—that is, both the ICC and other methods of capturing reliability are based on a comparison of some estimate of true score variance with total variance.

Continuous Versus Dichotomous Items

Although items may have a variety of response formats, we assume in this chapter that item responses consist of multiple-value response options. Dichotomous items (i.e., items having only two response options, such as "yes" and "no," or those having multiple response options that can be classified as "right" versus "wrong") are widely used in ability testing and, to a lesser degree, in other measurement contexts. Examples of dichotomous items include the following:

1. Zurich is the capital of Switzerland. True False
2. What is the value of pi?
   (a) 1.41
   (b) 3.14
   (c) 2.78

Special methods for computing reliability that take advantage of the computational simplicity of dichotomous responses have been developed. General

measurement texts, such as Nunnally and Bernstein's (1994), cover these methods in some detail. The logic of these methods for assessing reliability largely parallels the more general approach that applies to multipoint, continuous scale items. In fact, in some cases, the approach to assessing reliability for multiresponse items is an extension of an earlier approach developed for dichotomous-response items. In the interest of brevity, this chapter will make only passing reference to reliability assessment methods intended for scales made up of dichotomous items. Some characteristics of this type of scale are discussed in Chapter 5.

Internal Consistency

Internal consistency reliability, as the name implies, is concerned with the homogeneity of the items within a scale. Scales based on classical measurement models are intended to measure a single phenomenon. As we saw in the preceding chapter, measurement theory suggests that the relationships among items are logically connected to the relationships of items to the latent variable. If the items of a scale have a strong relationship to their latent variable, they will have a strong relationship to one another. Although we cannot directly observe the linkage between items and the latent variable, we can certainly determine whether the items are correlated to one another. A scale is *internally* consistent to the extent that its items are highly intercorrelated. What can account for correlations among items? There are two possibilities: Either items causally affect each other (e.g., Item A causes Item B), or the items share a common cause. Under most conditions, the former explanation is unlikely, leaving the latter as the more obvious choice. Thus, high inter-item correlations suggest that the items are all measuring (i.e., are manifestations of) the same thing. If we make the assumptions discussed in the preceding chapter (particularly the assumption that items do not share sources of error), we also can conclude that strong correlations among items imply strong links between items and the latent variable. Thus, a unidimensional scale or a single dimension of a multidimensional scale should consist of a set of items that correlate well with one another. Multidimensional scales measuring several phenomena—for example, the Multidimensional Health Locus of Control scales (Wallston et al., 1978)—are really families of related scales; each "dimension" is a scale in its own right.

Coefficient Alpha

Internal consistency is typically equated with Cronbach's (1951) coefficient alpha (α). We will examine alpha in some detail for several reasons. First, it is widely used as a measure of reliability. Second, its connection to the definition of reliability may be less evident than is the case for other measures of reliability (such as the alternate forms methods) discussed later. Consequently, alpha may appear more mysterious than other reliability computation methods to those who are not familiar with its internal workings. Finally, an exploration