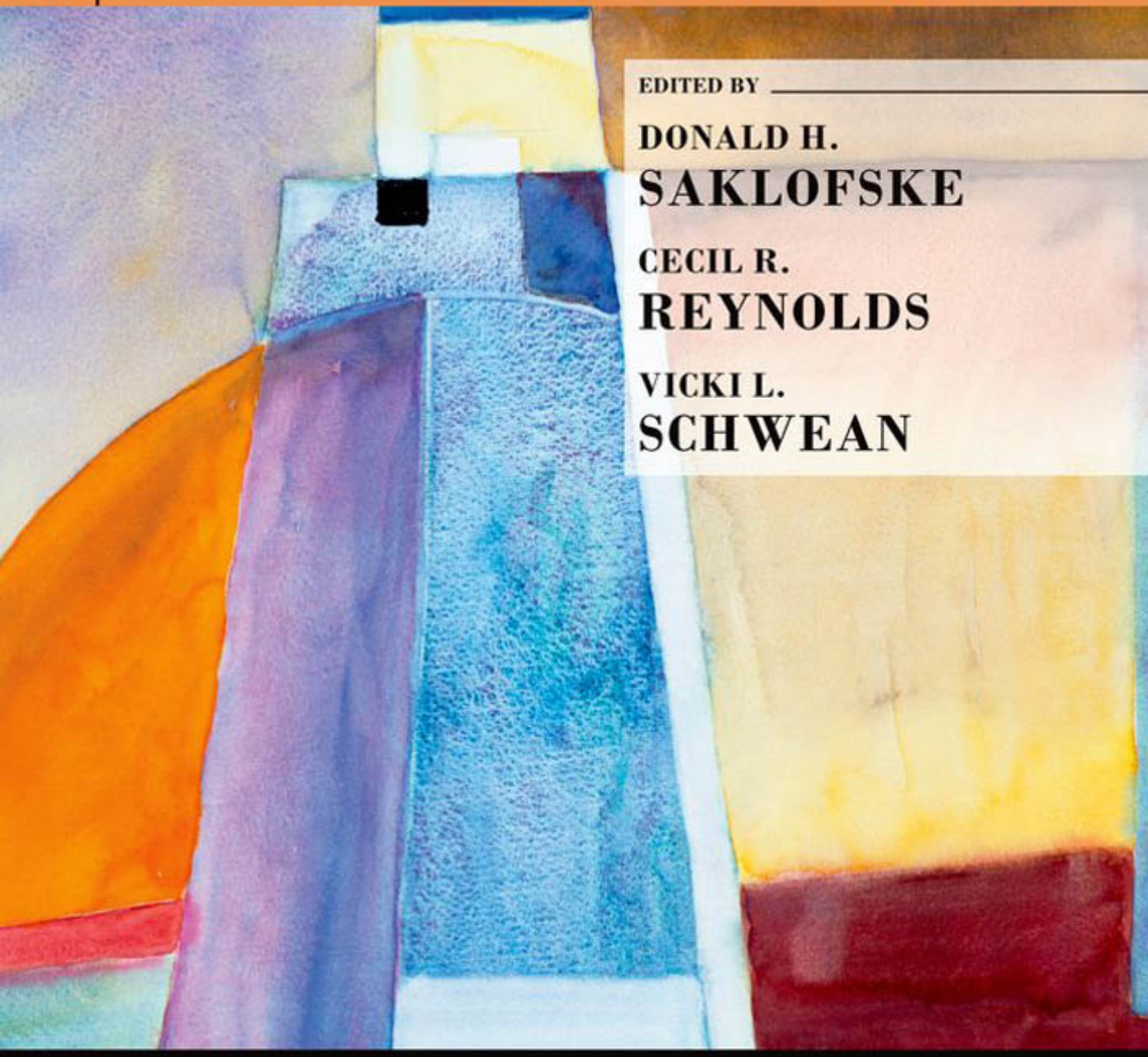
OXFORD LIBRARY OF PSYCHOLOGY



EDITED BY

DONALD H. SAKLOFSKE

CECIL R. REYNOLDS

VICKI L. SCHWEAN

The Oxford Handbook *of* CHILD PSYCHOLOGICAL ASSESSMENT

The Oxford Handbook of Child Psychological Assessment

OXFORD LIBRARY OF PSYCHOLOGY

EDITOR-IN-CHIEF:

Peter E. Nathan

AREA EDITORS:

Clinical Psychology
David H. Barlow

Cognitive Neuroscience
Kevin N. Ochsner and Stephen M. Kosslyn

Cognitive Psychology
Daniel Reisberg

Counseling Psychology
Elizabeth M. Altmaier and Jo-Ida C. Hansen

Developmental Psychology
Philip David Zelazo

Health Psychology
Howard S. Friedman

History of Psychology
David B. Baker

Methods and Measurement
Todd D. Little

Neuropsychology
Kenneth M. Adams

Organizational Psychology
Steve W. J. Kozlowski

Personality and Social Psychology
Kay Deaux and Mark Snyder

OXFORD LIBRARY OF PSYCHOLOGY

Editor in Chief PETER E. NATHAN

The Oxford Handbook of Child Psychological Assessment

Edited by

Donald H. Saklofske

Cecil R. Reynolds

Vicki L. Schwean

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press in the UK and certain other
countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

Library of Congress Cataloging-in-Publication Data
The Oxford handbook of child psychological assessment / edited by Donald H. Saklofske,
Cecil R. Reynolds, Vicki L. Schwean.
p. cm.
ISBN 978–0–19–979630–4
1. Behavioral assessment of children. 2. Behavioral assessment of teenagers. 3. Psychological
tests for children. 4. Child development—Testing. 5. Psychodiagnostics. I. Saklofske,
Donald H. II. Reynolds, Cecil R., 1952– III. Schwean, Vicki L.
BF722.3.O94 2013
155.4028′7—dc23
2012034712

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

SHORT CONTENTS

This page intentionally left blank

OXFORD LIBRARY OF PSYCHOLOGY

The *Oxford Library of Psychology,* a landmark series of handbooks, is published by Oxford University Press, one of the world's oldest and most highly respected publishers, with a tradition of publishing significant books in psychology. The ambitious goal of the *Oxford Library of Psychology* is nothing less than to span a vibrant, wide-ranging field and, in so doing, to fill a clear market need.

Encompassing a comprehensive set of handbooks, organized hierarchically, the *Library* incorporates volumes at different levels, each designed to meet a distinct need. At one level is a set of handbooks designed broadly to survey the major subfields of psychology; at another are numerous handbooks that cover important current focal research and scholarly areas of psychology in depth and detail. Planned as a reflection of the dynamism of psychology, the *Library* will grow and expand as psychology itself develops, thereby highlighting significant new research that will have an impact on the field. Adding to its accessibility and ease of use, the *Library* will be published in print and, later on, electronically.

The *Library* surveys psychology's principal subfields with a set of handbooks that capture the current status and future prospects of those major sub-disciplines. This initial set includes handbooks of social and personality psychology, clinical psychology, counseling psychology, school psychology, educational psychology, industrial and organizational psychology, cognitive psychology, cognitive neuroscience, methods and measurements, history, neuropsychology, personality assessment, developmental psychology, and more. Each handbook undertakes to review one of psychology's major sub-disciplines with breadth, comprehensiveness, and exemplary scholarship. In addition to these broadly-conceived volumes, the *Library* also includes a large number of handbooks designed to explore in depth more-specialized areas of scholarship and research, such as stress, health, and coping; anxiety and related disorders; cognitive development; or child and adolescent assessment. In contrast to the broad coverage of the subfield handbooks, each of these latter volumes focuses on an especially productive, more highly focused line of scholarship and research. Whether at the broadest or most specific level, however, all of the *Library* handbooks offer synthetic coverage that reviews and evaluates the relevant past and present research and anticipates research in the future. Each handbook in the *Library* includes introductory and concluding chapters written by its editor to provide a roadmap to the handbook's table of contents and to offer informed anticipations of significant future developments in that field.

An undertaking of this scope calls for handbook editors and chapter authors who are established scholars in the areas about which they write. Many of the nation's and world's most productive and best-respected psychologists have agreed to edit *Library* handbooks or write authoritative chapters in their areas of expertise.

For whom has the *Oxford Library of Psychology* been written? Because of its breadth, depth, and accessibility, the *Library* serves a diverse audience, including graduate students in psychology and their faculty mentors, scholars, researchers, and practitioners in psychology and related fields. Each will find in the *Library* the information they seek on the subfield or focal area of psychology in which they work or are interested.

Befitting its commitment to accessibility, each handbook includes a comprehensive index, as well as extensive references to help guide research. And because the *Library* was designed from its inception as an online as well as a print resource, its structure and contents will be readily and rationally searchable online. Furthermore, once the *Library* is released online, the handbooks will be regularly and thoroughly updated.

In summary, the *Oxford Library of Psychology* will grow organically to provide a thoroughly informed perspective on the field of psychology, one that reflects both psychology's dynamism and its increasing interdisciplinarity. Once it is published electronically, the *Library* is also destined to become a uniquely valuable interactive tool, with extended search and browsing capabilities. As you begin to consult this handbook, we sincerely hope you will share our enthusiasm for the more than 500-year tradition of Oxford University Press for excellence, innovation, and quality, as exemplified by the *Oxford Library of Psychology.*

Peter E. Nathan
Editor-in-Chief
Oxford Library of Psychology

ABOUT THE EDITORS

Donald H. Saklofske
Don Saklofske is a Professor, Department of Psychology, University of Western Ontario. He is editor of the Journal of Psychoeducational Assessment and Canadian Journal of School Psychology, Associate Editor of Personality and Individual Differences, and editor of the Springer Series on Human Exceptionality. Don is the current president of the International Society for the Study of Individual Differences.

Cecil R. Reynolds
A Distinguished Research Scholar at Texas A & M University, Dr. Reynolds is a Professor of Educational Psychology and a Professor of Neuroscience. He is well known for his work in psychological testing and assessment, and is the author or editor of more than 30 books, including The Handbook of School Psychology, the Encyclopedia of Special Education, and the Handbook of Psychological and Educational Assessment of Children. He also authored the widely used Test of Memory and Learning (TOMAL) and the Revised Children's Manifest Anxiety Scale. He has published a total of more than 300 scholarly works

Vicki L. Schwean
Vicki is currently Professor and Dean of Education, University of Western Ontario.

This page intentionally left blank

CONTRIBUTORS

Wayne Adams
Graduate Department of Clinical Psychology
George Fox University
Newberg, Oregon

Vincent C. Alfonso
Division of Psychological and Educational Services
Fordham University
Bronx, New York

Justin P. Allen
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas

Kathleen Hague Armstrong
Department of Child and Family Studies
University of South Florida
Tampa, Florida

Tiffany L. Arrington
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas

Stephen J. Bagnato
University of Pittsburgh School of Education
Children's Hospital of Pittsburgh
Pittsburgh, Pennsylvania

A. Lynne Beal
Private Practice
Toronto, Canada

Skylar A. Bellinger
Center for Child Health and Development
University of Kansas Medical Center
Kansas City, Kansas

Tanya Beran
Faculty of Medicine
University of Calgary
Calgary, Alberta, Canada

Jonas Bertling
Educational Testing Service
Princeton, New Jersey

Anthony Betancourt
Educational Testing Service
Princeton, New Jersey

Jeremy Burrus
Educational Testing Service
Princeton, New Jersey

Gary L. Canivez
Department of Psychology
Eastern Illinois University
Charleston, Illinois

Jenna Chin
Department of Counseling, Clinical, and School Psychology
University of California at Santa Barbara
Santa Barbara, California

Emma A. Climie
Faculty of Education
University of Calgary
Calgary, Alberta, Canada

Jessica Cuellar
Department of Psychology
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina

Scott L. Decker
Department of Psychology
University of South Carolina
Columbia, South Carolina

Erin Dowdy
Department of Counseling, Clinical, and School Psychology
University of California at Santa Barbara
Santa Barbara, California

Michelle A. Drefs
Faculty of Education
University of Calgary
Calgary, Alberta, Canada

Ron Dumont
School of Psychology
Fairleigh Dickinson University
Teaneck, New Jersey

Agnieszka M. Dynda
St. John's University
Queens, New York

Tanya L. Eckert
Department of Psychology
Syracuse University
Syracuse, New York

Liesl J. Edwards
University of Kansas Medical Center
Center for Child Health and Development
Kansas City, Kansas

Stephen N. Elliott
Learning Sciences Institute
Arizona State University
Tempe, Arizona

Monica Epstein
Department of Mental Health and Law Policy
University of South Florida
Tampa, Florida

Stephen E. Finn
Center for Therapeutic Assessment
Austin, Texas

Meghann Fior
Faculty of Medicine
University of Calgary
Calgary, Alberta, Canada

Erik L. Fister
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas

Dawn P. Flanagan
Department of Psychology
St. John's University
Queens, New York

James R. Flens
Private Practice
Brandon, Florida

Rex Forehand
Department of Psychology
University of Vermont
Burlington, Vermont

Craig L. Frisby
College of Education
University of Missouri
Columbia, Missouri

Mauricio A. Garcia-Barrera
Department of Psychology
University of Victoria
Victoria, British Columbia, Canada

Lauren B. Gentry
The University of Texas at Austin
Austin, Texas

Eugene Gonzalez
Educational Testing Service
Princeton, New Jersey

Jonathan W. Gould
Private Practice
Charlotte, North Carolina

Darielle Greenberg
Private Practice
Richardson, Texas

Matthew J. Grumbein
Leavenworth County Special Education Cooperative
Lansing USD 469
Lansing, Kansas

Ronald K. Hambleton
School of Education
University of Massachusetts Amherst
Amherst, Massachusetts

Jason Hangauer
University of South Florida
Tampa, Florida

Kimberly J. Hills
Department of Psychology
University of South Carolina
Columbia, South Carolina

Susan Homack
Private Practice
Rockwall, Texas

E. Scott Huebner
Department of Psychology
University of South Carolina
Columbia, South Carolina

Deborah J. Jones
Department of Psychology
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina

Diana K. Joyce
University of Florida
Gainesville, Florida

R. W. Kamphaus
Department of Psychology
Georgia State University
Atlanta, Georgia

Belinda N. Kathurima
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas

Alan S. Kaufman
School of Medicine
Yale University
New Haven, Connecticut.

James C. Kaufman
Learning Research Institute
California State University at San Bernardino
San Bernardino, California

Timothy Z. Keith
College of Education
The University of Texas at Austin
Austin, Texas

Ryan J. Kettler
Graduate School of Applied and Professional Psychology
Rutgers University
Piscataway, New Jersey

Sangwon Kim
Ewha Woman's University
Seoul, South Korea

H. D. Kirkpatrick
Forensic Psychologist

Eckhard Klieme
German Institute for International Educational Research
Frankfurt, Germany

Kathryn Kuehnle
Department of Mental Health and Law Policy
University of South Florida
Tampa, Florida

Patrick C. Kyllonen
Educational Testing Service
Princeton, New Jersey

Andrea Lee
School Psychology Program
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina

Jihyun Lee
National Institute of Education
Nanyang Technological University
Singapore

Minji Kang Lee
Psychometric Methods, Educational Statistics, and Research Methods
University of Massachusetts Amherst
Amherst, Massachusetts

Elizabeth O. Lichtenberger
Alliant International University
San Diego, California

Petra Lietz
Australian Council for Educational Research
Melbourne, Australia

Anastasiya A. Lipnevich
Queens College
City University of New York
New York, New York

Stephen W. Loke
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas

Benjamin J. Lovett
Department of Psychology
Elmira College
Elmira, New York

Patricia A. Lowe
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas

Carolyn MacCann
School of Psychology
The University of Sydney
Sydney, Australia

Marisa Macy
Department of Education
Lycoming College
Williamsport, Pennsylvania

David A. Martindale
Private Practice
St. Petersburg, Florida

Nancy Mather
College of Education
University of Arizona
Tucson, Arizona

Laura G. McKee
Department of Psychology
Clark University
Worcester, Massachusetts

Brian C. McKevitt
Department of Psychology
University of Nebraska at Omaha
Omaha, Nebraska

Jennifer Minsky
Educational Testing Service
Princeton, New Jersey

William R. Moore
University of Victoria
Victoria, British Columbia, Canada

Bobby Naemi
Educational Testing Service
Princeton, New Jersey
Jeaveen M. Neaderhiser
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas
Christopher R. Niileksela
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas
Samuel O. Ortiz
Department of Psychology
St. John's University
Queens, New York
Jonathan A. Plucker
Center for Evaluation and Education Policy
Indiana University
Bloomington, Indiana
Jennifer M. Raad
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas
Daniel J. Reschly
Peabody College
Vanderbilt University
Nashville, Tennessee
Cecil R. Reynolds
Department of Education & Human Development
Texas A & M University
College Station, Texas
Matthew R. Reynolds
Department of Psychology and Research in Education
University of Kansas
Lawrence, Kansas
Cynthia A. Riccio
Department of Education and Human Development
Texas A & M University
College Station, Texas
Richard D. Roberts
Educational Testing Service
Princeton, New Jersey
Christina M. Russell
Indiana University
Bloomington, Indiana
Donald H. Saklofske
Department of Psychology
University of Western Ontario
London, Ontario, Canada
W. Joel Schneider
Department of Psychology
Illinois State University
Normal, Illinois
Vicki L. Schwean
Faculty of Education
University of Western Ontario
London, Ontario, Canada
Jessica Oeth Schuttler
University of Kansas Medical Center
Center for Child Health and Development
Kansas City, Kansas
Jill D. Sharkey
The Gevirtz School
University of California, Santa Barbara
Santa Barbara, California
Bennett A. Shaywitz
The Yale Center of Dyslexia and Creativity
Yale University
New Haven, Connecticut
Sally E. Shaywitz
School of Medicine
Yale University
New Haven, Connecticut
Rune J. Simeonsson
School Psychology Program
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina
Steven N. Sparta
UCSD Medical School
Thomas Jefferson School of Law
University of California San Diego
San Diego, California
Kathy C. Stroud
Licensed Specialist in School Psychology
Michael L. Sulkowski
University of Florida
Gainesville, Florida
H. Lee Swanson
Graduate School of Education
University of California-Riverside
Riverside, California
Hedwig Teglasi
Department of Counseling, Higher Education, and Special Education
University of Maryland
College Park, Maryland

Deborah J. Tharinger
The University of Texas at Austin
Austin, Texas
Jennifer Twyford
University of California, Santa Barbara
Santa Barbara, California
Susan M. Unruh
Department of Counseling, Educational Leadership, and Educational & School Psychology
Wichita State University
Wichita, Kansas
Svenja Vieluf
German Institute for International Educational Research
Frankfurt, Germany
John O. Willis
Senior Lecturer in Assessment
Rivier College
Peterborough, New Hampshire
Jonathan Worcester
University of South Florida
Tampa, Florida

This page intentionally left blank

CONTENTS

This page intentionally left blank

PREFACE

Psychological assessment has paralleled the growth of psychology and its specialties since the appearance of the famous Galton tests, the founding of psychology beginning with establishment of Wundt's laboratory, and the successful application of Binet's ability tests. Whether measuring a specific sensory process (e.g., auditory discrimination), broader psychological constructs such as personality (e.g., Big 5), or an observable behavior (e.g., frequency of motor tics) or a latent trait such as intelligence, psychologists have always espoused the importance of measuring the constructs and variables that are the domain of psychological science and using the resulting information as part of the data that can facilitate and enhance decision making in psychological practice. It is not overstating to say that measurement and assessment are the cornerstones of psychology providing the tools and techniques for gathering information to inform our understanding of human behavior.

Precision in every sense of the word is key in psychological assessment. This begins with a description and operational definition of the trait or behavior under examination derived from the theory and research necessary to add empirical support. Following from this foundation is the development of scales that may include various tests (e.g., objective, self report, performance) as well as observation and interview methods to accurately measure (i.e., reliability, validity) the defined behaviors or traits. Standardizing these measures allows for even greater precision in administration, scoring, and interpretation. Data are gathered not only when the test is first published but in follow-up research that further allows for various comparisons of the individual's responses or test scores to normative and criterion interpretations, including change scores whether due to maturation or 'treatment'. Thus psychological measurement addresses the fundamental questions of "how much" and within the context of assessment, contributes to the additional questions of "what and why". Measures are extensions of theory- and research- based findings such that tests developed to measure intelligence are derived from various theories that have received empirical support. In turn, the findings can be used for a variety of 'applied' purposes - to explain, predict and change behavior.

A well used phrase in the measurement/assessment area is, "the more information and the better it is, the better the decision that will be made". Psychologists have created thousands of 'tests' over the past 100 years tapping such key cognitive constructs as intelligence and memory, personality factors such as extraversion and neuroticism, and conative measures including motivation and self efficacy. As psychological knowledge expands, so does the very need to measure and assess these 'new' variables. With the emergence of contemporary models such as emotional intelligence and theory of mind, new measures have quickly followed. Of course it is both theory but also the development of new data analysis techniques

such as structural equation modeling that has allowed us to determine how psychological constructs interact and even moderate or mediate the impact of particular factors on outcomes measures. In turn, this has enhanced the use of 'test batteries' to aid in the psychological assessment of a myriad of human 'conditions' ranging from depression and psychopathy, to learning disabilities and Attention Deficit-Hyperactivity Disorder. 'Clinical' assessment and diagnosis, necessary for determining the selection and application of the most appropriate evidence-based interventions, is grounded in the interface between a complex of key factors (both endogenous and exogenous) that can be obtained from our psychological tests and measures.

Although measurement and assessment are central to psychology, and all science-based disciplines and their resulting practices, psychological tests have been heavily criticized over the years. These criticisms not only come from other disciplines and the general public but also from psychologists themselves. For example, psychological test use has been challenged in the courts and the Response to Intevention (RTI) perspective that has gained momentum in education argues against a reliance on psychological tests for psychological and educational diagnosis. These are but two recent examples of the variability of opinion on assessment. But whether the attack comes from humanistic psychologists or radical behaviorists who might challenge the need for employing tests at all, the fact is that all psychologists engage in "assessment" through the gathering and analysis of data to aid decision making. Counseling psychologists rely heavily on 'talk' to determine a person's needs and issues whereas behaviorists are diligent in observing and measuring overt behaviors (without recourse to proposing underlying hypothetical factors) which can then be used to identify the antecedents and consequences relevant to the behavior in question. Psychoanalytically oriented psychologists may make greater use of projective techniques and free association as the 'data' for guiding their diagnosis and therapy decisions but still are engaged in the assessment process at all stages of their work with clients. These differences 'within' psychology show that assessment is not a static action but an ongoing process that starts with efforts to identify the 'issues' and continues as one observes changes related to everything from life events to therapy outcomes, including the need to reevaluate as new information comes to the fore.

The decisions that need to be made by psychologists can vary from traditional placement, selection and classification to program evaluation, early identification screening, and outcome prediction. Indeed, psychologists engage in a rather amazing array of assessments for many purposes. Psychological assessments and the measurement of various states, traits, and attributes have become valuable because in so many instances they reduce the error rates involved in predictions and decision-making. In fact, psychological assessment and the measurement process are useful only to the extent they can reduce error relative to reliance on other techniques. Determining the correct diagnosis to understand the presenting problems of a client (including the determination that there may be no pathology present), predicting who will be successful in a sales job, who will 'make it' academically in college, or whether medication has been effective in changing behavior, all require a most detailed and comprehensive assessment. In the forensic context, the extent of

functional impairment in a brain injury following a motor vehicle collision, which parent a child of divorce should reside within a custody agreement, and in capital murder cases in some USA states, who is eligible for the death penalty (defendants with intellectual disability cannot be executed) are a few examples of the many predictions and decisions to which psychological test data contribute in meaningful ways.

Another traditional view is that "tests are neutral; it is what we do with them that makes them useful-useless, informative-misleading, 'good-bad', or the like. These viewpoints clearly place psychological assessment in context. Psychological tests that assess the complexities of human behavior just don't appear from nowhere, nor does their use and application automatically follow from simply administering and scoring a test. Assessment employs multimethod-multimodal techniques that rely on scientific knowledge derived from research, theoretical constructs or latent traits and models of human behavior (normal development of social behavior to models of psychiatric classification such as the DSM series). Whatever the 'methods' of assessment, there must be a demonstration of their reliability and validity. This required psychometric support is necessary to weave assessment findings into our psychological knowledge of human behavior that then may lead to prevention and intervention techniques (primary, secondary, tertiary) intended to reduce psychological challenges and promote psychological health and wellness. This process requires a high degree of clinical knowledge and professional competency regardless of one's psychological orientation. Coupled with this is an adherence to the highest professional standards and ethical guidelines.

The editors of this volume are committed to 'best practices in psychological assessment' and while psychological assessment knowledge, techniques, and applications continues to 'improve, we are reassured by a position paper published by Meyer et al (2001) American Psychologist (2001, 56, 128–165) that summarized the literature on psychological assessment. Based on an extensive review of the published literature, it was concluded that: "psychological test validity is strong and compelling, psychological test validity is comparable to medical test validity, distinct assessment methods provide unique sources of information…". It is further stated that: "…a multimethod assessment battery provides a structured means for skilled clinicians to maximize the validity of individualized assessment" and that future investigations should "focus on the role of psychologists who use tests".

A very large literature has addressed the myriad of topics of relevance to psychological assessment. There are a number of journals devoted specifically to this topic including Psychological Assessment edited by Cecil Reynolds and the Journal of Psychoeducational Assessment edited by Don Saklofske. However, the continued growth and new developments in the assessment literature requires an ongoing examination of the 'principles and practices' of central importance to psychological assessment. In particular, the psychological assessment of children and youth has undergone some of the greatest developments, and those developments are the primary focus of this book.

This volume on assessment has been organized primarily, but not exclusively, around clinical and psychoeducational assessment issues. To ensure we are on solid

ground, the foundations that underlie current psychological assessment practices are revisited. For example, the mobility of people has led to major changes in the demographics of countries making cultural issues a major focus in assessment. Linked with these foundations are chapters addressing some of the fundamental principles of child assessment that particularly focus on ability, achievement, behavior and personality. Techniques and specific methods of practice can change rapidly, and we have paired such chapters where possible with the chapters (or sections within a chapter in some cases) from the two previous sections. Theory provides us with guidance in practice when techniques change, new methods are introduced, and new data are presented, as well as when we encounter new presenting issues and circumstances with patients or when asked new questions by referral sources as raised with some specific examples in the fourth section of this volume. A volume on methods that does not also focus on theory is a short-lived work. Here we hope to see theory integrated with research and practice that will enable you to read the chapters in this book, as well as future publications ,not just more profitably but critically as well.

We are especially grateful to all of our authors who wrote the informed and insightful chapters for this volume. Each is an expert who has contributed extensively to psychological assessment research and practice with children and youth and who individually and collectively have made this a book rich in content. While a number of people at Oxford University Press have had a role in this book, we are indebted to Chad Zimmerman, Sarah Harrington, and Anne Dellinger who have provided the necessary guidance and advice that has supported this book from proposal to publication. We also wish to extend our appreciation to Anitha Chellamuthu for guiding this book through the editing phases to publication.

Donald H. Saklofske
Cecil R. Reynolds
Vicki L. Schwean

PART 1

Foundations of Psychological Assessment

This page intentionally left blank

CHAPTER

1 The Role of Theory in Psychological Assessment

Darielle Greenberg, Elizabeth O. Lichtenberger, *and* Alan S. Kaufman

Abstract

This chapter reviews the role of theory in cognitive and neuropsychological assessment from a historical perspective. Theory has been applied to both test development and test interpretation, and it provides a strong framework for valid psychological assessments. Theory-based tests of the twenty-first century such as the Kaufman Assessment Battery for Children—Second Edition (KABC-II), Stanford-Binet Intelligence Test—Fifth Edition (SB-V), Das-Naglieri Cognitive Assessment System (CAS), Woodcock Johnson Test of Cognitive Abilities—Third Edition (WJ-III), and Differential Ability Scales—Second Edition (DAS-II) are highlighted as valid and reliable testing tools. Contemporary methods of test interpretation, including the Cross Battery Assessment approach and the Planning, Attention-Arousal, Simultaneous, and Success (PASS) model of processing, are presented as valid methods of interpretation based on theory. As noted from the chapter's historical perspective, incorporating theory in an assessment helps clinicians synthesize information that is gathered from the evaluation's multiple sources, and ultimately results in more accurate interpretations and interventions.

Key Words: theory, psychological assessment, cognitive, neuropsychological, testing, Cross Battery Assessment, PASS model

For centuries, professionals have been fascinated with the functions of the human body and brain. Attempts to measure brain function, specifically cognitive abilities, date back to 2200 B.C. in China. It is believed that the emperor gave formalized tests to his officers as a way to test for fitness of duty (Kaufman, 2009). With technological advances, significant strides have been made in the area of cognitive abilities and human intelligence. However, controversy regarding the components of these abilities and how to assess them still exists (see, e.g., Flanagan & Harrison, 2012).

The purpose of this chapter is to discuss the role of theory in psychological assessment from a historical perspective. The history is rich and has had an impact on contemporary test development and interpretation. What is meant by "psychological assessment"? Psychological assessment involves a synthesis of the information gathered from several sources, including psychological tests, family history, behavioral observations, and so forth, to understand or make statements regarding an individual's diagnosis, level of functioning, and treatment. Simply administering a test, such as the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV; Wechsler, 2003) or even a theory-based test like the Woodcock-Johnson III (WJ III; Woodcock, McGrew, & Mather, 2001b; Woodcock, McGrew, Schrank, & Mather, 2007) or Kaufman Assessment Battery for Children–Second Edition (KABC-II; Kaufman & Kaufman, 2004a), would be considered psychological testing, and the data collected from multiple other sources in addition to this one test would round

out a complete assessment. Theory has played a significant role in cognitive and neuropsychological assessments, and it is these types of assessments that are the focus of this chapter. Although we acknowledge the usefulness of theory in the development of other types of tools, such as group-administered tests, personality tests, or non-cognitive tests, our particular discussion will center around the role of theory in developing and interpreting tests of cognitive ability. The role of theory in psychological cognitive and neuropsychological assessment is two-pronged. The first prong is the development of tests from theory, and the second is the interpretation of tests from theory.

Historical roots and landmarks

Before describing the modern role of theory in test interpretation and development, a historical review of the period from 1500 to 1970 is warranted. A timeline of historical landmarks in psychological assessment appears in Table 1.1.

Table 1.1 Timeline of Select Historical Landmarks in Psychological Assessment

2200 B.C.	Chinese emperors gave formalized tests to their officials as part of a standardized civil service testing program.
A.D. 1575	Juan Huarte published *Examen de Ingenios* (The Tryal of Wits) in which he tried to demonstrate the connection between physiology and psychology.
1799	Jean-Marc Itard worked to rehabilitate "Victor," a young wild boy found in the woods. Itard assessed differences between normal and abnormal cognitive functioning.
1644	Thomas Willis, an English physician, detailed the anatomy of the brain.
1800	Franz Gall created *phrenology*, or the idea that the prominent bumps on a person's skull determined his personality and intelligence.
1861	Pierre Broca discovered that the speech-production center of the brain was located in the ventro-posterior region of the frontal lobes (now known as "Broca's area").
1874	Carl Wernicke found that damage to the left posterior, superior temporal gyrus resulted in deficits in language comprehension. This region is now referred to as "Wernicke's area."
1837	Edouard Seguin established the first successful school for children with mental retardation.
1838	Jean Esquirol proposed that mental retardation was distinct from mental illness. He suggested that mental disabilities could be categorized into different levels.
1879	William Wundt founded the first psychological laboratory in Germany.
1884	Francis Galton theorized that intelligence was based on sensory keenness and reaction time. He set up a laboratory that used tests to measure these physical and mental abilities.
1888	James McKeen Cattell opened a testing laboratory at the University of Pennsylvania, and his work helped establish mental measurement in the United States.
1904	Charles Spearman proposed a two-factor theory of intelligence that included a general factor (g) and specific (s) factors.
1905	Albert Binet and Theodore Simon developed an intelligence test for screening school-age children.
1909	E. L. Thorndike proposed that intelligence was a cluster of three mental abilities: social, concrete, and abstract.
1917	Robert Yerkes and Lewis Terman developed the Army Alpha and Army Beta, group-administered intelligence tests.
1933	Louis Thurstone used a factor-analytic approach to study human intelligence.

(*continued*)

Table 1.1 (Continued)

1935	Ward Halstead established the first laboratory in America devoted to the study of brain–behavior relationships.
1939	David Wechsler published the Wechsler-Bellevue Intelligence Scale.
1949/1955	David Wechsler published the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Adult Intelligence Scale (WAIS).
1959	J. P. Guilford proposed a Structure of Intellect model of intelligence.
1963	Raymond Cattell and John Horn proposed a theory of crystallized and fluid intelligence, expanding on Cattell's work in 1941.
1979	Alan Kaufman published "Intelligent Testing with the WISC-R," which launched the assessment field into merging theory into test interpretation.
1983	Alan and Nadeen Kaufman published the Kaufman Assessment Battery for Children (K-ABC).
1985	John Horn expanded the G*f*-G*c* model to include ten abilities.
1986	Robert L. Thorndike et al. published the Stanford-Binet—Fourth Edition, which was designed to conform to *Gf-Gc* theory.
1989	Richard Woodcock revised the 1977 Woodcock-Johnson Psych-Educational Battery (WJ, which was not based on theory, to develop the WJ-R, founded on 7 Broad Abilities posited by Horn's *Gf-Gc* theory.
1990	Colin Elliott published the Differential Ability Scale (DAS), which was based on g theory.
1993	John Carroll proposed a three-stratum theory of cognitive abilities, including general ability (level III), broad abilities (level II), and narrow abilities (level I).
1994	J. P. Das, Jack Naglieri, and John Kirby propose the Planning, Attention, Simultaneous, Successive (PASS) theory of intelligence.
1997	Kevin McGrew proposed an integrated Cattell-Horn and Carroll model of cognitive abilities, which was refined by Dawn Flanagan, Kevin McGrew, and Samuel Ortiz in 2000.
1997	Jack Naglieri and J. P. Das published the Cognitive Assessment System (CAS), which is based on the PASS theory of intelligence.
2000	Dawn Flanagan and colleagues developed the Cross-Battery approach to test interpretation.
2001	Woodcock-Johnson–3rd ed. was published, which was based on a CHC theoretical model.
2003	Stanford-Binet–5th ed. was published, which was based on a CHC theoretical model; WISC-IV was published, based on cognitive neuroscience research and theory
2004	Kaufman Assessment Battery for Children–2nd ed. was published, which was based on a dual (CHC and Luria) theoretical model.
2007	Colin Elliott published the Differential Ability Scale—Second Edition (DAS-II), which was based on CHC theory.
2008/2012	Pearson published the latest versions of Wechsler's scales, the WAIS-IV (2008) and WPPSI-IV (2012); all of Wechsler's fourth editions are based on cognitive neuroscience research and theory, especially concerning fluid reasoning, working memory, and processing speed.

Historical Antecedents Before the Nineteenth Century—Juan Huarte, Jean-Marc Itard, and Thomas Willis

Psychological assessment has its roots mainly in the nineteenth century. However, before the 1800s, there were the influential works of men such as Juan Huarte de San Juan and Jean-Marc Gaspard Itard. The sixteenth century was the beginning of the modern era, which brought about economic, political, social, and religious changes. Scientific innovations were booming. In 1575, Juan Huarte, a Spanish physician, published *Examen de Ingenios* (The Tryal of Wits) in which he tried to demonstrate the connection between physiology and psychology. This publication was considered the best-known medical treaty of its time (Ortega, 2005). Huarte believed: 1) Cognitive functions were located in the brain; 2) cognitive functions were innate; 3) human understanding was generative; 4) qualitative differences existed between humans and animals; and 5) language was a universal structure. He also theorized that language was an index of human intelligence and suggested the idea of testing to understand intelligence. Huarte's ideas greatly influenced modern psycholinguistics, organizational psychology, and psychological assessment (Ortega, 2005). Needless to say, his beliefs were revolutionary for his time.

Over two decades later, during the eighteenth century, philosophers and scholars began to question the laws, beliefs, and ideas of the aristocracy. In 1799, the work of Jean-Marc Itard drew public attention for his work with a feral young boy, "Victor," who was found in the woods. Physicians who examined Victor described him as "deaf," "retarded," "a mental defective," and "hopelessly insane and unteachable" (Ansell, 1971; Lane, 1986; Lieberman, 1982). Itard disagreed and believed that Victor's deficiencies were not the result of mental deficiency, but rather due to a lack of interaction with others. For five years, he attempted to "rehabilitate" Victor using an intense education program at the Institute of Deaf Mutes. Itard's aims were to increase his socialization, stimulation, and education. Although Itard was not successful in making Victor "normal," Victor was able to speak and read a few words and follow simple directions. Itard's program was perhaps the first of what we call today an Individualized Educational Program or Plan (IEP).

During these times, physicians were not only responsible for medically examining people like Victor, but they were also in charge of studying and explaining the relationship between brain function and behavior (known today as *neuropsychology*) (Boake, 2008). Although not a physician, Rene Descartes, one of the greatest philosophers, was the first to note that the brain was the most vital organ in mediating behavior. He struggled with understanding and explaining the mind–body connection. After seeing an animated statue of St. Germaine, he theorized that the "flow of animal spirits" through nerves caused the body to move, which led to behaviors (Hatfield, 2007). This theory is known today as the *mechanistic* view of behavior. Descartes believed that although the body and mind interacted, they were, indeed, separate entities.

In 1664, an English physician by the name of Thomas Willis was the first to detail the anatomy of the brain. He is considered to be one of the greatest neuroanatomists of all time and the founder of clinical neuroscience (Molnar, 2004). After studying many patients and dissecting their brains, he described two types of tissue in the brain: gray and white matter. Agreeing with Descartes, he theorized that the white matter was made up of channels that dispersed the "spirits" produced by the gray matter. Willis was also convinced that the brain structures themselves influenced behavior.

Nineteenth-Century Contributions from Brain Research—Franz Gall and Pierre Paul Broca

Around the 1800s, in Austria, physician Franz Gall introduced the idea that the brain was made up of separate organs that were responsible for certain traits, such as memory and aggressiveness. He created *phrenology* or the idea that one could examine the prominent bumps on a person's skull and determine his or her personality and intelligence; a larger brain meant greater intelligence.

Although incorrect about the connection between bumps and intelligence, Gall sparked interest in the area of brain localization (or the idea that specific areas of the brain were responsible for specific functions). As advances in medicine took place, modest progress in understanding human anatomy was made. Prior beliefs had inaccurately attributed behavior to "spirits," while Gall's theories were dismissed as absurd. However, "the field was not ready for behavioral localization" (Maruish & Moses, 1997, p. 34).

After attending a conference, Pierre Paul Broca, a French physician, focused on understanding how brain damage affected people. While working in a hospital, Broca came into contact with a patient who had lost his use of speech, although he could still comprehend language. Because the

patient could only say and repeat the word "tan," he became known as Tan. After Tan died in 1861, Broca performed an autopsy and found a lesion on the left side of the brain's frontal cortex. Other patients like Tan were found to have the same damaged area. From these patients, Broca postulated that the brain's left side of the frontal cortex was responsible for processing language. This region of the brain would later become known as *Broca's area*. Broca's lesion-method, which involved localization of brain function by studying the anatomy of the brain lesion, became an accepted tool for understanding the brain–behavior relationship.

Several years later, German physician Carl Wernicke suggested that not all the functions of language processing were in the area Broca described. During his work on the wards of the Allerheiligen Hospital, he found that patients who sustained damage to or had lesions on the superior posterior portion of the left hemisphere also experienced problems with language comprehension. This area was later named *Wernicke's area*. In 1874, Wernicke published a model of language organization, describing three types of language centers: 1) motor language (damage to this center produced the speech production problems described by Broca); 2) sensory language (damage to this area produced comprehension deficits); and 3) a pathway between these two centers (damage resulted in impairments in repetition) (Mariush & Moses, 1997).

Nineteenth Century Contributions from Research on Mental Deficiency—Jean Esquirol and Edouard Seguin

Along with attention to brain function localization, interest in criminals, mental illness, and mental disabilities (and the differences between them) arose. Thanks to the works of Jean-Etienne Dominique Esquirol and Edouard Sequin, mental disability was no longer associated with insanity (Aiken, 2004). Esquirol theorized that persons with mental illness actually lost their cognitive abilities. In contrast, he determined that those who were called "idiots" never developed their intellectual abilities, and he proposed several levels of mental disability (i.e., morons, idiots, etc). He also believed them to be incurable. Eduardo Sequin was a student of Itard and Esquirol. Sequin disagreed with Esquirol and believed that mental deficiencies were caused by sensory isolation or deprivation and could be mitigated with motor and tactile stimulation (Winzer, 1993). Agreeing with Itard that children with mental disabilities could learn, Sequin expanded Itard's work into three main components: 1) motor and sensory training; 2) intellectual training; and 3) moral training. During the French Revolution, Sequin fled to the United States. He continued his work and established several schools devoted entirely to teaching children with mental retardation. Along with promoting understanding of those who had mental deficiencies, Esquirol and Sequin's work fostered a continued curiosity about intelligence and intelligence testing.

The Birth of IQ Tests in the Late 1800s—Francis Galton and James McKeen Cattell

Western society experienced many changes in culture and technology in the late 1800s. Compulsory-education laws in the United States and Europe and the rise of psychology as a quantitative science were precursors to the introduction and measurement of intelligence (Thorndike, 1997). Before the compulsory-education law, only children whose families came from higher social strata (or who were interested) attended school. The curriculum was set to meet the standards and needs of *these* students. As one can imagine, not everyone was educated. The majority of American society included people and parents who were uneducated or who were unable to speak English (due to the large number of immigrants). Giving access to public education was a way to improve literacy and assimilate immigrants. Thus, the new laws resulted in heterogeneity in the student body and a dramatic increase in student failure rates (Thorndike, 1997). Due to the astonishing failure rates, leaders believed education should not be wasted on those would not benefit, so they devised plans to "weed out" the children who were most likely to fail—intelligence testing was one method.

Along with the educational changes came the rise of psychology as a quantitative science. Gustav Fechner, Herman Ebbinghaus, Sir Francis Galton, and James McKeen Cattell were among the early forerunners who believed mental abilities could be measured (Sattler, 2008; Wasserman, 2012). While Fechner believed he had discovered "the physics of the mind," Ebbinghaus developed a way to empirically study memory and mental fatigue. In England, Sir Francis Galton believed that people were born with a blank slate and that they learned through their senses. He theorized that intelligence was based on sensory keenness and reaction time; so, people who had more acute senses were more intelligent. He developed tests to measure these physical and mental abilities and set up a laboratory in 1884,

which was open to the public. In the announcement of his lab, called the Anthropometric Laboratory, he stated that one of its purposes was to serve "those who desire to be accurately measured in many ways, either to obtain timely warning of remediable faults in development, or to learn their powers" (Sattler, 2008, p. 216). However, the idea of such a laboratory was not a novel one. William Wundt is credited with the establishment of the first psychological laboratory, in Germany in 1879. Galton, with the help of his friend the mathematician Karl Pearson, was also formidable in originating the concepts of *standard deviation*, *regression to the mean*, and *correlation*. Unfortunately, his assumptions and the results of his tests were often not supported by the statistics he developed. Because of his contributions, nevertheless, Galton is often called "the father of the testing movement" (Ittenbach, Esters, & Wainer, 1997).

Galton's assistant, James McKeen Cattell, is responsible for coining the term *mental test* and for bringing Galton's ideas to the United States (Boake, 2002; Ittenbach et al., 1997; Wasserman, 2012). Cattell was interested in studying individual differences in behavior. He believed in the importance of measurement and experimentation and established his own laboratory in Pennsylvania. He developed 50 different measures to assess sensory and motor abilities, although these measures did not differ significantly from Galton's tasks. Important to the history of assessment, Cattell realized the usefulness of tests as a way to select people for training and diagnostic evaluations. As such, he attempted to bring together a battery of tests. Cattell provided us with a standard way to measure human intellectual ability rather than keeping the field of psychology as an abstract discipline (Thorndike, 1997).

The Dawn of the Twentieth Century and the Dynamic Contributions of Alfred Binet

At the end of the nineteenth century, after being publicly embarrassed for his failed work in the area of hypnosis, Frenchman Alfred Binet turned his attention to the study of intelligence. With his two daughters as his subjects, he created and played a series of short games with them. From these encounters, he theorized that intelligence involved more complex mental abilities than just the senses. Binet believed that intelligence was equated with common sense, and called intelligence "judgment . . . good sense . . . the faculty of adapting one's self to circumstances" (American Psychological Association (APA), 2004, p. 1). Binet believed that intelligence was multifaceted and could be measured in three ways: 1) The medical method (anatomical, physiological, and pathological signs of inferior intelligence); 2) the pedagogical method (school-acquired knowledge); and 3) the psychological method (direct observations and measurements of intelligent behavior) (Foschi & Cicciola, 2006). In 1894, he devoted much of his time to researching the mental and physical differences among schoolchildren and became the director of Laboratory of Physiological Psychology in France.

By 1904, Binet was associated with a group of parents and professionals called the Free Society for the Psychological Study of the Child. This group was concerned with school failure rates. The compulsory-education laws in France impacted the government's ability (and private institutions') to provide education to all children. The result was a national system of screening exams for secondary and university education students (Schneider, 1992). The exams did not create a problem for those who advanced, but did for those considered "abnormal" due to their inability to be educated. Children who failed were deemed to belong to one of two categories: 1) Those who could not learn, and 2) those who could learn but would not do so. Those who could not learn were labeled "stupid," while the latter were referred to as "malicious." Binet's involvement with this organization led to his appointment to the French Ministry of Public Instruction, a committee created to identify "abnormal" children. With his main objective to differentiate "normal" children from the "retarded" ones, he created the "metric scale of intelligence" (Schneider, 1992, p. 114). This new approach was *not* to measure sensory or motor reaction times, but rather to measure a child's response to questions. He organized questions based on a series of increasing complexity and assumed that those who answered the more complex questions displayed higher intellectual levels. His original scale, *Measuring Scale of Intelligence*, was introduced in 1905 with the help of Victor Henri and Theodore Simon. The scale comprised 30 items measuring what he believed encompassed intelligence, such as visual coordination, naming objects in a picture, repeating a series of numbers presented orally, constructing a sentence using three given words, giving distinctions between abstract terms, etc. His test was used exclusively to determine whether children needed specialized classes. According to Binet, children who demonstrated intellectual retardation for at least two years were candidates for the classes. Along with the first

IQ test, Binet introduced the important notation of error. He realized that measuring intelligence was not completely accurate and that his tests provided only a sample of an individual's behavior. Binet's original scale and its revisions that followed (1908, 1911, and 1916) "served as both a model of form and source of content for later intelligence tests" (Boake, 2002). The Stanford-Binet Scale (1916) and its 1937 and 1960 revisions became the dominant measures of intelligence in the United States for a half-century.

The Dawn of the Twentieth Century and Charles Spearman's Theory of General Intelligence (g)

The contributions of English psychologist Charles Spearman cannot be overlooked. As a student of Wundt and influenced by Galton, Spearman was intrigued with the concept of human intelligence. While doing his research, he noted that all mental abilities were correlated to each other in some way. He concluded that scores on a mental ability test were similar—a person who performed well on one test would perform well on another (Deary, Lawn, & Bartholomew, 2008). He concluded that intelligence was a general ability that could be measured and expressed as a numerical value. Spearman believed that intelligence was made up of *general ability* or *g,* plus one or more *specific* or *s* factors, and proposed a general-factor or *g* theory. He stated:

> G means a particular quantity derived from statistical operations. Under certain conditions the score of a person at a mental test can be divided into two factors, one of which is always the same in all tests, whereas the other varies from one test to another; the former is called the general factor or G, while the other is called the specific factor. This then is what the G term means, a score-factor and nothing more. . . . And so the discovery has been made that G is dominant in such operations as reasoning, or learning Latin; whereas it plays a very small part indeed in such operation [*sic*] as distinguishing one tone from another . . . G is in the normal course of events determined innately; a person can no more be trained to have it in higher degree than he can be trained to be taller. (Deary et al., 2008, p. 126)

This theory was revolutionary and considered to be the first of many. In 1927, Spearman noted positive correlations (or positive manifold) among cognitive tests explained by psychometric *g* (Reynolds, 2012). When he compared children with normal ability to those with low ability, he observed that correlations were stronger in low ability groups compared to high ability groups. Spearman theorized "as a general rule the effects of psychometric g on test scores decrease as g increases, likening it to the law of diminishing returns from economics" (Reynolds, 2012, p. 3). This phenomenon has become known as Spearman's law of diminishing returns (SLODR). Along with these theories, Spearman refined the use of correlation statistics. Using factor analysis, he improved test reliability by using a correction formula to deal with the errors in his observations that obscured the "common intellective factor" (von Mayrhauser, 1992). Although his theory was criticized, Spearman's use of statistical factor analysis remains an important part of contemporary research and test development.

The Growth of the Binet and Nonverbal Tests in America in the Early Twentieth Century

Along with Binet, other individuals were studying and pursuing the measurement of intelligence. Two men, in particular, were influential pioneers—Henry Goddard and Lewis Terman. Henry Goddard is often considered the first "school psychologist" (Thorndike, 1997). In 1905, he was the director of the Vineland Training School for retarded children and was interested in their unique abilities. Although he wanted to measure the abilities of his students, no measure was available. His search led him to France, where he met Binet. Although he was skeptical, he translated the Binet-Simon scale from French to English and successfully used the scale on his students. In 1908, he introduced an adapted version of the scale, making minor revisions and incorporating standardization (2,000 American children were used). His version was used specifically to evaluate those with mental retardation. While Goddard translated and promoted the Binet-Simon scale, Lewis Terman expanded, standardized, and revised the scale. Terman was responsible for the tentative revision of the Binet-Simon scale in 1912 and the Stanford Revision and Extension of the Binet-Simon Scale in 1916. Terman is also known for renaming the *mental quotient* that Stern developed in 1914. The idea behind this *intelligence quotient* was that the use of a ratio provided a better measurement of mental retardation than the difference between two ages, because the difference did not mean the same thing at different ages (Sattler, 1992).

The beginning of World War I (WWI) initiated the need to evaluate millions of potential American soldiers for "fitness for duty." This seemed an impossible

task, given the number of recruits (some of whom were immigrants) and the fact that the only measures of abilities were based on an individual administration. In 1917, Robert Yerkes and Lewis Terman led a team that developed the group-administered intelligence tests known as Army Alpha and Army Beta (Thorndike, 1997). The Army Alpha was given to the "literate" group, which covered mostly verbal abilities. Army Beta, which involved mostly nonverbal skills, was administered to the "illiterate" group or the group that performed badly on the Army Alpha. The Beta group (composed of mostly immigrants) had more difficulty performing well on the test, resulting in their rejection by the Army to serve as soldiers in WWI.

After the war, a heated debate ensued regarding the validity of the Army testing and the Stanford-Binet. Those involved were outraged about the prejudicial statements of the results of the Army testing, which claimed that individuals from different regions (North vs. South) and of ethnic minorities were inferior (Goddard was largely responsible for questionable interpretation of the test data that led to the racist claims). At the core of the debate was the nature of intelligence, a familiar controversy that began years earlier.

Twentieth-Century Opponents of Spearman's g Theory

Shortly after Spearman's theory was introduced, a debate regarding the nature of intelligence began. Critics believed that Spearman's theory was too simplistic. Thus, in 1909, Edward Lee Thorndike and his colleagues (Lay and Dean) tested the *g* hypothesis and concluded from their analysis, that they were almost tempted to replace Spearman's *g* theory by the equally extravagant theory that "there is nothing whatever common to all mental functions, or to any half of them" (R. M. Thorndike, 1997, p. 11). E. L. Thorndike believed that intelligence was a cluster of three mental abilities: 1) social (people skills); 2) concrete (dealing with things); and 3) abstract (verbal and mathematical skills) (Shepard, Fasko, & Osborne, 1999). While critics like Thorndike continued to question and denounce Spearman's theory, Spearman endlessly sparred with his critics, maintaining that his theory was sound. Although never resolved, this heated debate continued for almost 20 years.

By 1936, the Stanford-Binet was widely accepted in the United States as the standard for measuring intelligence (Roid & Barram, 2004). Finally, Spearman had "proven" his theory. But, much to his chagrin, E. L. Thorndike disagreed again. Thorndike criticized tests similar to Stanford-Binet for measuring only one aspect of intelligence; he continued to insist that intelligence was not a single construct, but much more complex.

Between 1918 and 1938, additional tests (such as Kohs' Block Design Test and the Bender Visual Motor Gestalt Test) were developed and published in response to the debate, but only a few theories (e.g., Thurstone's multiple factor analytic approach) were introduced (Thorndike, 1997). Challenging Spearman's theory, Louis Thurstone (1938) believed that intelligence was not a unitary trait and assumed that intelligence was systematically organized. Using factor analysis, he identified factors including verbal fluency, perceptual speed, inductive reasoning, numeracy, rote memory, deductive reasoning, word fluency, and space or visualization skills. He believed that each factor had equal weight in defining intelligence and labeled these factors *primary mental abilities*.

David Wechsler's Innovations in the 1930s

While many individuals were debating the Army testing issue, David Wechsler was preparing to "reinvent the wheel." Wechsler's contributions to the field of psychological assessment are unmistakable. While waiting to serve in the Army, Wechsler came in to contact with Robert Yerkes. Later, he was the assigned psychologist who administered the Army Alpha and Army Beta to recruits. As he gave the tests, he began to observe the weaknesses of these tools and was determined to use his strong clinical skills and statistical training to develop a new and improved test. Wechsler attributed the misdiagnosis of civilians as having low mental abilities to the heavy emphasis on verbal skills. He hypothesized that if civilians were evaluated on other levels, their abilities would be judged "normal." He believed:

> Intelligence is an aspect of behavior; it has to do primarily with the appropriateness, effectiveness, and the worthwhileness of what human beings do or want to do . . . it is a many-faceted entity, a complex of diverse and numerous components. . . . Intelligent behavior . . . is not itself an aspect of cognition. . . . What intelligence tests measure, what we hope they measure, is something much more important: the capacity of an individual to understand the world about him and his resourcefulness to cope with its challenges. (Wechsler, 1975, p. 135)

When he became chief psychologist at Bellevue Psychiatric Hospital in 1932, he needed a test that could be applied to his population. He stated that the Stanford-Binet scales helped in determining whether an individual had any special abilities or disabilities, but that its application was geared more toward children and adolescents than adults and that the profile interpretation was complicated and unstandardized (Boake, 2002). Creating a standardized measure, statistical in nature, for use with adults was his mission.

In 1939, after a seven-year project, Wechsler introduced his first scale—the Wechsler-Bellevue. He included many tasks from other tests, including the Army Alpha, Army Beta, Army Individual Performance Scale, and Stanford-Binet. He deemphasized previous heavy reliance on verbal skills by introducing nonverbal tasks along with verbal tasks. His selection and development of tasks was based on his belief that intelligence was part of a person's personality and comprised "qualitatively different abilities" (Sattler, 1992, p. 44). "[Wechsler's] aim was not to produce a set of brand new tests but to select, from whatever source available, such a combination of them as would best meet the requirements of an effective adult scale" (Boake, 2002, p. 397). His standardization sample included individuals ranging from seven to 59 years of age who lived in the New York area. By the 1940s, Wechsler's test had gained credibility and was widely used. Wechsler refined and revised his scales until his death in 1981. The scales continue to be modified, even today (as seen by the Wechsler Intelligence Scale for Children–Fourth Edition [WISC-IV], the Wechsler Adult Intelligence Scale–Fourth Edition [WAIS-IV]), and the Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition [WPPSI-IV], although they still remain tied—to some extent—to their original scales. Unlike earlier editions of Wechsler's scales, the fourth editions are based on cognitive neuroscience research and theory, especially within the domains of fluid reasoning, working memory, and processing speed. Furthermore, Wechsler's impact on the contemporary field of assessment remains profound, particularly in transforming the field of intelligence testing psychometric measurement to clinical assessment (Kaufman, in press; Wasserman, 2012).

Mid–Twentieth-Century Contributions from Neuropsychology

While some individuals were emphasizing the concept of intelligence and how it was to be measured, others were interested in the relationship between the brain and behavior. Until the 1930s or so, the field of neuropsychology had been dominated by physicians (Boake, 2008). In 1935, Ward Halstead established the first laboratory in America devoted to the study of the brain–behavior relationship in humans. He was interested in understanding how brain damage affected cognitive, perceptual, and sensorimotor functioning. Because intelligence tests did not help quantify these deficits, he observed the daily activities of several patients and determined that their deficits were varied. Most notable were the loss of adaptive functioning and loss of flexibility of thought. Based on these observations, he compiled a battery of tests to administer in order to understand and examine the deficits. Several years later, Halstead collaborated with his former student Ralph Reitan to develop the Halstead-Reitan Battery. Reitan was responsible for researching and ultimately revising the battery. From his results, he developed indices of brain damage.

In Russia, Alexander Luria worked from a different angle. Luria developed a model of brain organization in which he theorized that brain–behavior relationship could be broken down into components he called *functional systems* (Sbordone & Saul, 2000). He believed that each area of the brain played a specific role in behavior. His theory "was acknowledged as brilliant and insightful, but was seen as forbiddingly complex and impractical for the average clinician" (Hebben & Milberg, 2009, p. 19).

Mid–Twentieth-Century Contributions from Raymond Cattell, John Horn, and J. P. Guilford

The revisions of the Wechsler-Bellevue Scale gave way to the development of additional tests and theories of intelligence between the 1940s and the 1970s. In 1941, Raymond Cattell introduced a dichotomous theory of cognitive abilities. He theorized that there were two types of intelligence—*crystallized* and *fluid* (Horn & Noll, 1997). Crystallized intelligence, G*c*, involved acquired skills and knowledge based on the influences of a person's culture. In contrast, fluid intelligence, or G*f,* referred to nonverbal abilities not influenced by culture.

For two decades, Cattell's theory, and theories in general, were largely overlooked. However, John Horn, a student of Cattell, was responsible for the resurgence and expansion of Cattell's theory, in 1965. Working together and utilizing Thurstone's work, Horn and Cattell theorized that crystallized and fluid intelligence also involved abilities such as

visual processing (G*v*), short-term memory (G*sm*), long-term memory (G*lr*), and processing speed (G*s*). In 1968, Horn added auditory processing (G*a*) and refined the descriptions of other abilities (Flanagan, Ortiz, & Alfonso, 2007). The theory remains in use today as a framework for test developers and approaches to test interpretation (which is discussed later).

In 1967, J. P. Guilford's Structure of Intellect (SOI) became one of the major theories used in the field of intellectual assessment (Kaufman, 2009). Rejecting Spearman's view, Guilford believed that intelligence was composed of multiple dimensions: *operations* (general intellectual processes, such as the ability to understand, encode, and retrieve information), *contents* (how the information is perceived, such as auditory or visual) and *products* (how the information is organized, such as units and classes). This theory was innovative as it implied that there were more types of intelligence (120) than just the *g* described by Spearman. Today, the theory is utilized in the field of learning disabilities and gifted assessments. Linda Silverman, a leading expert in the field of gifted assessment, stated:

> Guilford's model was well received by educators, particularly those who decried the narrowness of some of the older conceptions of intelligence. The concept of a number of intelligences left room for everyone to be gifted in some way. But the model and the methodology have met with severe criticism within the field of psychology. . . . These researchers claim that there is not enough evidence to support the existence of the independent abilities Guilford has described. (Silverman, personal communication, July 8, 2008)

Theory-based tests in the twenty-first century

With an understanding of the historical landmarks, we now turn our attention to the role of theory in test development. Over the past many centuries, our fascination with human cognitive abilities has led to many dramatic developments in the measurement of, and theories related to, intellectual abilities. The links between brain and human behavior and subsequent developments linking neurological pathways and cognitive thought processes have expanded our knowledge of how best to measure human abilities. Physicians, psychologists, researchers, and legislators alike have had a role in shaping psychological assessment. Following this historical path from 2200 B.C. through the end of the twentieth century, we have learned that the field of intellectual assessment is continually evolving. We have highlighted some of the earlier theories related to the assessment of cognitive abilities, and we will now turn to the more modern theories that have shaped both test development and interpretation into the twenty-first century.

To date, the Kaufman Assessment Battery for Children–Second Edition (KABC-II; Kaufman & Kaufman, 2004a), the Stanford Binet, Fifth Edition (SB5; Roid, 2003b), the Cognitive Assessment System (CAS; Naglieri & Das, 1997a), the Woodcock Johnson–Third Edition (WJ-III; Woodcock et al., 2001b; 2007), and the Differential Ability Scales–Second Edition (DAS-II; Elliott, 2007a) are all testing tools that have been based on theory.

Kaufman Assessment Battery for Children—Second Edition (KABC-II)

Drs. Alan and Nadeen Kaufman first introduced their Kaufman Assessment Battery for Children (K-ABC) in 1983. Their philosophy on theory and assessment was innovative and empirically based. The original K-ABC, a measure of intelligence and achievement for children aged 2½ to 12½, significantly differed from traditional tests (including the Wechsler, Woodcock-Johnson, and Stanford-Binet scales) in that it was rooted in neuropsychological theory (i.e., Luria-Das). The scales were divided in two processes: *sequential* and *simultaneous*. Those children who used sequential processing were described as solving problems in a specific, linear order, regardless of content. In contrast, children using simultaneous processing were described as solving problems in a spatial, holistic manner. This aspect of testing had not emerged until the K-ABC, even though theories of intelligence had mentioned the role of the brain. Another essential aspect of the K-ABC was its use with minority children. Cultural bias in psychological assessment has been the subject of longstanding debate among practitioners and researchers. Research had indicated that African-American children performed 15 to 16 points lower on the Wechsler scales than Caucasian children. The K-ABC significantly reduced this difference (by half) and it was said to be "culturally fair."

The KABC-II (Kaufman & Kaufman, 2004a) is used to evaluate the processing and cognitive abilities of children and adolescents aged three to 18 in a clinical, psychoeducational, or neuropsychological setting. It can be also used in conjunction with other assessment tools to identify mental

retardation, intellectual giftedness, and learning disabilities. The KABC-II remains a culturally sensitive tool. Data show that Caucasians and African Americans continue to show reduced differences in global scores relative to other tests of intelligence (Kaufman et al., 2005).

The KABC-II is based on Luria's neuropsychological theory, and also uses the Cattell-Horn-Carroll (CHC) theory. The KABC-II was drastically revised from the original K-ABC. Along with subtest changes and its foundation in a dual theoretical model, the KABC-II gives the examiner the freedom to choose which of two global scores (one based on Luria theory and one based on CHC theory) is the most appropriate one for each person tested—an option that is not afforded by any other assessment tools (Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005). The choice is based on what is best suited to the child's background and reason for referral. For example, if the child is from a bilingual background, the Kaufmans suggest using the Luria Model or MPI (Mental Processing Index). In addition, if the child has or may have a learning disability in reading, they suggest using the CHC Model, which yields the Fluid Crystallized Index (FCI).

A brief review of these two theories is important to the understanding of the theory-based scales (i.e., MPI and FCI). The FCI is the global scale based on the CHC theory and measures general cognitive ability (Kaufman et al., 2005). The CHC theory is a combination of Horn-Cattell's (1968) G*f*-G*c* theory and Carroll's (1993) three-stratum theory (Flanagan, 2000; Schneider & McGrew, 2012). As previously mentioned, Cattell theorized that intelligence was divided into two abilities: fluid and crystallized. Crystallized intelligence or G*c* involved abilities that were acquired through formal education and culture. In contrast, fluid intelligence or G*f* consisted of inductive and deductive reasoning abilities that were influenced by biological and neurological factors. This theory was quite different from the verbal–performance dichotomy used by the Wechsler tests. In 1965, Horn elaborated on the G*f*-G*c* theory to include the following additional cognitive abilities: visual processing (G*v*), short-term memory (G*sm*), long-term memory (G*lr*), and processing speed (G*s*). In later years, he refined G*v*, G*s*, and G*lr* and added auditory processing (G*a*), quantitative knowledge (G*q*) and reading and writing (G*rw*). Horn believed that intelligence was composed of these equally weighted abilities. Figure 1.1 depicts CHC Broad Abilities classifications.

John Carroll's (1993) three-stratum theory is an extension of the G*f*-G*c* theory and other theories. From the results of numerous hierarchical factor-analyses based on correlational data, he theorized that intelligence or cognitive abilities have multiple levels or strata—Stratum I (narrow abilities), Stratum II (broad abilities), and Stratum III (general ability) (Kamphaus, 2008). Stratum I includes specific abilities, such as quantitative reasoning (the ability to reason inductively and deductively), listening ability (the ability to listen and comprehend), and spelling ability (the ability to spell). Stratum II involves the combinations of narrow abilities that form broader abilities, such as crystallized intelligence, fluid intelligence, quantitative knowledge, and so forth. For example, the broad ability crystallized intelligence or G*c* refers to the acquired knowledge based on formal education and culture. The narrow abilities of G*c* include skills such as language development, lexical knowledge, listening ability, and general (verbal) information. Finally, Stratum III encompasses general ability, or what have been labeled "general intelligence."

In contrast to the FCI, the MPI is a global scale based on Luria's model. This scale measures mental-processing ability (a child's ability to solve problems) and excludes language ability and word knowledge (Kaufman et al., 2005). As the name suggests, the Luria model is based on the work of Luria in the 1970s. Luria believed that the brain's basic functions are represented in three "blocks"—Block 1 being arousal and attention, and corresponds to

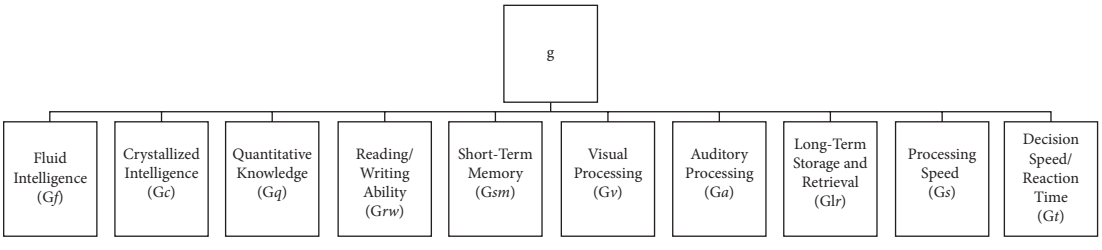


Figure 1.1 CHC Broad Abilities (Stratum II).

the reticular activating system; Block 2 being analyzing, coding, and storing information, and corresponds to the occipital, parietal, and temporal lobes; and Block 3 being executive functions, planning, and programming behavior, and corresponds to the anterior portion of the frontal lobes. Luria also believed that these "blocks" must work together in order for new material to be learned effectively. After information enters the brain, Block 2 is responsible for sending that information to Block 3. Realizing the importance of these systems, the Kaufmans included subtests measuring auditory and visual synthesis (such as requiring a child to point to a series of pictures in the correct order, corresponding to a series of words given by the examiner), as well as subtests that measure simultaneous processing that require use of Block 2 and Block 3 (such as requiring a child to point to a picture that does not go with the others around it).

The KABC-II was standardized on a sample of 3,025 children, stratified according to 2001 U.S. Census data. Reliability and validity data provide support for the psychometric properties of the test. Current literature, along with the test manual, indicates that it is a stable tool (Kaufman & Kaufman, 2004a; Kaufman et al., 2005). Internal consistency coefficients range from .69 to .97, test-retest coefficients range from .74 to .95, and validity coefficients range from .15 to .91. Like the K-ABC, the KABC-II is useful for evaluating minority children. The structure of the tool includes 18 subtests (such as copying the examiner's exact sequence of taps on the table with fist, palm, or side of the hand; assembling several blue and yellow triangles to match a picture of an abstract design; etc.) and yields one to five scales, depending on the child's age and interpretive approach used. For example, at age levels 7 to 18, ten core tasks are administered, yielding either MPI or FCI, either four scales (MPI) or five scales (FCI), and the Planning/G*f* scale.

Fletcher-Janzen and Lichtenberger (2005) commented on the KABC-II's strengths and weaknesses in the areas of test development, administration and scoring, and test interpretation. In terms of test development, the KABC-II has several strengths and weaknesses. Its strengths include the following:

(1) It is based on dual theoretical models (Luria and CHC);

(2) It allows evaluators to choose the theoretical model;

(3) It evaluates a wide range of children and adolescents (ages 3–18);

(4) It allows evaluators to understand cognitive abilities in the context of academics, as it is normed with Kaufman Test of Educational Assessment–Second Edition (KTEA-II; Kaufman & Kaufman, 2004b);

(5) Its norms reflect a sample of ethnic minority responses (approximately 66%);

(6) It has ample floors and ceilings on nearly all subtests;

(7) It permits an evaluator to accept a correct response, regardless of the mode of communication (signing, writing, Spanish, etc.);

(8) The materials are well organized, sturdy, and novel; and

(9) It gives out-of-level norms for evaluating young children who might meet floors and ceilings too soon (Fletcher-Janzen & Lichtenberger, 2005).

In contrast, the KABC-II has several weaknesses in the area of test development, including the following: 1) Does not measure auditory processing (G*a*) and processing speed (G*s*); 2) record forms are complex; and 3) bonus points are used on three subtests, which confounds the measures (Fletcher-Janzen & Lichtenberger, 2005).

In terms of strengths of administration and scoring, the KABC-II:

(1) contains sample and teaching items that can be given in the child's native language,

(2) allows the examiner to explain items in child-specific language if the child does not understand,

(3) has short, simple instructions,

(4) has limited subjective scoring items,

(5) contains subtests that are presented in both visual and auditory forms, and

(6) has a supplemental computer scoring and interpretation software.

Weaknesses include the following: 1) Scoring on some subtests requires special attention to avoid clerical errors; 2) discontinue rules are not consistent from subtest to subtest; and 3) some children may have difficulty understanding the grammar items on Rebus (Fletcher-Janzen & Lichtenberger, 2005).

For the KABC-II, interpretation strengths are as follows:

(1) Luria and CHC models are the foundation;

(2) Use of the CHC model works well for cross-battery assessment;

(3) Interpretation is dependent on global scales and scale indexes;

(4) The interpretation system provides the evaluator with a continuous prompt to check hypotheses with other evidence;

(5) The manual provides mean MPI and FCI, scale index, and subtest scores for ethnic minority groups;

(6) Record form provides room to note basic analysis and strengths and weaknesses;

(7) Out-of-level norms are available for gifted and lower functioning children;

(8) Allows assessment of immediate and delayed memory;

(9) Allows assessment of learning and crystallized knowledge;

(10) A nonverbal index can be calculated and interpreted for children who have difficulty with oral communication (Fletcher-Janzen & Lichtenberger, 2005).

Interpretation weaknesses include: 1) the Knowledge/G*c* subtests do not allow evaluators to assess expressive language, and 2) some comparisons cannot be made because of age limits on some subtests, namely "Story Completion" and "Rover" (Fletcher-Janzen & Lichtenberger, 2005).

In one study that investigated the KABC-II's consistency with the CHC theory, Matthew Reynolds and his colleagues used the standardized sample (ages 3–18) as their participant pool (Reynolds, Keith, Fine, Fisher, & Low, 2007). Multiple-sample analyses were performed. Results showed the KABC-II measures the same construct across all ages. In addition, for school-age children, the test generally matches the five CHC broad abilities it is proposed to measure. The test provides a "robust measure of *g* and strong measures of G*c*, G*v*, G*lr*, and G*sm*, and both *g* and the broad abilities are important to explaining variability in subtest scores" (Reynolds et al., 2007, p. 537). However, some inconsistencies were found in Gestalt Closure, Pattern Reasoning, and Hand Movements. The subtest Gestalt Closure appeared to measure crystallized intelligence (G*c*) in addition to, or perhaps instead of, visual processing (G*v*). The subtest Pattern Reasoning appeared to measure visual processing (G*v*) in addition to fluid reasoning (G*f*). Finally, the subtest Hand Movements measures fluid reasoning (G*f*) in addition to short-term memory (G*sm*). In terms of clinical applications, Fletcher-Janzen and Lichtenberger (2005) report that the KABC-II is effective for individuals who are deaf or hard of hearing, autistic, have speech and language disorders, mental retardation, ADHD, and learning differences.

Sex differences in cognitive abilties in children ages 6 to 18 have been found for the KABC-II (Reynolds, Keith, Ridley, Patel, 2008). In this study, multi-group higher-order analysis of mean and covariance structures (MG-MACS) and multiple indicator-multiple cause (MIMIC) models were used on the standardization sample. Results indicated that boys showed a mean advantage in latent visual-spatial ability (*Gv*) at all ages and in latent crystallized ability (*Gc*) at ages 6 to 16. In contrast, girls scored higher on the latent, high-order *g* factor, at all ages, but these results were statistically significant at only ages 6 to 7 and 15 to 16.

Researchers have investigated the application of other theories as they relate to the structure of KABC-II (Reynolds, Keith, & Beretvas, 2010; Reynolds & Keith, 2007). For example, in one study, Reynolds and Keith (2007) used the standardization sample for ages 6 to 18 to confirm the presence of SLODR. Confirmatory factor analysis was performed. Results indicated that SLODR was present, and "its presence was not dependent on the hierarchical model of intelligence. Moreover, [the] findings suggest that SLODR acts on *g* and not on the broad abilities" (Reynolds & Keith, 2007, p. 267). In another study by Reynolds et al. (2010), a factor mixture model was performed on the standardization sample to eliminate the previous division of participants into separate groups. The results also offered support for SLODR, "most notably the *g* factor variance was less in high *g* mean classes" (Reynolds et al., 2010, p. 231). Reynolds (2012) stated that although the presence of SLODR has been detected in several batteries, its effects on "the measurement of intelligence and interpretation of test scores is less well-understood." (p. 4).

For the most up-to-date research summaries of the KABC-II, consult Reynolds et al. (2010) and Singer, Lichtenberger, Kaufman, Kaufman, and Kaufman (2012).

The Stanford Binet, Fifth Edition (SB5)

Along with the KABC-II, the Stanford Binet is another theory-based assessment tool. Its history is long, dating back to Binet and Simon in 1905. The Stanford Binet–Fifth Edition (SB5; Roid, 2003b) is based on a five-factor hierarchical cognitive model, a combination of theories developed by Carroll, Cattell, and Horn now known as the CHC model (Roid & Barram, 2004). Roid retained the theory that *g* comprises verbal and nonverbal abilities. The SB5 is the first intellectual battery to cover five cognitive factors: fluid reasoning, knowledge,

quantitative reasoning, visual-spatial reasoning, and working memory, in both domains (verbal and nonverbal). Therefore, the SB5 yields a Full Scale IQ, Verbal IQ, Nonverbal IQ, plus the five factor indexes on each domain (Roid & Barram, 2004).

The SB5 is designed to assess an individual's general intellectual ability between the ages of three and 85 and above. It was standardized and stratified on a large sample (N = 4,800; ages 2–96) based on the 2001 U.S. Census data. Reliability and validity data provide support for the psychometric properties of the test (Roid & Barram, 2004). For example, internal consistency coefficients range from .90 to .98.

As the SB5 is a fairly new instrument, researchers need more time to explore it. However, strengths and weaknesses have emerged in terms of test development and standardization, administration and scoring, and test interpretation and application (Roid & Barram, 2004). In terms of test development and standardization, the SB5 has the following strengths: 1) Large norm sample; 2) large age range; 3) in-depth field testing and fairness reviews; 4) content-validity studies of CHC aligned factors; 5) use of item response theory; and 6) linkage with Woodcock-Johnson III Tests of Achievement (Roid & Barram, 2004). In contrast, weaknesses include: 1) it does not assess all CHC model factors; 2) it does not include many clinical and/or special group data; and 3) it correlates with only the WJ-III Achievement (Roid & Barram, 2004).

The SB5 has many strengths and weaknesses in terms of administration and scoring. Its strengths include the following:

(1) Levels are tailored to the examinee's ability;
(2) Scoring metrics are similar to other batteries;
(3) It is a child-friendly test;
(4) New Change-Sensitive Scores are used;
(5) IQ score levels have been extended on both extremes (10 to 40 and 160 to 225);
(6) Record forms are well-designed;
(7) Helpful examiner pages are included in item books; and
(8) There is an optional computer-scoring program that is easy to use (Roid & Barram, 2004).

In contrast, administration and scoring weaknesses involve the following: 1) levels may be confusing to evaluators; 2) shifting between subtests may be difficult for evaluators; 3) extended IQs are only available for Full Scale IQ; 4) nonverbal subtests do not have pure pantomime administration; 5) computer-scoring program is not included with the kit; and 6) nonverbal knowledge may need expressive language skills (Roid & Barram, 2004).

In terms of interpretation and application, the SB5 has strengths including the following:

(1) The assessment of working memory improves diagnoses,
(2) The contrast between verbal and nonverbal subtests is useful,
(3) A comprehensive interpretive manual is included,
(4) Progress can be noted by using Change-Sensitive Scores,
(5) Early prediction of learning disabilities can be made by using Working Memory, Knowledge, and Quantitative Reasoning scores, and
(6) Extended IQs are used for assessment of giftedness and mental retardation (Roid & Barram, 2004).

The weaknesses in this area are: 1) nonverbal subtests require receptive and expressive language skills and 2) more studies of classroom application are needed (Roid & Barram, 2004).

Canivez (2008) investigated the SB5's link to theory by conducting orthogonal higher-order factor structure of the test. His participants included the three youngest age groups from the original standardization sample (N = 1,400 2–5-year-olds; 1,000 6–10-year-olds; and 1, 200 11–16-year-olds. The results of the study indicated that the SB5 "fundamentally measures general, global intelligence (Stratum III; Carroll, 1993). When examining the 10 SB5 subtest correlation matrices for the three youngest age groups, there was no evidence to suggest the presence of more than one factor as proposed by Roid. . . . No evidence of a five factor model was found" (Canivez, 2008, pp. 538–539).

Investigators have also looked into the effectiveness of the SB5 in assessing giftedness, autism spectrum disorders, preschool children, attention-deficit/hyperactivity disorder, autism, and working memory (e.g., Canivez, 2008; Coolican, Bryson, Zwaigenbau, 2008; Leffard, Miller, Bernstien, DeMann, Mangis, & McCoy, 2006; Lichtenberger, 2005; Minton & Pratt, 2006; Newton, McIntosh, Dixon, Williams, & Youman, 2008). Coolican and colleagues (2008) investigated the utility of the SB5 on children with autism spectrum disorders. Their participants included 63 children (12 girls, 51 boys) with a diagnosis of autism, Asperger's syndrome, and pervasive developmental disorder not otherwise specified (PDDNOS). Ninety percent of the children completed the SB5. Their results revealed a broad range of functioning;

individuals earned Full Scale IQs (FSIQs) ranging from 40 to 141. In addition, a higher percentage of children had stronger nonverbal skills than verbal skills. Minton and Pratt (2006) tested 37 students in grades two through six in Idaho. They concluded that elementary school students who were gifted or highly gifted scored significantly lower on the SB5 than on the WISC-III. This result suggests that using the two or three standard deviations from the mean as a cutoff for giftedness vs. nongiftedness was too high.

For the most up-to-date research summaries of the SB5, consult Roid and Pomplin (2012).

THE WOODCOCK JOHNSON–THIRD EDITION (WJ-III)

Although the original Woodcock Johnson was not theory-based, the Woodcock-Johnson Psycho-Educational Battery–Revised (WJ-R, Woodcock & Johnson, 1989) was grounded in Horn-Cattell theory. The latest revision, the Woodcock-Johnson III Tests of Cognitive Abilities (WJ III COG; Woodcock, Johnson, & Mather, 2001b; Woodcock et al., 2007), is based on Cattell-Horn-Carroll (CHC) theory and is designed to measure intellectual abilities of individuals from age five to 95. All three levels (stratum I, II, and III) of the CHC theory are represented on the WJ III, although the primary purpose of the test is to accurately measure broad CHC factor scores (Stratum II) (Schrank, Flanagan, Woodcock, & Mascolo, 2002). Sanders, McIntosh, Dunham, Rothlisberg, and Finch (2007) noted, "Creating tests that measured the CHC abilities allowed for analysis of within-individual variability and provided additional ipsative interpretative information" (p. 120).

The WJ III COG is divided into two major components—the standard battery and the extended battery. For the standard battery, seven cognitive factors, including comprehension-knowledge, long-term retrieval, visual-spatial thinking, auditory processing, fluid reasoning, processing speed, and short-term memory, are assessed along with the general intellectual ability. Three additional cognitive performance cluster scores can be calculated, which include verbal ability, thinking ability, and cognitive efficiency. Not all 20 tests are administered, but rather those subtests that are relevant to information the examiner seeks, as well as to the referral question. For example, the first seven subtests are administered for general intellectual ability. However, if an evaluator is wondering about an individual's short-term memory (G*sm*), the two additional subtests can be administered (Schrank et al., 2002).

A useful aspect of the WJ III COG is that it has been normed with the WJ III Tests of Achievement (WJ III ACH; Woodcock, McGrew, & Mather, 2001a) so that examiners can compare cognitive and achievement abilities. The WJ III ACH consists of 22 subtests that evaluate five areas, including reading, oral language, mathematics, written language, and academic knowledge (such as science). Like the WJ III COG, the WJ III ACH is divided into a standard and extended battery. Administering nine subtests will allow for a Total Achievement score to be obtained for children age five or older. Both the WJ III COG and WJ III ACH were normatively updated in 2007, which involved a recalculation of norms for subtests and clusters (Schrank & Wendling, 2012).

The WJ III COG was standardized on a sample of 8,818 ranging from age two to 95+ and selected from more than 100 "geographically and economically diverse communities" (Schrank et al., 2002). The psychometric properties of the test indicate a stable tool (McGrew & Woodcock, 2001; Schrank et al., 2002). For the standard battery, individual test reliabilities range from .81 (Test 3: Spatial Relations) to .94 (Test 5: Concept Formation). For the extended battery, individual test reliabilities range from .74 (Test 19: Planning) to .97 (Test 18: Rapid Picture Naming). Median cluster reliability statistics range from .88 (Short-Term Memory) to .98 (General Intellectual Ability–Extended). Test-retest reliability coefficients range from .73 to .96. Convergent and discriminate validity coefficients range from .20 to .60.

The WJ III COG has several strengths and weaknesses. Its strengths include the following: 1) the battery is based on empirically strong theory of cognitive abilities; 2) interpretation of its results offers important information regarding cognitive strengths and weaknesses; 3) it is conformed with the WJ III Tests of Achievement and provides actual discrepancy norms; 4) the tool is technically stable; and 5) the materials are well made. In contrast, its weaknesses include complexity of administration and interpretation, lack of hand scoring abilities, and the need for additional research for the clinical clusters (Schrank et al., 2002).

One illustrative study that linked the WJ III COG to CHC theory was conducted by Taub and McGrew (2004), who performed confirmatory factor analysis of the battery and determined its cross-age invariance. The WJ III COG standardization sample served as the data for this study. Three sets of confirmatory factor analyses were

performed. Results of the analyses provide support for the factorial invariance of the WJ COG when the 14 tests contributing to the calculation of the examinee's GIA and CHC factors scores are administered. Support is provided for the WJ III COG theoretical factor structure across five age groups (ages 6 to 90+) (Taub & McGrew, 2004, p. 72). Researchers have also looked into the effectiveness of the WJ III COG in assessing learning disabilities and attention problems (Leffard, Miller, Bernstien, DeMann, Mangis, & McCoy, 2006; Schrank et al, 2002). Schrank and his colleagues (2002) delineate several WJ III discrepancy procedures to assist in identifying specific learning disabilities, including ability-achievement, predicted achievement/achievement discrepancy, general intellectual ability/achievement discrepancy, oral language ability/achievement discrepancy, intra-ability discrepancy, intracognitive discrepancy, intra-achievement discrepancy, and intra-individual discrepancy.

Using the Woodcock-Johnson and Kaufman tests, Scott Barry Kaufman and his colleagues (2012) investigated whether cognitive *g* and academic achievement *g* are the same as the conventional *g* (Kaufman, Reynolds, Liu, Kaufman, & McGrew, 2012). From previous research, we know that IQ-achievement correlations are moderate to high, but that 50 to 75 percent of the variance in achievement is unaccounted for by cognitive ability. Many factors have been found to impact academics. Some of the variance is measurement error, whereas other variance is accounted for by such factors as student characteristics, school environments, and curriculum. They used two large nationally representative data sets and two independent test batteries. Second-order latent factor models and multi-group confirmatory factor analysis were used. The results indicated that *COG-g* and *ACH-g* are not the same as *g*. They are distinct but highly related constructs. And, importantly, Kaufman et al. (2012) gave strong support to the CHC theory-based structure of both the KABC-II and KABC-II).

For the most up-to-date research summaries of the WJ III COG and its 2007 normative update, consult S. B. Kaufman et al. (2012), Schneider and McGrew (2012), and Schrank and Wendling (2012).

THE DIFFERENTIAL ABILITY SCALES—2ND EDITION

The Differential Ability Scales (DAS; Elliott, 1990a) was developed by Colin Elliott from their predecessor, the British Ability Scales (BAS; Elliott, 1983a, 1983b), to focus on "specific abilities rather than on 'intelligence'" (Elliott, 1997, p. 183). The second revision of the Differential Ability Scales (DAS-II; Elliott, 2007a) was designed to "address processes that often underlie children's difficulties in learning and what scientists know about neurological structures underlying these abilities" (Dumont, Willis, & Elliott, 2009, p. 5). The theoretical underpinning of the tool is not based on a single theory, but it has been connected to various neuropsychological processing models and the Cattell-Horn-Carroll theory, which has already been described above (Stavrou & Hollander, 2007).

The DAS-II is designed to evaluate children from the ages of two to 17. The test consists of 20 subtests and is divided into two overlapping age-level batteries—the Early Years (2:6–6:11) and the School Years (5:0–17:11). The Early Years battery is even further divided into a lower (2:6–3:5) and upper level (3:6–6:11). The battery yields an overall composite score labeled General Conceptual Ability (GCA), as well as several additional cluster scores, including Verbal Ability, Nonverbal Reasoning, and Spatial Ability. The Verbal Ability Cluster is a measure of crystallized intelligence or G*c*, the Nonverbal Reasoning Cluster is a measure of fluid intelligence or G*f*, and the Spatial Ability Cluster is a measure of visual-spatial ability or G*v*. These clusters make up the core subtests. Other subtests, known as the diagnostic subtests, measure memory skills, processing speed, and school readiness.

The DAS-II was standardized and normed on 3,480 children living in the United States based on the October 2002 census. The psychometric properties of this tool indicate that it is a stable tool (Dumont, Willis, & Elliott, 2009; Stavrou & Hollander, 2007). Average internal consistency reliabilities range from .77 to .95. Test-retest reliability coefficients range from .83 to .92. In addition, the DAS-II has satisfactory concurrent validity (Dumont, Willis, & Elliott, 2009; Stavrou & Hollander, 2007). Mean overall correlation was .80.

Regarding the strengths and weaknesses of the DAS-II, the strengths include but are not limited to the following:

(1) The General Conceptual Ability Score;
(2) the Special Nonverbal Composite;
(3) ability to administer the nonverbal subtests in Spanish and American Sign Language;
(4) evaluation of differential abilities;
(5) use of Cattell-Horn-Carroll theory;
(6) fairly easy administration and scoring;
(7) child-centered;

(8) diagnostic subtests and clusters; and
(9) ability to evaluate learning differences (Dumont, Willis, & Elliott, 2009).

In contrast, weaknesses include but are not limited to the following: 1) norming that only extends to 17 years, 11 months; 2) it is a test of cognitive ability, not an IQ test; 3) it is a complex test that requires training; and 4) additional testing is required to understand the expressive language skills for younger children (Dumont, Willis, & Elliott, 2009).

Timothy Keith and colleagues have conducted confirmatory factor analyses of the DAS-II across age levels, using standardization data for ages 4–17 years, to support the CHC theoretical basis for the DAS-II for the Early Years and School-Age batteries (Keith, Low, Reynolds, Patel, & Ridley, 2010). These results confirmed the "robustness of the structure across age levels" (Elliott, 2012, p. 347).

Sex differences in cognitive abilities in children ages 5–17 have been found for the DAS-II (Keith, Reynolds, Roberts, Winter, and Austin, 2011). In this study, multi-group mean and covariance structural equation modeling was used on the standardization sample. Girls showed advantages on processing speed (*Gs*) across all ages (especially ages 8–13) and free-recall memory, a narrow ability of long term retrieval (*Glr*), for some age groups. In contrast, boys showed an advantage on visual-spatial ability (*Gv*) for most ages, ranging from less than 1 point at ages 8–10 to almost 5 points at ages 14–15. Younger girls showed an advantage on short-term memory (*Gsm*). Statistically significant sex differences were not found on latent comprehension-knowledge (*Gc*) or the latent *g* factor.

Researchers have explored the application of other theories as they relate to DAS-II, such as SLODR (Reynolds, 2012; Reynolds, Hajovsky, Niileksela, & Keith, 2011). Recently, Reynolds (2012) provided a deeper understanding of how SLODR impacts the measurement of intelligence and interpretation of test scores. The purposes of his study were: (a) to determine whether the *g* loadings of the composite scores were linear, and (b) if they were nonlinear, to demonstrate how SLODR affects the interpretation of these loadings. Using the norming sample, he performed linear and nonlinear confirmatory factor analysis. Several important contributions were made, such as (a) *Gf* was unaffected by SLODR (*Gc*, *Gv, Gsm*, and *Gs* decreased as *g* increased), and (b) "*g* loadings should be viewed as *g* level dependent" (p. 23).

For the most up-to-date research summaries of the DAS-II, consult Reynolds et al. (2011) and Elliott (2012).

The role of theory in contemporary test interpretation

With an understanding of the role of theory in test development, we can now shift our focus to the role of theory in contemporary test interpretation. Until recently, the importance of using theory in test interpretation had not been universally accepted or acknowledged. According to Randy Kamphaus and his colleagues (Kamphaus, Petoskey, & Morgan, 1997; Kamphaus, Winsor, Rowe, & Kim, 2012), theory was not applied to test interpretation until the late 1970s. Theory-based test interpretation has evolved significantly since the early days of Binet, who used a measuring approach. Four "waves" have been delineated in terms of the history of test interpretation: 1) Quantification of a general level; 2) clinical profile analysis; 3) psychometric profile analysis; and 4) applying theory to intelligence test interpretation (Kamphaus et al., 1997; Kamphaus et al., 2012).

The First Wave—Quantification of a General Level

Until the 1900s, identification of mental abilities was strictly medical or physical, such as "idiocy" or "imbecility." The first wave (quantification of a general level) began with Alfred Binet. As described in the previous section, in response to compulsory-education laws and increased failure rates among schoolchildren, Binet was appointed to the French Ministry of Public Instruction. His job was to develop a way to differentiate normal children from retarded ones. His 1905 *Measuring Scale of Intelligence* was created for this purpose. The interpretation was not based on a theory, but rather on two categories—whether the child was "normal" or "retarded."

By the 1920s, other descriptive terms and ranges were utilized: for example, those with IQ scores of 50 to 74 were classified as "Morons," IQ scores of 95 to 104 were described as "Average," and IQ Scores of 125 to 149 were classified as "Superior" (Levine & Marks, 1928).

Terman delineated a different classification system from Binet. His categories ranged from "Definite feeble-mindedness" to "Near genius or genius" (Davis, 1940). During World War II, Wechsler attempted to apply a description of intelligence based on statistical frequencies and distance from the mean (i.e., 50% of people who earned IQ scores of 91 to 110 were in the Average range of intelligence). Today, we continue to use a classification system, but we understand and recognize that

it is only the first step to a meaningful interpretation of test results.

The Second Wave—Clinical Profile Analysis

During the mid-1940s, the use of clinical profile analysis replaced the classification system. The contribution of psychoanalytic theory to this wave is instantly recognizable, with David Rapaport, Merton Gill, and Roy Schafer being the major contributors. In 1940, Rapaport was appointed head of the Psychology and Research Departments at the famous Menninger Clinic. His interest was in understanding schizophrenia, and he did not deny that these individuals had impairments in intellectual functioning. Although Rapaport criticized the field for the lack of theory application, he justified using psychological testing in psychiatric settings (Lerner, 2007). He used a battery of tests, including the Wechsler-Bellevue Scale, Rorschach, Thematic Apperception Test, etc., and applied psychoanalytic theory to interpretation the results of each test.

Rapaport eventually collaborated with Gill and Schafer to propose a new approach (*Diagnostic Psychological Testing*) to test interpretation (Sugarman & Kanner, 2000). They believed that an IQ level had almost no diagnostic significance in their clinical work (Wiggins, Behrends, & Trobst, 2003). So they emphasized the quantitative "interrelations" among subtest scores and the qualitative aspects of individual item responses in order "to demonstrate that different types of maladjustment tend to have different distinguishing and recognizable impairments of test performance" (Wiggins et al, 2003, p. 57). Their five principles were:

(1) Every single subtest score and response was significant and representative;

(2) A comparison of the successes and failures led to further understanding of the examinee;

(3) Subtest scores were related to each other and were representative of the subject;

(4) Both the Verbal score and the Performance score was significant to the examinee's overall makeup;

(5) The data must be considered in light of other data.

The importance of scatter analysis was also described. *Scatter analysis* referred to "the relationship of any two scores or any single score to the central tendency of all the scores" (Wiggins et al., 2003, p. 58). For Rapaport and his colleagues, the Vocabulary subtest served as baseline for subtest comparisons because of its centrality and stability. They suggested that a profile could be indicative of diagnoses such as "simple schizophrenia" or "depressives" (psychotic and neurotic). They stated, "A large percentage of schizophrenics scored relatively low on the arithmetic subtest, while they scored high on digit span. This pattern is a reversal of what is the usual pattern in neurotics, depressives, and normals" (Schafer & Rapaport, 1944, p. 280). In addition, a significant discrepancy between Digit Span Forwards and Digit Span Backwards was indicative of a psychotic process (Wiggins et al., 2003).

The Third Wave—Psychometric Profile Analysis

Access to computers and statistical software launched the third wave—"psychometric profile analysis." The major contributors of this wave included Jacob Cohen (1959), Alexander Bannatyne (1974), and Alan Kaufman (1979). Cohen (1957, 1959) conducted the first factor analyses of the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) and the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955). He used the standardization sample data reported in both manuals for his analyses. For the WAIS, Cohen (1957) identified *g* (general intellectual functioning) along with five Factors Scores, including Factor A (Verbal Comprehension), Factor B (Perceptual Organization), Factor C (Memory), Factor D (Picture Completion), and Factor E (Digit Symbol). His Factor Scores were obtained by averaging the subtests said to measure these abilities. Factor A was obtained by averaging scores from the Information, Comprehension, Similarities, and Vocabulary subtests. Factor B was obtained by averaging the Block Design and Object Assembly subtest scores. Finally, Factor C was obtained by averaging the Arithmetic and Digit Span subtests. Factor D and Factor E were considered "minor factors." In 1959, he analyzed the data from the WISC. His results were closely related to those obtained for the WAIS. In addition to the Verbal, Performance, and Full Scale IQs, five Factors Scores were discovered: 1) Factor A (Verbal Comprehension I); 2) Factor B (Perceptual Organization); 3) Factor C (Freedom from Distractibility); 4) Factor D (Verbal Comprehension II); and 5) Factor E (an unlabeled quasi-specific factor). Cohen indicated that Factor A seemed to involve aspects of verbal knowledge acquired by formal education, including facts (Information), verbal categorization (Similarities), and manipulation of numbers (Arithmetic). He distinguished Factor A from Factor D. Factor B required tasks on

nonverbal skills, involving the interpretation and/or organization of stimuli presented visually against a time limit. These skills included Block Design, Object Assembly, Mazes, and Picture Arrangement. Factor C involved tasks that required attention and concentration, including Digit Span, Mazes, Picture Arrangement, Object Assembly, and Arithmetic. Finally, Factor D involved the use of judgment and included Comprehension, Picture Completion, Vocabulary, and Similarities. Along with obtaining Verbal, Performance, and Full Scale IQs, Cohen delineated other Factor Scores helpful in interpreting an individual's intelligence. He also noted that his studies of the WISC and WAIS provided "insight into the process of intellectual maturation via the comparative analysis of the factorial structures for the three age groups" (Cohen, 1959, p. 285).

Like Cohen, Bannatyne (1974) offered an alternative interpretive system for the Wechsler scales. His reorganization was created in response to attempts to understand the results of the learning disabled (LD) student. The traditional Verbal, Performance, and Full Scale method did not account for the poor performances on certain subtests (i.e., Information and Vocabulary) and adequate performance on the Digit Span subtest. Bannatyne suggested analyzing these students' performances based on Spatial (ability to recognize spatial relationships and manipulate objects in space), Conceptual (ability to use general verbal language), Sequential (ability to retain visual and auditory information), and Acquired Knowledge categories (Webster & Lafayette, 1980). He proposed that a child with dyslexia would obtain a good spatial score and a poor sequencing score (Henry & Wittman, 1981). Although these categories appeared to have high reliability, inconsistent results were found among researchers (Kaufman, 1981). Kaufman noted, "One should not conclude, however, that Bannatyne's recategorizations are irrelevant to LD assessment: that would be far from the truth. Although the groupings do not facilitate differential diagnosis, they still provide a convenient framework for understanding the LD child's assets and deficits" (Kaufman, 1981, p. 522).

From Cohen's work, Kaufman (1979) constructed a systematic method for using the first three factors to interpret the scales of the Wechsler Intelligence Scale for Children–Revised (WISC-R; Wechsler, 1974). He believed:

> The focus is the child, with interpretations of the WISC-R and communication of the results in the context of the child's particular background, behaviors, and approach to the test items. . . . Global scores are deemphasized, flexibility and insight on the part of the examiner are demanded, and the test is perceived as a dynamic helping agent rather than as an instrument for placement, labeling. . . . (Kaufman, 1979)

Kaufman's approach to test interpretation was based on three premises: 1) The WISC-R subtests assess what the person has learned; 2) the subtests are examples of behavior and not comprehensive; and 3) WISC-R evaluates mental functioning under fixed experimental conditions (Kaufman, 1979). This new approach included starting with the most general and global score (Full Scale IQ) and working to the more specific levels (a single subtest) until all meaningful hypotheses about the individual were revealed. He provided case report examples as a way to illustrate his method. Kaufman was the first to merge research and theory with testing. He noted the importance of taking into consideration physical, cultural, and language factors.

The Fourth Wave—Applying Theory to Intelligence Test Interpretation

Kaufman's interpretation method launched the fourth wave of test interpretation—applying theory to intelligence testing. The best contemporary models of theoretical interpretation include the Cross-Battery Assessment approach (Flanagan, McGrew, & Ortiz, 2000) and the Planning, Attention-Arousal, Simultaneous, and Success model of processing (Naglieri & Das, 1997a). The remainder of this section will be devoted to describing these approaches.

Theory-Based Approaches in the Twenty-first Century

CROSS-BATTERY ASSESSMENT APPROACH TO INTERPRETATION

In the late 1990s, Dawn Flanagan, Kevin McGrew, and Vincent Ortiz introduced the Cross-Battery Assessment approach (XBA). They believed that the traditional "verbal" and "nonverbal" interpretative framework presented by Wechsler was ineffective in meeting the needs of contemporary theory and knowledge regarding intelligence and intelligence test batteries (Flanagan, McGrew, & Ortiz, 2000). Their review of volumes of theories and intelligence tests found that an integration of Horn-Cattell G*f*-G*c* theory (Horn, 1991, 1994) and the three-stratum level theory of cognitive abilities provided "the most comprehensive and empirically

supported model of the structure of intelligence currently available" (Flanagan & McGrew, 1997, p. 315). They also found that no single intelligence test battery successfully operationalized the G*f*-G*c* theory or measured all major broad G*f*-G*c* abilities. This integration of theories resulted in the XBA. Flanagan and her colleagues state that the XBA "narrows the gap between practice and cognitive science" (Flanagan & McGrew, 1997, p. 314) and provides assessment professionals with a more "valid and defensible way of deriving meaning from test scores than that provided by the traditional (and largely atheoretical) Wechsler Scale approach" (Flanagan, 2000, p. 295). Furthermore, the XBA is a method of systematically analyzing broad and narrow abilities as a "cluster" rather than by individual subtests, identifies cognitive strengths and weaknesses, aids in the understanding of the relationship between cognitive and academic constructs, and provides a framework to enhance communication between professionals. They also believe that until new test batteries are developed, it is essential that professionals utilize the XBA.

The XBA is based on three pillars—contemporary CHC theory, broad CHC ability classifications, and narrow CHC ability classifications. These pillars are utilized to increase the validity of intellectual assessment and interpretation. The first pillar uses the Cattell-Horn-Carroll theory, which is the most comprehensive and empirically supported model of cognitive abilities (Flanagan & McGrew, 1997). The CHC theory was previous described.

The second pillar of XBA is the CHC broad or Stratum II classifications of cognitive and achievement tests (e.g., G*c* or crystallized intelligence). Flanagan and her colleagues analyzed all subtests from the major intelligence and achievement batteries and classified them according to particular CHC broad abilities or processes. Currently, there are over 500 broad and narrow abilities classifications (Flanagan, Ortiz, & Alfonso, 2007). Having knowledge of which tests measure what abilities helps the clinician "organize tests into . . . clusters that contain only the measures that are *relevant* to the construct of interest" (Flanagan, Ortiz, & Alfonso, 2007, p. 23). For example, measuring short-term memory/working memory (G*sm*-MW) is assessed by subtests such as Wechsler's Arithmetic, the Kaufman Assessment Battery for Children's (KABC-II's) Word Order, and the Woodcock-Johnson Third Edition's (WJ-III's) Numbers Reversed (Flanagan et al., 2007).

The third pillar of XBA is the inclusion of the CHC narrow (Stratum I) classifications of cognitive and achievement tests according to content, format, and task demand (e.g., language development, listening ability, etc.). Flanagan and her colleagues (2007) believed that this layer is necessary to further improve assessment and interpretation validity and to ensure that underlying constructs are represented. The authors provide examples of construct representation and construct underrepresentation. They believe that the latter occurs when the assessment is "too narrow and fails to include important dimensions or facets of a construct" (Flanagan, Ortiz, & Alfonso, 2007, p. 26). An example of construct underrepresentation is the Concept Formation subtest on the Woodcock-Johnson III Tests of Cognitive Abilities (WJ III) because it measures only one narrow ability of fluid intelligence (G*f*). Therefore, according to Flanagan and her colleagues, at least one other measure of G*f* is needed to ensure appropriate representation of the construct. A clinician would need to use the Analysis-Synthesis test in conjunction with the Concept Formation test. In contrast, an example of construct representation is the Verbal Comprehension Index (VCI) of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV; Wechsler, 2008), because this index includes Vocabulary, Similarities, and Information, all of which represent aspects of crystallized intelligence (G*c*).

The guidelines, implementation, and stages of interpretation for this cross-battery approach are detailed and specific. The steps include the following:

(1) Select a primary intelligence battery;
(2) Identify the broad CHC abilities or processes measured by the primary battery;
(3) Select tests to measure the narrow CHC abilities not measured by the primary battery;
(4) Administer all tests;
(5) Enter the data into the XBA computer program;
(6) Follow the guidelines from the results of the program.

The Cross-Battery Assessment Data Management and Interpretive Assistant, a computer program, has been designed to assist the evaluator (Flanagan, Ortiz, & Alfonso, 2007). The latest resource on the clinical application of the XBA and on research studies conducted on the approach is an excellent chapter by Flanagan, Alfonso, and Ortiz (2012).

With any approach, there are strengths and weaknesses. Strengths of the cross-battery approach include its use of modern theory, improved

communication among professionals, is a way to evaluate children with specific learning disabilities and cultural language differences, gives professional flexibility, and has computer-programmed assistance. The XBA affords professional flexibility. The guidelines of the approach allow evaluators to glean different types of data specific to the purpose of the evaluation. In terms of modern theory, the XBA is based on "the most empirically supported and well-validated theory of the structure of cognitive abilities/processes, namely Cattell-Horn-Carroll (CHC) theory. . . . By utilizing this theoretical paradigm, the XBA approach has the advantage of being current and in line with the best available scientific evidence on intelligence and cognitive abilities/ processes" (Flanagan, Ortiz, & Alfonso, 2007, p. 212). This modern theory, in turn, provides professionals with a classification system for clear, valid, and specific communication of an individual's performance similar to the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) for clinicians.

Along with professional flexibility, use of modern theory, and improved communication among professionals, the XBA offers a promising system to evaluate individuals with specific learning disabilities (SLD) and those who are culturally and linguistically different (CLD) (Flanagan, Ortiz, & Alfonso, 2007). The many different definitions, measures, and interpretation approaches to learning disabilities have led to difficulties in evaluating individuals with SLD. The authors of the XBA delineate four levels (with sublevels) that must be met for a definite diagnosis of SLD, as follows:

(1) At Level I-A, a normative deficit in academic functioning is required;

(2) Level I-B, confounding factors (such as insufficient instruction, emotional disturbance, medical conditions, etc.) are considered and determined not to be the primary cause of the academic deficit(s);

(3) Level II-A, a normative deficit in a cognitive ability/process is required;

(4) Level II-B, confounding factors (such as insufficient instruction, emotional disturbance, medical conditions, etc.) are considered and determined not to be the primary cause of either the academic or the cognitive deficit(s);

(5) Level III, underachievement is demonstrated by an empirical or logical relationship between the cognitive and academic deficit(s) and by evidence of otherwise normal functioning, such as mild mental retardation;

(6) Level IV, there must be evidence of deficits in activities of daily life that require the academic skill (Flanagan, Ortiz, & Alfonso, 2007).

In addition to its strengths, the XBA has potential weaknesses, including its norm sample, complexity, time-consuming aspect, and lack of a standardization framework. First, there is no internal norm group, making the validity of the approach questionable. The XBA authors reply to and address this issue in their book *Essentials of Cross Battery Assessment* (Flanagan, Ortiz, & Alfonso, 2007). They believe that the XBA did not need a norm group since the tools used in each battery are valid. In addition, the authors suggest that examiners use assessment tools that were normed within a few years of each other, which leads to greater chances of the norming samples' being similar. A second weakness of the XBA is its complexity. This alleged weakness is seen as a strength by Flanagan and her colleagues. They believe that holding evaluators to a high standard of theory and interpretation is essential. In addition to norming issues and complexity, the XBA is seen as time-consuming. The approach requires more administration, scoring, and hypothesizing than traditional methods. Along with revisions to the approach, a computerized program has been developed to reduce the time required. Finally, when utilizing the XBA, subtests are given out of order or omitted, thus can be seen as violating standardized administration procedures.

Research on the XBA is less plentiful than the plethora of research studies on CHC theory. Researchers have been applying CHC approach to cognitive abilities for many years, and most notably, in relation to academic achievement, including reading, writing, and mathematics (e.g., Flanagan, 2000; Flanagan et al., 2012; Floyd, Keith, Taub, & McGrew, 2007; Floyd, McGrew, & Evans, 2008; Schneider & McGrew, 2012; Taub, Floyd, Keith, & McGrew, 2008). Floyd and his colleagues (2008) investigated the contributions of CHC cognitive abilities to explaining writing achievement. Their participants included the norming sample used for the WJ-III Tests of Cognitive Abilities (Woodcock, McGrew, Mather, 2001b). From a simultaneous multiple regression, the researchers were able to determine that comprehension-knowledge, processing speed, short-term memory, long-term memory, auditory processing and phonemic awareness, and fluid reasoning demonstrated moderate to strong effects on writing achievement (basic skills and written expression).

Flanagan (2000) investigated the validity of CHC approach with elementary school students. Her sample included 166 students from the Woodcock-Johnson Psycho-Educational Battery–Revised (WJ-R; Woodcock & Johnson, 1989) technical manual. These children were given the WJ-R Tests of Cognitive Ability (Extended battery) and Achievement, as well as the Wechsler Intelligence Scale for Children–Revised (WISC-R; Wechsler, 1974). Structural equation modeling was used. Findings demonstrated that the *g* factor underlying the Wechsler-based CHC cross-battery model "accounted for substantially more variance in reading achievement (25%) than the *g* factor underlying the atheoretical Verbal Comprehension-Perceptual Organization-Freedom from Distractibility (VC-PO-FD) Wechsler model" (Flanagan, 2000, p. 295). Results indicated that a Wechsler-based CHC cross-battery approach is "an effective way of ensuring valid representation of multiple cognitive abilities, especially those that have been found to affect significantly the development of reading skills" (Flanagan, 2000, p. 296).

NAGLIERI-DAS PASS APPROACH

Another contemporary and sound theory in test interpretation is the Nalieri-Das PASS approach. In the late 1970s, J. P. Das linked Luria's work to the field of intelligence by suggesting that intelligence be seen as a cognitive construct (Naglieri, 1997). According to Luria, cognitive processing occurred in three separate, but necessary units: 1) regulating of cortical tone and maintenance of attention; 2) receiving, processing, and storing of information; and 3) programming, regulating, and directing mental activity (Das, Naglieri, & Kirby, 1994). Das then described this relationship in terms of the information integration model (Das, Kirby, & Jarman, 1979). Later, Jack Naglieri and Das collaborated to develop the PASS (Planning, Attention, and Simultaneous and Successive processing) theory of cognitive processing. They believed in the importance of Luria's theory, but "focused more on the cognitive processing components rather than their specific neurological locations" (Das, Kirby, & Jarman, 1979). According to this approach, intelligence has three processes—attentional (cognitive activity), informational (simultaneous and successive), and planning.

The first process examined by the theory is attention, which is located in the brainstem and lower cortex (Kirby & Das, 1990). This process allows a person to "respond to a particular stimulus and inhibit responding to competing stimuli" (Naglieri, 1997, p. 249). The major forms of attention include arousal and selective attention. For Das and Naglieri, selective attention was of more interest than arousal, as arousal is assumed. According to the theory, *attention* refers to "specifically directed cognitive activity as well as resistance to the distraction of the competing stimuli" (Naglieri, 1997, p. 250) and is determined by both arousal and planning (Kirby & Das, 1990). Attention and arousal have been linked to task performance, which influences the informational and planning processes (Kirby & Das, 1990).

The information processes include simultaneous and successive processing, which typically operate collaboratively (Kirby & Das, 1990). The major difference is that simultaneous processing allows for "the integration of stimuli into groups where each component of the stimulus array must be interrelated to every other," and successive planning allows for "the integration of stimuli that are serial-ordered and form a chainlike progression" (Naglieri, 1997, p. 250). In essence, with successive processing, the stimuli are not interrelated; rather, each stimulus is related only to the one it follows. Information that is processed simultaneously is said to be "surveyable," because the stimuli are related and can be examined either during the activity (such as copying a design) or through recall (reproducing the design from memory) (Naglieri & Sloutsky, 1995). Simultaneous processing takes place when stimuli are perceived, remembered, or conceptualized and, thus, applied during both verbal and nonverbal tasks. Successive processing is tied to skilled movements, such as writing, because the specific skill requires a series of movements that are in a specific order (Naglieri, 1997; Naglieri & Sloutsky, 1995).

According to the theory, the planning processes use attention and information processes along with knowledge to help an individual identify and utilize the most effective solution(s) to a problem(s). This system is believed to be located in the prefrontal areas of the brain (Kirby & Das, 1990) and includes abilities such as developing a plan of action, evaluating the plan's effectiveness, impulse control, regulation of voluntary actions, and speech (Naglieri, 1997). It is the *how* of the system; how to solve problems.

The PASS processes form a "functional system that has interrelated interdependent components that are closely related to the base of knowledge, and developmental in nature and influenced by the cultural experiences of the individual" (Naglieri &

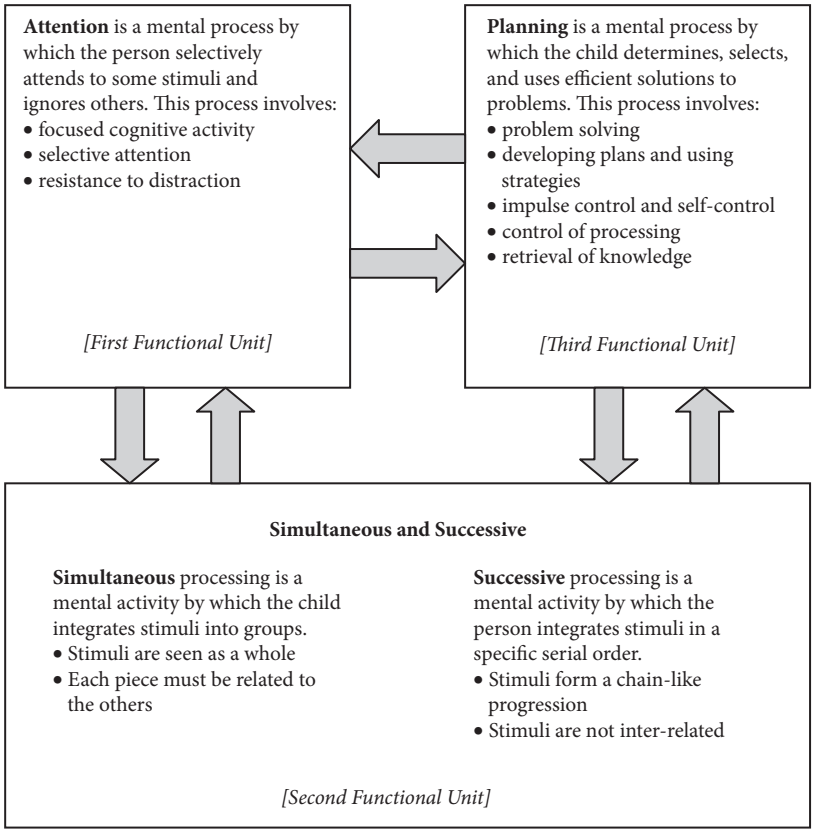


Figure 1.2 The Cognitive Processes of PASS Theory. The PASS processes are dynamic in nature and form an interrelated, interdependent system (as noted by the arrows in the figure).

Sloutsky, 1995, p. 14). The system is interactive; all components work together to perform nearly all of our everyday life tasks. It provides an understanding of cognitive activities (i.e., how individuals learn, think, and/or solve problems). The figure below (Fig. 1.2) describes how the system functions.

Researchers have investigated the use of the PASS theory in evaluating learning disorders, attention deficit/hyperactivity disorder, and mental retardation (e.g., Das, 2002; Kirby & Das, 1990; Kroesbergen, Van Luit, & Naglieri, 2003; Naglieri, 1997, 2001; Naglieri, Das, & Goldstein, 2012; Naglieri & Otero, 2012; Naglieri, Salter, & Edwards, 2004). Naglieri and his colleagues (2004) assessed the PASS characteristics of children with attention and reading disabilities. One hundred and eleven children were administered the Cognitive Assessment System (CAS; Naglieri & Das, 1997a). Results indicated that the children with attention disabilities scored lower on the Planning scale than children in regular education. Children with reading disabilities scored lower on the Successive scale than children in regular education and children with attention disabilities. These children also scored lower on the Simultaneous scale than children in regular education. Das (2002) linked dyslexia with successive-processing deficits. He found that individuals with this specific reading disability make "phonological errors while reading real or made-up words or are slow in reading them (i.e., are slow decoders), or are both slow and inaccurate." (Das, 2002, pp. 31–32).

Conclusions

Psychological assessment involves a synthesis of the information gathered from several resources to understand or make statements regarding an individual's diagnosis, level of functioning or disability, and strategies for intervention or treatment. The history of assessment has its roots in many cultures, dating back to 2200 B.C. Each country focused on different aspects of understanding intelligence and developing measures to assess intelligence. Assessment abounds with many different theories, adaptations, and methods for interpretation that continue to change.

This chapter has explored the role of theory in psychological assessment, which is two-pronged—theory in test development and theory in test interpretation. Theoretically based test development and interpretation provides a strong framework for valid psychological assessments. In terms of test development, the KABC-II, SB5, CAS, WJ III COG, and DAS-II are all valid and reliable testing tools. We believe that the most valid and reliable contemporary methods of test interpretation include the Cross Battery Assessment approach (XBA; Flanagan et al., 1997; Flanagan et al., 2012) and the Planning, Attention-Arousal, Simultaneous, and Success (PASS) model of processing (Naglieri & Das, 1994; Naglieri et al., 2012). We encourage and challenge researchers and practitioners alike to continue developing tests and methods of interpretation based on theory, and to rely on the diverse theory-based instruments for the assessment of children, adolescents, and adults just as they continue rely on Wechsler's scales (Flanagan & Kaufman, 2009; Lichtenberger & Kaufman, 2013).

References

Aiken, L. R. (2004). *Assessment of intellectual functioning* (2nd ed.). New York: Springer.

Ansell, C. (1971). Wild child (*L'enfant sauvage*). *Professional Psychology*, *2*(1), 95–96.

American Psychological Association. (2004). *Intelligence and achievement testing: Is the half full glass getting fuller?* Retrieved 10/10/08 from http://www.psychologymatters.org/iqtesting.html.

Bannatyne, A. (1974). Diagnosis: A note on recategorization of the WISC scaled scores. *Journal of Learning Disabilities*, *7*, 272–274.

Boake, C. (2008). Clinical neuropsychology. *Professional Psychology: Research & Practice*, *39*(2), 234–239.

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical & Experimental Neuropsychology*, *24*(3), 383–405.

Canivez, G. L. (2008). Orthogonal higher order factor structure of the Stanford-Binet Intelligence Scales–5th ed., for children and adolescents. *School Psychology Quarterly*, *23*(4), 533–541.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, England: Cambridge University Press.

Cohen, J. (1957). A factor-analytically based rationale for the Wechsler Adult Intelligence Scale. *Journal of Consulting Psychology, 21*(6), 451–457.

Cohen, J. (1959). The factorial structure of the WISC at ages 7–6, 10–6, and 13–6. *Journal of Consulting Psychology*, *23*(4), 285–299.

Coolican, J., Bryson, S. E., & Zwaigenbaum, L. (2008). Brief report: Data on the Stanford-Binet Intelligence Scales (5th ed.) in children with autism spectrum disorder. *Journal of Autistic Developmental Disorders*, *38*, 190–197.

Das, J. P. (2002). A better look at intelligence. *Current Directions in Psychological Science*, *11*(1), 28–33.

Das, J. P., Kirby, J. R., & Jarman, R. F. (1979). *Simultaneous and successive cognitive processes*. New York: Academic Press.

Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). *Assessment of cognitive processes: The PASS theory of intelligence*. Needham Heights, MA: Allyn & Bacon.

Davis, F. B. (1940). The interpretation of IQs derived from the 1937 revision of the Stanford-Binet Scales. *Journal of Applied Psychology*, *24*(5), 595–604.

Deary, I. J., Lawn, M., & Bartholomew, D. J. (2008). Conversations between Charles Spearman, Godrey Thomson, and Edward L. Thorndike: The international examinations inquiry meetings, 1931–1938. *History of Psychology*, *11*(2), 122–142.

Dumont, R., Willis, J. O., & Elliott, C. D. (2009). *Essentials of the DAS-II Assessment*. New York: Wiley.

Elliott, C. D. (1983a). *The British Ability Scales. Manual 1: Introductory handbook*. Windsor, England: NFER-Nelson.

Elliott, C. D. (1983b). *The British Ability Scales. Manual 2: Technical handbook*. Windsor, England: NFER-Nelson

Elliott, C. D. (1990a). *Differential Ability Scales*. San Antonio, TX: Psychological Corporation.

Elliott, C. D. (1997). The Differential Ability Scales. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 183–208). New York: Guilford.

Elliott, C. D. (2007a). *Differential Ability Scales, 2nd ed.: Administration and scoring manual*. San Antonio, TX: Harcourt Assessment.

Elliott, C. D. (2012). The Differential Ability Scales—Second Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 336–356). New York: Guilford Press.

Flanagan, D. (2000). Wechsler-based CHC cross-battery assessment and reading achievement: Strengthening the validity of interpretation drawn from Wechsler test scores. *School Psychology Quarterly*, *15*(3), 295–329.

Flanagan, D. P., Alfonso, V. C., & Ortiz, S. O. (2012). The cross-battery assessment approach: An overview, historical perspective, and current directions. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 459–483). New York: Guilford Press.

Flanagan, D. P. & Harrison, P. L. (Eds.) (2012). *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.). New York: Guilford Press.

Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC-IV assessment* (2nd ed.). Hoboken, NJ: Wiley.

Flanagan, D., & McGrew, K. (1997). A cross battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 314–325). New York: Guilford.

Flanagan, D., McGrew, K., & Ortiz, S. (2000). *The Wechsler intelligence scales and CHC theory: A contemporary approach to interpretation*. Boston: Allyn & Bacon.

Flanagan, D., Ortiz, S., & Alfonso, V. (2007). *Essentials of cross-battery assessment* (2nd ed.). New York: Wiley.

Fletcher-Janzen, E., & Lichtenberger, E. O. (2005). Strengths and weaknesses of the KABC-II. In A. S. Kaufman, E. O. Lichtenberger, E. Fletcher-Janzen, & N. L. Kaufman (Authors). *Essentials of KABC-II assessment* (pp. 168–175). Hoboken, NJ: Wiley.

Floyd, R. G., Keith, T. Z., Taub, G. E., & McGrew, K. S. (2007). Cattell-Horn-Carroll cognitive abilities and their effects on reading decoding skills: *G* has indirect effects, more specific abilities have direct effects. *School Psychology Quarterly*, *22*(2), 200–233.

Floyd, R. G., McGrew, K. S., & Evans, J. J. (2008). The relative contributions of the Cattell-Horn-Carroll cognitive abilities in explaining writing achievement during childhood and adolescence. *Psychology in the Schools*, *45*(2), 132–144.

Foschi, R., & Cicciola, E. (2006). Politics and naturalism in the 20th-century psychology of Alfred Binet. *History of Psychology*, *9*(4), 268–289.

Hatfield, G. (2007). Did Descartes have a Jamesian theory of emotions? *Philosophical Psychology*, *20*(4), 413–440.

Hebben, N., & Milberg, W. (2009). *Essentials of neuropsychological testing* (2nd ed.). Hoboken, NJ: Wiley.

Henry, S. A., & Wittman, R. D. (1981). Diagnostic implications of Bannatyne's recategorized WISC-R scores for identifying learning disabled children. *Journal of Learning Disabilities*, *14*(9), 517–520.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werber, & R. W. Woodcock (Eds.), *Woodcock-Johnson technical manual* (pp. 197–232). Chicago: Riverside Publishing.

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (pp. 443–451). New York: Macmillan.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: G*f*-G*c* theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York: Guilford.

Ittenbach, R. F., Esters, I. G., & Wainer, H. (1997). The history of test development. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 17–31). New York: Guilford.

Kamphaus, R. W. (2008). *Clinical assessment of child and adolescent intelligence* (2nd ed.). New York: Springer-Verlag.

Kamphaus, R. W., Petoskey, M. D., & Morgan, A. (1997). A history of intelligence test interpretation. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 32–47). New York: Guilford.

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2012). A history of intelligence test interpretation. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 56–70). New York: Guilford Press.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R*. New York: John Wiley.

Kaufman, A. S. (1981). The WISC-R and learning disabilities assessment: State of the art. *Journal of Learning Disabilities*, *14*(9), 520–526.

Kaufman, A. S. (2009). *IQ testing 101*. New York: Springer.

Kaufman, A. S. (in press). Biography of David Wechsler. In F. Volkmar (Ed.), *Encyclopedia of autistic spectrum disorders*. New York: Springer.

Kaufman, A. S., & Kaufman, N.L. (2004a). *Kaufman Assessment Battery for Children–2nd ed. (K-ABC-II)*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (2004b). *Kaufman Test of Educational Achievement–2nd ed. (KTEA-II): Comprehensive Form*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E., & Kaufman, N. (2005). *Essentials of KABC-II Assessment*. New York: Wiley.

Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (2012). Are cognitive *g* and academic achievement *g* one and the same *g*? An exploration on the Woodcock-Johnson and Kaufman tests. *Intelligence, 40*, 123–138.

Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the Differential Abilities Scale-II: Consistency across ages 4–17. *Psychology in the Schools, 47*, 676–697.

Keith, T. Z., Reynolds, M. R., Roberts, L. G., Winter, A. L., Austin, C. A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the Differential Ability Scales – Second Edition. *Intelligence, 39*, 389–404.

Kirby, J. R., & Das, J. P. (1990). A cognitive approach to intelligence: Attention, coding, and planning. *Canadian Psychology*, *31*(3), 320–333.

Kroesbergen, E. H., Van Luit, J. E. H., & Naglieri, J. A. (2003). Mathematical learning difficulties and PASS cognitive processes. *Journal of Learning Disabilities*, *36*(6), 574–582.

Lane, H. (1986). The wild boy of Aveyron and Dr. Jean-Marc Itard. *History of Psychology*, *18*(1–2), 3–16.

Leffard, S. A., Miller, J. A., Bernstein, J., DeMann, J. J., Mangis, H. A., & McCoy, E. L. B. (2006). Substantive validity of working memory measures in major cognitive functioning test batteries for children. *Applied Neuropsychology*, *13*(4), 230–241.

Lerner, P. M. (2007). On preserving a legacy: Psychoanalysis and psychological testing. *Psychoanalytic Psychology*, *24*(4), 208–230.

Levine, A. J., & Marks, L. (1928). *Testing and intelligence and achievement*. New York: Macmillan.

Lieberman, L. M. (1982). Itard: The great problem solver. *Journal of Learning Disabilities*, *15*(9), 566–568.

Lichtenberger, E. O. (2005). General measures of cognition for the preschool child. *Mental Retardation & Developmental Disabilities Research Review*, *11*, 197–208.

Lichtenberger, E. O., & Kaufman, A. S. (2013). *Essentials of WAIS-IV assessment* (2nd ed.). Hoboken, NJ: Wiley.

Mariush, M. E., & Moses, J. A. (1997). *Clinical neuropsychology: Theoretical foundations for practitioners*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McGrew, K. S., & Woodcock, R. W. (2001). Woodcock-Johnson III *Technical manual*. Itasca, IL: Riverside.

Minton, B. A., & Pratt, S. (2006). Gifted and highly gifted students: How do they score on the SB5? *Roeper Review*, *28*(4).

Molnar, Z. (2004). Thomas Willis (1621–1675), the founder of clinical neuroscience. *Nature Reviews Neuroscience*, *5*, 329–335.

Naglieri, J. A. (1997). Planning, attention, simultaneous, and successive theory and the Cognitive Assessment System: A new theory-based measure of intelligence. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 247–267). New York: Guilford.

Naglieri, J. A. (2001). Using the Cognitive Assessment System (CAS) with learning-disabled children. In A.S. Kaufman & N. L. Kaufman (Eds.), *Specific learning disabilities and difficulties in children and adolescent psychiatry* (pp. 141–177). New York: Cambridge University Press.

Naglieri, J. A., & Das, J. P. (1997a). *Cognitive assessment system*. Chicago, IL: Riverside Publishing Company.

Naglieri, J. A., Das, J. P., & Goldstein, S. (2012). Planning-Atten tion-Simultaneous-Successive: A cognitive-processing-based theory of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 178–194). New York: Guilford Press.

Naglieri, J. A., & Otero, T. M. (2012). The Cognitive Assessment System: From theory to practice. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 376–399). New York: Guilford Press.

Naglieri, J. A., Salter, C. J., & Edwards, G. H. (2004). Using the Cognitive Assessment System (CAS) with learning-disabled children. In A. S. Kaufman and N. L. Kaufman (Eds.), *Specific learning disabilities and difficulties in children and adolescents: Psychological assessment and evaluation* (pp. 141–177). Cambridge, England: Cambridge University Press.

Naglieri, J. A., & Sloutsky, V. M. (1995). Reinventing intelligence: The PASS theory of cognitive functioning. *The General Psychologist*, *31*(1), 11–17.

Newton, J. H., McIntosh, D. E., Dixon, F., Williams, T., & Youman, E. (2008). Assessing giftedness in children: Comparing the accuracy of three shortened measures of intelligence to the Stanford-Binet Intelligence Scales, 5th ed. *Psychology in the Schools*, *45*(6), 523–536.

Ortega, J. V. (2005). Juan Huarte de San Juan in Cartesian and modern psycholinguistics: An encounter with Noam Chomsky. *Psicothema*, *17*(3), 436–440.

Reynolds, M. R. (2012). Interpreting the *g* loadings of intelligence test composite scores in light of Spearman's law of diminishing returns. Manuscript accepted for publication in *School Psychology Quarterly*.

Reynolds, M. R., & Keith, T. Z. (2007). Spearman's law of diminishing returns in hierarchical models of intelligence for children and adolescents. *Intelligence, 35*, 267–281.

Reynolds, M. R., Keith, T. Z., & Beretvas, N. (2010). Use of factor mixture modeling to capture Spearman's law of diminishing returns. *Intelligence, 38*, 231–214.

Reynolds, M. R., Keith, T. Z., Fine, J. G., Fisher, M. E., & Low, J. A. (2007). Confirmatory factor structure of the Kaufman Assessment Battery for Children–2nd ed.: Consistency with Cattell-Horn-Carroll theory. *School Psychology Quarterly*, *22*(4), 511–539.

Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence for higher-order MG-MACS and MIMIC models. *Intelligence, 36*, 236–260,

Roid, G. H. (2003b). *Stanford-Binet Intelligence Scales–5th ed.*. Itasca, IL: Riverside Publishing.

Roid, G. H., & Barram, R. A. (2004). *Essentials of Stanford-Binet Intelligence Scales (SB5) assessment*. New York: Wiley.

Roid, G. H., & Pomplun, M. (2012). The Stanford-Binet Intelligence Scales, Fifth Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 249–268). New York: Guilford Press.

Sanders, S., McIntosh, D. E., Dunham, M., Rothlisberg, B. B., & Finch, H. (2007). Joint confirmatory factor analysis of the Differential Ability Scales and the Woodcock-Johnson Tests of Cognitive Abilities–3rd ed. *Psychology in the Schools*, *44*(2), 119–138.

Sattler, J. M. (1992). Historical survey and theories of intelligence. In J. M. Sattler, *Assessment of children: Revised and updated 3rd ed.* (pp. 37–60). San Diego: Jerome M. Sattler, Publisher.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations (5th ed.)*. San Diego: Jerome M. Sattler, Publisher.

Sbordone, R. T., & Saul, R. E. (2000). *Neuropsychology for health care professionals and attorneys: 2nd ed.*. CRC Press.

Schafer, R., & Rapaport, D. (1944). The scatter: In diagnostic intelligence testing. *A Quarterly for Psychodiagnostic & Allied Studies*, *12*, 275–284.

Schneider, W. (1992). After Binet: French intelligence testing, 1900–1950. *Journal of the History of the Behavioral Sciences*, *28*, 111–132.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99–144). New York: Guilford Press.

Schrank, F. A., Flanagan, D. P., Woodcock, R. W., & Mascolo, J. T. (2002). *Essentials of WJ III Cognitive Abilities Assessment*. New York: Wiley.

Schrank, F. A., & Wendling, B. J. (2012). The Woodcock-Johnson III Normative update. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 297–335). New York: Guilford Press.

Shepard, R., Fasko, D., & Osborne, F. (1999). Intrapersonal intelligence: Affective factors in thinking. *Education*, *119*, 663.

Singer, J. K., Lichtenberger, E. O., Kaufman, J. C., Kaufman, A. S., & Kaufman, N. L. (2012). The Kaufman Assessment Battery for Children—Second Edition (KABC-II) and the Kaufman Test of Educational Achievement—Second Edition (KTEA-II). In D. P. Flanagan & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (3rd ed., pp. 269–296). New York: Guilford.

Sugarman, A., & Kanner, K. (2000). The contribution of psychoanalytic theory to psychological testing. *Psychoanalytic Psychology*, *17*(1), 3–23.

Stavrou, E., & Hollander, N. L. (2007). Differential Ability Scales–2nd ed. (DAS-II). *The School Psychologist, Fall*, 120–124.

Taub, G. E., Floyd, R. G., Keith, T. Z., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly*, *23*(2), 187–198.

Taub, G. E., & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance of the Woodcock-Johnson Tests of Cognitive Abilities III. *School Psychology Quarterly*, *19*(1), 72–87.

Thorndike, R. M. (1997). The early history of intelligence testing. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 3–16). New York: Guilford.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Von Mayrhauser, R. T. (1992). The mental testing community and validity: A prehistory. *American Psychologist 47*(2), 244–253.

Wasserman, J. D. (2012). A history of intelligence assessment: The unfinished tapestry. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 3–70). New York: Guilford Press.

Webster, R. E., & Lafayette, A. D. (1980). Distinguishing among three subgroups of handicapped students using Bannatyne's recategorization. *The Journal of Educational Research*, *73*(4), 237–240.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children*. San Antonio, TX: Psychological Corporation.
Wechsler, D. (1955). *Wechsler Adult Intelligence Scale*. San Antonio, TX: Psychological Corporation.
Wechsler, D. (1974). *Wechsler Intelligence Scale for Children—Revised*. San Antonio, TX: Psychological Corporation.
Wechsler, D. (1975). Intelligence defined and undefined: A relativistic approach. *American Psychologist*, 135–139.
Wechsler, D. (2003). *Wechsler Intelligence Scale for Children – 4th ed.*. San Antonio, TX: Psychological Corporation.
Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–4th ed.*. San Antonio, TX: Psychological Corporation.
Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence – 4th ed.*. San Antonio, TX: Pearson.
Wiggins, J. S., Behrends, R. S., Trost, K. K. (2003). *Paradigms of personality assessment.* Guilford Press.
Winzer, M. (1993). *History of special education: From isolation to integration*. Washington, DC: Gallaudet University Press.
Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Tests of Cognitive Ability—Revised*. Chicago: Riverside.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001a). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001b). *Woodcock-Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside.
Woodcock, R. W., McGrew, K. S., Schrank, F. A., & Mather, N. (2007). *Woodcock-Johnson III Normative Update*. Rolling Meadows, IL: Riverside.

CHAPTER

2 Testing: The Measurement and Assessment Link

Scott L. Decker

Abstract

This chapter broadly reviews measurement theory, scale development, testing, and assessment. The chapter is divided into two broad areas to represent distinct phases of testing involving test development and test application. A model is provided to illustrate the integrated role of testing with measurement and assessment components. Theories of measurement are reviewed with the use of the Item Response Theory, not only for the purpose of objective measurement, but as a basic model to analyze how personal attributes interact with test stimuli. Interpretive phases of tests within an assessment process are described, which include decision-making, prescriptive action, and social outcomes.

Key Words: psychological testing, measurement, item response theory, decision-making

Introduction

Measurement theory, scale development, testing, and assessment are all important contributors to test development and and test application. Despite the detailed research in each of these areas, there are few models which focus on the integration and interrelationship across these components. The model described in this chapter is used to illustrate the integrated role of testing with measurement and assessment components. Furthermore, measurement theories are discussed to illustrate the importance of objective in the analysis of how personal attributes interact with test stimuli. Extensions of the model to the interpretive phases of tests within an assessment process are also described.

The chapter is divided into two conceptual sections: (a) pre-application or development stage of testing, and (b) the application stage of testing. During the test development stage, theory and measurement are used for the purpose of understanding the test (i.e., developing construct validity). During the application stage, the test is used to understand the object it was designed to measure. The purpose of dividing the chapter into these two sections is to provide a better integration of the numerous components in assessment, which include aspects of theory, measurement, measurement models, testing, decision-making, diagnosis, and prescriptive action. Additionally, these two sections coincide with contemporary categories of validity (Embretson, 1983). These concepts form a layer of interconnected concepts. For example, testing depends on measurement, which includes scaling, which in turn depends on the theoretical basis of a construct. *Assessment* is the integration of multiple sources of information for the purpose of making a judgement that leads to a prescriptive action. A *test*, or testing, is a device used to measure behavior which provides information in the assessment process. Measurement theory provides a critical foundation for constructing tests as measurement tools. Test interpretation is a part of a decision-making process in which some action, such as an intervention, is to be implemented. Interventions influence outcomes which are evaluated by social goals and values.

Fundamental Issues in Testing

Psychology is replete with conceptual ideas of the inner workings of mental phenomena and postulated causes of behavior. Lacking, however, is the objective measurement of many of these theories and constructs. As a result, paradigms in psychology wax and wane. Often, the constructs of one theory rename the constructs of another theory. Few theories, however, provide objective measures by which to test the theoretical propositions of the theory. Stated differently, few objective measures exist to test the theoretical propositions of most psychological theories. As such, measurement has been described as the "Achilles' heel" of psychological research (Pedhazur & Schmelkin, 1991).

Like in psychology in general, there is disagreement within the specific area of psychological measurement. Measurement models differ in Classical Test Theory from Item Response Theory (to be addressed later). Debate on the role of measurement scale of a test to permissible use of statistical procedures has raged for almost half a century (J. Stevens, 1996). As an additional confounding influence, different researchers use different terminology to describe similar aspects of testing. Terms such as "assessment instruments" are used, although some view assessment and testing as very different. Are "instruments" and "tests" equivalent? Similarly, the definition of "measurement model" and "measurement scale" are used interchangeably.

An additional difficulty in discussing testing, measurement, or assessment is that all of these topics are interrelated. This leads to an extraordinary complexity involved with each of these topics. As a possible consequence, these topics are often extensively written about, but in isolation and disconnected from the other topics. Similarly, many standard assessment textbooks provide adequate coverage on each of these topics but often do not provide an integration of the different components. Often, extensive psychometric evidence for the tests is provided, but applications of the test are dismissed and the test user is left to figure out how to appropriately apply the test, using "clinical judgement" (Kamphaus & Campbell, 2006; Sattler, 2001). Because the ultimate application for test usage has been left unspecified, this may have partially contributed to a growing dissatisfaction with the use of norm-referenced testing. As a result, context-based methods of testing that attempt to more directly link assessment from a specific context (e.g., functional behavior analysis, curriculum-based assessment, portfolio assessment) have grown in popularity, although they have substantially less psychometric rigor.

Foundations of Testing: Measurement

The foundation of testing is measurement. One important historical root of measurement in the behavioral sciences can be traced to Krantz, Luce, Suppes, and Tversky's (1971) *Foundations of Measurement*. The three-volume set provided the basis for what has become known as the *representational theory of measurement* (Krantz, Luce, Suppes, & Tversky, 1971). In representational theory, measurement involves understanding an empirical observation that can be recoded in terms of mathematical structures (Luce & Suppes, 2001). Simplifying, "measurement" includes an object of measurement or the measurement of an object attribute. Object attributes, presumed to vary across different objects, can be coded, or represented, with different numerical values. The initial coding of empirical phenomena with numerals is qualitative. The abstraction of the phenomenon into numerals that are used as numbers in a number system that has quantitative properties is the basis of measurement. Significant debate in this fundamental step has long been a characteristic in the history of measurement. Using physical sciences as a model and reserving measurement for what we now would consider interval and ratio scales, Campbell (1920) insisted that all measurement must satisfy certain properties such as concatenation or additivity. Because psychological measurement rarely demonstrated such properties, Campbell concluded psychology could not be considered a science (Campbell, 1920).

This influenced the development of formal definitions of scaling. Stevens's definition of measurement as "the assignment of numbers to aspects of objects or events according to one or another rule or convention" is perhaps the most popular definition of measurement (Stevens, 1968, p. 850). Stevens's scaling is the assignment of numbers to a sample of behavior along a dimension characterized by some unit of metric. Stevens suggested four types of metrics that continue to be popular: *nominal*, *ordinal*, *interval*, and *ratio* (Stevens, 1946). Given their ubiquity in psychology, they will only be discussed briefly. *Nominal* amounts to naming or classify objects or persons into one or more categories. In nominal measurement of attributes, an attribute is either present or not. *Ordinal* involves the detection of an attribution and the rank ordering of the attribution. That is, object can be rank

ordered (high to low) by the number assignment to the attribute. *Interval* measurement entails not only rank ordering but the "amount" or quantity of difference, with constant or equal amounts between number assignments. Finally, *ratio* includes interval properties but also includes a true "zero" point for the absence of the attribute. Weight and height are two examples of ratio measures, and the widespread use of these measures and being ratio is not coincidental. For a more exhaustive review, see Pedhazur and Schmelkin, 1991.

Various aspects of these early conceptual models of measurement in behavioral science have continued to be debated for over half a century (Gaito, 1980; Guttman, 1977; Lord, 1946; Michell, 1986). However, Stevens's influence on the definition of measurement can clearly be seen in modern conceptualizations of measurement. Townsend and Ashby (1984) described measurement as a process of assigning numbers to objects in such a way that interesting qualitative empirical relations among the objects are reflected in the numbers as well as in the properties of the number system (Townsend & Ashby, 1984), or similarly to objects of measurement (Reynolds, 1986). Additionally, many modern approaches have sided with Stevens by including classification (nominal) and ranking (ordinal) as types of measurement (Pedhazur & Schmelkin, 1991).

Furthermore, various other methods of measurement have been described, but most capture concepts similar to those described in the Representational Theory of Measurement. For example, some have made a distinction between a *natural variable* and a *scaled variable* (Krantz et al., 1971). A natural variable is a variable that is defined by using the actual objects of interest that does not depend on abstract symbols such as numbers, on which scaled variables do depend (Reckase, 2000). Natural variables are directly observable from the objects of interest, whereas scaled variables are not. Natural variables can be conceptualized as detectable from direct observation and are discrete in that the observed event can be classified or be distinguished from other events or the absence of the event. For example, an observable event must be detected such that a determination of whether it is present/absent, or yes/no, can be made, or different gradations can be determined. Scaled variables are a conversion of these observable events into a metric of measurement by some rule. The scaled variable can be a raw score or a raw score corrected for a subject's developmental age.

Applications of Psychometric Models in Measurement

Psychological theory describes the attributes of the object of measurement, the different values an attribute may have, and a causative explanation for differences in values across objects. Often in psychology the object of measurement is a person. The attributes that are of interest are behavior and mental processes that influence behavior. The attributes that are measured are dictated by theory. Similarly, the types of prompts or questions used in testing are dictated by theory. Different theories have different types of emphasis for different attributes. A *construct* is defined as the concept or characteristic that a test is designed to measure. Because constructs are unobservable, different lines of valid evidence are needed to provide information relevant to a specific interpretation of test scores. Furthermore, validity is generally the degree to which evidence and theory support the interpretation of test scores and is considered the most fundamental factor in evaluating a test (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999).

Measurement theory provides only an abstraction of attributes, whereas theory describes attributes in detail. Additionally, measurement describes the process of quantifying attribute values but does not describe what values should be quantified. The meaning of the number may differ by the type of number used to represent the attribute. For example, people share physical attributes (e.g., height, weight) but differ in the values of these attributes (100 lbs, 200 lbs).

Scale is a term used to describe the transformation of behavioral performance, typically in response to questions, into numbers and how to present the questions in order to get the best measurement. Formally, *scale* is the set of rules that are used to assign a number to an attribute (Thorndike, 2005). A familiar scale of measurement for the attribute of length in physical objects is the assignment of inches (the basic measurement unit) from a ruler or tape measure. Another common metric is temperature, which may use either Fahrenheit or Celsius measurement units.

Scaling in behavioral measurement is "messier" than in the physical sciences, or, one might say, it involves a larger degree of error. Scaling in psychology typically involves the assignment of numbers to behavioral responses. The behavioral responses are typically from predetermined stimuli with set rules for assigning numbers. Examples of such scaling are eliciting responses that can be scored as correct or incorrect, and adding the number of correct

responses to get an overall score for a set of items. Scaling may also involve a set of rules to transform the raw score into another measurement scale, such as normative score (percentile, standard, or normalized).

In scaling behavioral measurements, there are different frames of references, or measurement models, that can be used. The two main paradigms are *random sampling theory* and *item response theory* (Suen, 1990), although this is sometimes referred to as *classical test theory* or *item response theory* (Embretson & Reise, 2000). Random sampling theory, which involves both classical test theory and generalizability theory, is based on a true score model. The premise is in any testing situation no person can be exposed to all the possible items within a construct domain. Therefore, the limited sample of items provides an *observed score,* which is viewed as an approximation to a *true score*. These psychometric models essentially address the problem of how to generalize from a sample to a larger population. Classical test theory has been the dominate paradigm until recently. Classic test theory has numerous limitations. Some of the most important involve estimating item difficulties, sample-dependent statistics, single reliability estimates, and problems in comparing scores across similar tests.

Due to these and other limitations in classical test theory, item response theory (IRT) has become the most frequently used psychometric paradigm, especially in test development. IRT is a type of latent trait model that presumes a unitary dimension to describe the attribute that is being measured. A benefit to IRT models is that the model scales behavioral responses based on the joint interaction of a person's ability with the item difficulty (Figure 2.1). The basic idea is that when a person's ability is greater than the item difficulty, then the person has a higher probability of correctly answering the item. Conversely, when the item difficulty is greater than the person's ability, the person should incorrectly answer the item. When the item difficulty and the person's ability are equally matched, there is a 50/50 chance of getting the item correct. This basic relationship is modeled with a logistical curve (Figure 2.2).

Model-based measurement in IRT is fundamentally different from classical test theory. Similar to the measurement of physical objects, one does not need to invent a new "ruler" for every object investigated. Instead, the ruler or tape measure is used as an existing model. Item response models work on a similar premise. Although IRT models differ in the number of parameters used in the model, only the Rasch (one-parameter) model will be described here. There are also two- and three-parameter models that include parameters for item discrimination and guessing. These models are not presented, because they are extensions of the basic Rasch model, and some have argued the adding these additional parameters compromises aspects of objective measurement (Wright & Stone, 1979). The Rasch model describes the outcome of a person's ability interacting with a stimulus (item) with some difficulty that results in a binary outcome, such as pass/fail, correct/incorrect, etc. The underlying model is a logistics curve that models success and failure based on a person's ability and an item's difficulty. Unlike in classical test theory, the values for the item difficulties are not sample-dependent, just as the units of measurement do not change on a ruler based on the object being measured. The probabilistic outcome is a function of the difference in the person's ability (B) and the difficulty of the item (D). Rasch (1960) described the specific ordinal relationship to describe probabilities of a test simply as "a person having a greater ability than

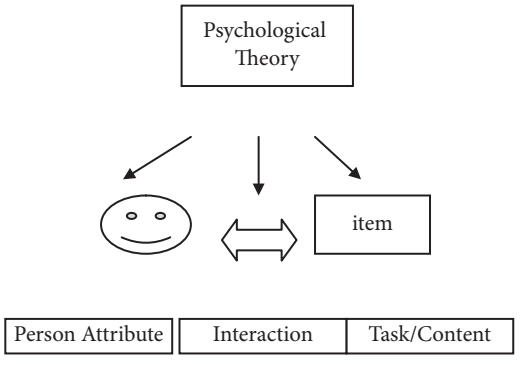


Figure 2.1 Schematic of theory in specifying person by task interaction.

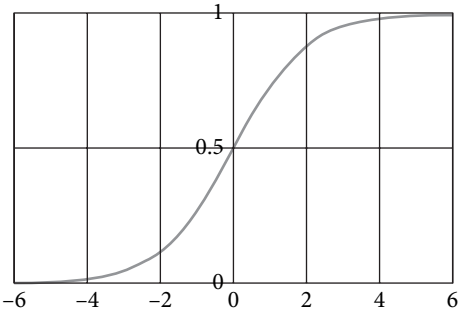


Figure 2.2 Logistical curve of ability with probability of response.

another person should have the greater probability of solving any item of the type in question (p. 117). Formally, when B = D, the probability of a correct response is 50/50. When B > D, the probability of a correct response increases from .5, and decreases when B < D. Formally, the probability of a correct answer is given in the following equation:

$$P(x = 1) = e^{(B-D)}/1 + e^{(B-D)}$$

where e is the natural log function (2.7183) which is raised to the difference in the person's ability (B) and the item difficulty (D). The resulting units of measurement are described as *logits*, which are typically set to the arbitrary value of 0 as the mean. Suppose someone with a logit ability of 3 completed a spelling item that was calibrated to have a difficulty of 1. Using equation 1, the probability of correctly answering the item can be determined as:

$$P(1) = 2.7182^{(3-1)} / 1 + 2.7182^{(3-1)} = .88$$

Similarly, if a person with a logit ability of 3 interacted with an item calibrated with a difficulty of 4, obviously, the probability of success would be much less than in the previous example; and more specifically:

$$P(1) = 2.7182^{(3-4)} / 2.7182^{(3-4)} = .27$$

The relationship of different ability–difficulty differences can be viewed in Table 2.1. Object measurement is the repetition of a measuring unit and describes a constancy in measurement not dependent on the sample or measurement instrument. Notice from Table 2.1, the probability of success is the same for the same differences in measurement regardless of the value of the measurement. For example, there is a probability of .73 regardless of whether the person's ability/item difference is 3-2, or 2-1, or -1-(-2).

Table 2.1 Probability Outcomes Based on Person Ability and Item Difficulty

B-D	P(x = 1)
3.0	.95
2.0	.88
1.0	.73
0.0	.50
–1.0	.27
–2.0	.12
–3.0	.05

Although conjoint measurement, which enables equal interval scaling (Stevens, 1946), technically includes Weak Order, Independence, Double Cancellation, Solvability, and Archimedean Condition (Kyngdon, 2008), the Rasch model's fulfillment of these properties, or approximate fulfillment, has led many to conclude it is the only measurement model in psychology that provides interval scaling (Andrich, 1988; Bond & Fox, 2001; Embretson, 1999; Embretson & Reise, 2000; Woodcock, McGrew, & Mather, 2001; Wright & Stone, 1979). Some disagree with these claims since there is still difficulty in verifying the equal interval nature of the actual underlying psychological or causal process of behavioral responses (Kyngdon, 2008). Additionally, some have argued IRT metrics are still arbitrary until observed scores, no matter the form, are mapped onto meaningful behaviors (Blanton & Jaccard, 2006). Regardless, such probabilistic features that are not sample-dependent represent a substantial improvement in psychometric measurement from that of its historic predecessor, classical test theory (Embretson, 2006).

This transformation of test values that provide indicators of behavior to a measurement scale is the quintessential distinction in testing as the "use of an instrument" from testing as "measurement." The degree to which a test adequately "measures" a construct, rather than provides an arbitrary representation, is directly related to the degree to which valid inferences can be made on the change in amount of a construct. Thus, the issue of understanding the measured representation of psychological constructs is not just a technical issue relevant for quantitative psychologists, but is the foundation in which all concepts in psychology that involve constructs, which is the nearly the whole of psychology. In practice, items calibrated with the Rasch model are selected to have different difficulties that adequately cover the range of ability. The scale with selected items is then used for practical applications.

Psychological measurement, which involves both psychological theory and measurement, will continue to evolve. As demonstrated by measurement models, psychology will always involve the analysis of a person–item interaction, where the item may be an item or some other contextual variable identified or derived by psychological theory. Although far from perfect, the foundations of measurement

with representational theory of measurement and the application of measurement models as used in IRT probably represent the pinnacle, or near pinnacle, of measurement in the behavioral sciences. It is difficult to imagine what new purely measurement developments could occur that would fundamentally change psychological measurement beyond that provided by IRT models.

Testing

Keeping the complex nature of measurement in mind, testing can now be more directly addressed. A *test* is an evaluative device used to sample an examinee's behavior in a specified domain that is scored using a standardized process (AERA et al., 1999). The objective of testing is to describe a characteristic of a subject as a numerical score to represent the quantity of the characteristic (Suen, 1990). Objects of measurement are psychological constructs. When used in assessment, tests are used to obtain information and reduce uncertainty (McFall & Townsend, 1998).

Although a test can be simply defined as a device for scoring behavior, the intricacies in this process are complex. A test is the assembly of stimuli that elicit behavioral responses from a test taker in which behavioral responses are numerically coded. The stimuli are typically calibrated, or ordered by difficulty, to form a scale that measures an attribute of an object (i.e., personal characteristic). The selection of the test stimuli or content is theoretically based. Additionally, a test provides information on the status of an attribute by recording some observable event or behavior. Linking recorded observations from the test to a measurement unit is an aspect of scaling.

Testing, as a component of psychological assessment, typically provides a measurement of a person's attribute (i.e., mental process). Multiple tests are used to measure different attributes to provide a comprehensive assessment to assist in the assessment decision-making process.

Behavioral responses are scaled by recording behaviors, usually with a predetermined response format, representative of the construct. Constructs have a dimension; that is, higher or lower amounts of a construct. The dimension represents the range of values to describe individual differences in attribute values across different objects. Objects are multidimensional (i.e., multi-characteristics) but are typically measured by unidimensional tests. Different attribute levels, as indicated by score values, are then examined or correlated with other attribute values from different constructs as well as with important outcomes or events. For example, intelligence tests measure intelligence by combining multiple subtests measuring some theoretical attribute of intelligence. Scores are corrected for age-related variance and converted to a scale of a mean of 100 and a standard deviation of 15. The differences in levels of intelligence across different people result in a distribution, typically normal or Gaussian, across the measurement scale. Correlational methods can investigate the relationship of variations in intelligence with variations in other variables such as personality and academic achievement. The question as to whether intelligence can be represented by a single variable, and the nature of that single variable, is not an issue of measurement. Rather, this is an issue of theory and validity. Similarly, the accuracy and stability of assigning numbers to represent differences in attributes is an issue of reliability, which influences measurement but is not measurement.

Testing in Assessment

At the time of writing this chapter, the Supreme Court of the United States made a decision in which testing was at the center of the lawsuit. In the *Ricci v. DeStefano* case (decision made in June, 2009), 20 firefighters sued the city of New Haven, Connecticut, alleging that they were discriminated against. Firefighters promotions were determined based on a test, but the test scores resulted in perceived disproportionate number of promotions of white firefighters. As a result, the test was declared invalid, and the results were discarded for fear of a lawsuit by the non-white firefighters. However, discarding the result of the test was also viewed as discrimination—against the white firefighters (and one Hispanic firefighter), and resulted in a lawsuit. Ultimately, the Supreme Court ruled that the city should not have thrown out the exam, arguing that by doing so, the city was using race as a criterion for promotion, which violated Title VII of Civil Rights Act of 1964, in which employment decisions cannot be made based on race.

The point of mentioning this Supreme Court case is not to state an opinion on the verdict or address the issue of test bias (see Reynolds, Lowe, & Saenz, 1999, for more on test bias). The point here is to simply illustrate the complex and numerous layers of meaning involved in testing, which extends beyond just a device for measuring. Tests are always developed and administered for some purpose. The purpose is usually driven by some social need (e.g., promotion, intervention, or classification).

In each situation, judgement is required based on a decision-making process. The judgement then results in some action that satisfies the social need. Furthermore, social benefit, or perceptions of social benefit, may influence not just assessment but test development. In the New Haven firefighters' case, the city officials were required to make a judgement, first on test scores and then on the permissible use of test scores, which influenced social outcomes. As such, judgement and decision-making, as well as the resulting actions and the outcomes of those actions, are important components in assessment and provide an important link between theory, measurement, testing, assessment, and social outcomes.

Figure 2.3 depicts the integrative influence in the relationship between testing, measurement and assessment. As the role of theory, measurement, measurement scale, and assessment have been discussed previously in this chapter, the remainder of the chapter will cover *judgement*, *prescriptive action*, and *social outcome*. Although clinical judgement research is readily available, the process of translating a judgement into some action is not. Often, the action taken is more contextually derived and is difficult to determine in the abstract. Similarly, social outcomes are important but rely upon some action, which in turn relies upon judgement and assessment. In most treatment utility paradigms, assessment and decision-making are taken for granted and not represented in the models. Often such research demonstrates the utility of a behavioral intervention using a single-subject design and concludes that the change in baseline during the intervention did not require any sophisticated cognitive or personality tests. Not included, but important within an applied context, is "why," or the justification, an intervention was deemed to be needed and "why" the particular intervention was chosen. Such processes in behavioral research have remained covert mental processes of the experimenters.

Figure 2.3 represents a cyclical process that suggests that the major components of measurement, testing, and assessment are interrelated. Testing, in development or application, is interconnected to theory, measurement, and social values and consequences.

Not intended to be a unitary model of construct validity, the present model intended to (1) emphasize the sequential relationship of key stages in the application of tests in an assessment process, and (2) to emphasize the interrelatedness of these key stages. The Nomological Network (Messick, 1995) model consists of distributed but connected

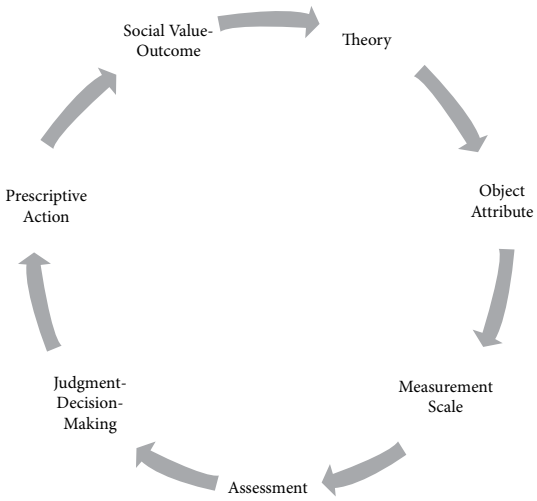


Figure 2.3 Sequential cycle of measurement, testing, and assessment.

nodes. Although this is accurate, it may not capture the sequential nature of the assessment process nor the sequential process of measurement or decision-making and how they are interconnected. Theory guides scale development, which influences which measures are used in a particular assessment. Similarly, judgements are the result of an assessment process and lead to an action that is "theoretically" believed to have a desirable social outcome; thus social value and theory are connected. Thus, the end of the chain of reasoning in testing loops back to the beginning in that it influences the actual design of the test. This is also partially idealized. For example, test development typically starts and ends with the accuracy in measuring a construct. As suggested by Figure 2.3, test development may also benefit by starting with (1) what is the social value of measuring a particular attribute, (2) what action or intervention can be taken based on information about an attribute, or (3) how can decisions be made based on a measurement of an attribute. Here, test development begins with the end in mind.

Assessment

Assessment is a broader term than *testing* that involves the integration of test information with other sources of information (AERA et al., 1999). Assessment is a framework for constructing a unified meaning from various sources of information. Assessment goes beyond test scores and involves a multi-step and multidimensional evaluation of an individual.

Assessment marks the point at which a test, constructed via the methodology previously presented, is used as a tool of investigation rather than being the focus of the investigation. Assessment is:

> concerned with the clinician who takes a variety of test scores, generally obtained from multiple test methods, and considers the data in the context of history, referral information, and observed behavior to understand the person being evaluated, to answer the referral questions, and then to communicate findings to the patient, his or her significant others, and referral sources. (Meyer et al., 2001, p. 143)

Because contextualized decision-making is required, assessment is not a completely objective process. As Matarazzo (1990) described in his APA presidential speech, assessment is "a highly complex operation that involves extracting diagnostic meaning from an individual's personal history and objectively recorded tests scores . . . it is the activity of a licensed professional, an artisan familiar with the accumulated findings of his or her young science. . . ." (p. 1000).

Assessment is often described in multiple stages. Sattler (2008) described assessment as an 11-step process that includes collecting data from multiple sources that include both formal testing procedures as well as observations, and clinical judgement. McFall and Townsend (1998) described assessment as consisting of eight layers that integrated various aspects involved in assessment. Layer 1 consisted of postulates, which were assumptions, beliefs, or values. Layer 2 was a formal theoretical model. Layer 3 was described as referents or observable instantiations, Layer 4 was instrument methods; followed by Layer 5 of measurement model; Layer 6, data reduction; Layer 7, data analysis; and Layer 8, interpretation and inference. A loop connects Layer 8 with Layer 2 to demonstrate the influence of inferences on the questions that gave rise to the assessment process (McFall & Townsend, 1998). According to McFall and Townsend (1998) the purpose of assessment was one of obtaining information and reducing uncertainty.

Most models of assessment generally describe the process of transforming test data into usable information as part of "test interpretation." Sattler notes that test interpretation is the most challenging step in the assessment process, and it involves integrating assessment data, making judgements, and exploring implications (Sattler, 2008). Interpretation of test scores to provide meaningful information is central in the assessment process. Blanton and Jaccard (2006) indicated meaning from test scores "must be established through research that links specific scores to the observable events that are relevant to the underlying psychological dimension of interest" (Blanton & Jaccard, 2006, p. 33). Similarly, scaled scores are believed to aid interpretation by indicating how a given score compares to those of other test takers (AERA et al., 1999). Procedural, objective, algorithmic methods for deriving "meaning" from test scores are generally not recommended because of the complexities involved with assessment, which include linking validity studies to a contextual purpose.

Interpretation of test scores is connected to the validity evidence for a test. According to the Standards, "Test scores ideally are interpreted in light of the available normative data, the psychometric properties of the test, the temporal stability of the constructs being measured, and the effect of moderator variables and demographic characteristics" (AERA) et al., 1999, p. 121). Tests are valid to the degree in which evidence supports inferences from the test. The evidence to support inferences is based on validity evidence; thus test validity is central to test interpretation. *Validity* refers to the degree that evidence and theory support the interpretations of test scores (AERA et al., 1999).

Models of test validity have evolved over time to more accurately represent the nuances of process involved in the application of tests in assessment. Traditional validity research amounted to obtaining evidence that the test was measuring what it was suppose to measure (Campbell & Fiske, 1959). Here, construct validity was the central focus and obtained primarily by evidence of a test's correlation with other tests with a similar label, and no, or lower, correlations with tests having a different label. Construct validity has been traditionally viewed as empirically established after the test was constructed (Cronback & Meehl, 1955). The "meaning" of test scores was determined by their relation with other variables, which formed what was termed a "nomological network." Similarly, the nomological network model of validity attempted to expand construct validity to incorporate other aspects of the assessment process (Cronbach & Meehl, 1955). This view was criticized at its inception as confusing "meaning" with "significance" (Bechtoldt, 1959). Over time, the validity concept has become "encrusted with additional meanings" and is likely to require revisions (Schwager, 1991).

Additionally, nomological networks have been difficult to define (Cronbach, 1988). Embretson (1985)

attempted to provide clarification by distinguishing construct representation from nomothetic span. *Construct representation* involves evidence to understand the processes, strategies, and knowledge that persons use to respond to test items and how these behaviors are represented by test scores. *Nomothetic span* is the evidence to support how individual differences as represented by test scores are related to external variables and the utility of those relationships. Different researchers promote different types of validity and use different terminology. For example, *nomological span* (Embretson, 1983) is synonymous with *external validity* (Cook & Campbell, 1979), which in turn is referred to as a *nomological network* (Messick, 1995), which creates additional problems for relating validity evidence to test interpretation.

Such criteria attempt to bridge the gap in assessment between using tests in data collection and making inferences leading to judgement with tests. Additionally, such criteria imply published validity on a test will make the connection for the clinician. Unfortunately, published research on most psychological tests does not provide such guidance.

In an attempt to focus less on the mechanical issues of construct validity, Messick (1980) has attempted to better address the connection between test "interpretation" and the social consequences of tests. Messick unification is described in the following diagram:

Other researchers have begun to de-emphasize construct validity, which has traditionally been viewed as the core pillar of assessment, and placed more emphasis on the social consequences aspect. The focus on social outcomes as the credentialing criteria of usefulness has been termed *treatment utility*. Treatment utility is "the degree to which assessment is shown to contribute to beneficial treatment outcome" (Hayes, Nelson, & Jarrett, 1987, p. 963). This functional approach would argue the only utility of testing is the degree to which it is associated with change in some valued social outcome. This approach places "treatment validity" at the core of validity (Fuchs, Fuchs, & Speece, 2002) and has led some to suggest norm-referenced tests should be discontinued due to a lack of treatment validity (Gresham, 2002; Reschly & Grimes, 2002).

Table 2.2 Messick's View of the Interaction of Test Interpretation with Prescriptive Action and Social Outcome Variables

	Test Interpretation	Test Use
Evidential basis	Construct validity (CV)	CV + relevance/utility (R/U)
Consequential basis	CV + value implications (VI)	CV + R/U + VI + social consequences

Adapted from Messick, 1980.

Contemporary models of validity are therefore fragmentary. Integrating these different points of view has been difficult. The social consequences, such as treatment benefit, should be more highly weighted than in traditional models. However, sole focus on social outcomes creates numerous problems (Decker, 2008; Reynolds, 1986). As Messick (1980) states,

> What matters is not only whether the social consequences of test interpretation and use are positive or negative, but how the consequences came about and what determined them. In particular, it is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity, such as construct under-representation or construct-irrelevant variance. (p. 748)

In Figure 2.3, processes are represented as beginning in assessment, judgement, prescriptive action, and social outcome. In support of treatment utility theories, much can be gained by first asking, "What is the benefit?" However, in support of construct theories, even social benefit involves "theory"; how the benefit came about is important, not just whether or not the benefit occurred, and some form of decision-making that informs prescriptive action is part of all interventions.

Attributes exist in individuals as finite, discrete properties but are measured as continuous variables labeled as constructs. Test interpretation involves numerous scores on continuous scales. However, the social value of assessment is one of deriving a discrete judgement. Thus, assessment requires the judgement of discrete probabilities from continuous scales to map to some prescriptive outcome. Currently, the categorical interpretation of test scores is arbitrarily given by dividing the normal distribution curve into ordinal categories (e.g., below average, average, above average). Indeed, qualitative outcomes appear to be the most informative to patients (depressed, not depressed) and may provide a linkage from test scores to real-world outcomes.

Judgement and Decision-Making

The complexities of making judgements from assessment data are vast due not only to all the issues

in measurement theory, the theory of the construct being measured, and situational factors during testing, but also due to how these factors impact, or are impacted by, the contextual issues involved in applied assessment applications, which include social consequences. This gets to the interrelatedness of these different concepts. Unfortunately, the complexity increases as it is at this point that interpretation is defined by cognitive events of the clinician and thus a large number of new variables become influential. The role of clinical interpretation as part of the assessment process is perhaps the most important link in the chain (McFall & Townsend, 1998). However, it is also mostly described in qualitative terms such as "integrative," "holistic," "comprehensive," and "synthesis." Granted, this is due in part to the vast complexities involved that do not readily lend themselves to statistical modeling. Additionally, contemporary models of validity do not specify how clinicians should make evaluative decisions based on a certain context. Yet, evaluative clinical decisions are the primary mechanism that leads to prescriptive action and in turn to outcome. As a result, such processes are unaccounted for and remain implicit and ambiguous, or are determined to be irrelevant.

Perhaps part of the reason why "interpretation" and validity are often not specified or specified in multiple ways is because of the complexity and challenges involved. Part of the challenge is that validity research, often conducted under controlled conditions, may not always be relevant for the contextual issues in applied practice. Additionally, clinical decision-making is intimately a part of the assessment process, and different types of validity are prioritized based on the decision-making demands of researchers or clinicians (Kazdin, 2003). One validates, not a test, but an interpretation of data arising from a specified procedure (Cronbach, 1971). Recall that test validity is in part determined by evidence suggesting the test is measuring what it is "intended" to measure. However, "intentions" change based on contextual situations. A test can be perfectly valid and reliable but have no link or implication for real-world processes that are relevant to a clinical situation.

The primary confounding problem that has historically plagued assessment is the lack of integration between the different components involved in measurement and assessment within an applied context, such as treatment. Historically, psychological or cognitive measures have been developed to fit a particular theory, and evidence is provided to validate a test as a measure of a theory. Traditionally, the outcome of testing was interpretation. The actual applications of how many practitioners would be using the instrument to make decisions has been of secondary value. True, test developers cannot anticipate all the possible uses for a particular test. However, these tests are then used by practitioners, and it is left up to the practitioner to know how to apply the test to assist him in making a decision in a specific context.

A review of research on clinical judgement is beyond the scope of this chapter (for a review see Garb, 1998; Garb & Schramke, 1996). In this chapter, two descriptions of cognitive phenomena in test interpretation will be discussed. Based on social-psychological research beginning with Solomon Asch in the 1940s, social psychologists have extensively researched how individuals develop overall impressions or judgements based on the accumulation of data (Lewicka, 1997). Lewicka (1988) distinguished between "diagnostic" and "prospective" processes, which have also been termed "categorical" versus "piecemeal" or top-down/bottom-up (Fiske & Pavelchak, 1986). Diagnostic inferences involve inferring a category membership based on specific features of an object (attribute→category); whereas prospective inference infers features of an object based on its category membership (category→attribute). Diagnostic inferences are bottom-up and data-driven; whereas prospective inferences are top-down and theory-driven. Essentially, observations are categorized to form concepts. Concepts in turn help us understand observations. In assessment, clinician judgement is influenced by the degree to which observations and concepts are used. When a concept is formed prior to data collection and not supported by data collection but maintained in prescriptive action, this is called *bias*.

During testing, clients are provided scores on various dimensions that represent attributes derived from theoretical constructs. In assessment, clinicians use scores on dimensions in supporting both diagnostic and prospective judgements that justify prescriptive actions in a social context. The whole foundation of clinician judgement resulting in social benefit rests on the mechanics of measurement, beginning with the assignment of a number to an attribute of a person.

It is important for clinicians to be aware of the type of decision that is required in an assessment process. This involves thinking through the referral question and clearly stating the problem,

determining possible outcomes of an assessment, and determining how outcomes will be prioritized based on assessment data. This provides an important connection between clinical judgement and prescriptive action that leads to social outcomes. Structuring assessment judgement outcomes is helpful in this process.

Although psychological assessments can be used for numerous goals (e.g., measure client attribute; determine disability, strengths, and weakness, etc.), we will limit the scope here to classification. Classification systems establish rules for placing individuals within a specific class and provide a means of investigating correlates of class membership, such as treatment outcomes (Sattler, 2001). In an assessment context where the clinician is asked to provide a diagnostic judgement, the judgement can be one of meter reading (Blanton & Jaccard, 2006), will be used to specify an interpretive statement made by the clinician that is a direct translation of a test score to another scaled frame of reference. For example, norm-referenced tests have charts indicating qualitative labels of, for example, "above average," "average," "below average." "Inferred" interpretation is inferential. It involves direct interpretations of a test in which but goes beyond the test data. Diagnostic judgements—disability classification, for example—use numerous sources of data, none of which would directly lead to a clinical judgement.

As a simple example, suppose a clinician is asked to determine whether a person has a disorder or not (criterion classification), based on one data point that provides a positive or negative indicator. This interaction can be modeled in a classification matrix. Data can accurately classify individuals by suggesting they have the condition, when they do, or they do not have the condition, when they do not. The other outcomes could be that the data incorrectly indicate the person either has or does not have the condition when the converse is true. Notice this example is simplistic in that psychological test data rarely provide a binary outcome of disorder/no disorder. Additionally, there are complexities involved in determining true criterion status. (Interested readers may consult the following sources for more detailed aspects of this process: Elwood, 1993; Franklin & Krueger, 2003; Gigerenzer, 2002.) However, the example is intentionally simplified for demonstration.

Table 2.3 Decision-Making Matrix for Determining Judgement and Test Correspondence

Criterion Classification		
Test Results		
	+	–
+	A	B
–	C	D
	A + C	B + D

A = Sensitivity (A/A + C)
B = False positive
C = False negative
D = Specificity (D/B + D)
Positive predictive power = A/A + B
Negative predictive power = D/C + D

In such a scenario, a classification matrix describing the hypothetical outcomes of the test may be useful. Although such a matrix is frequently found in many assessment textbooks, such information is rarely reported for standardized commercial instruments used by psychologists (Elwood, 1993). There are several challenges to the use of such tables in practice. One problem, the *base rate problem*, has long been recognized and results from most of the clinical conditions' having a low prevalence rate (Meehl & Rosen, 1955). In such conditions, positive predictive values almost always suggest classifying an individual as not having the disorder despite test data.

One means of overcoming this limitation is to use Bayesian methods (Franklin & Krueger, 2003). Bayes's method is useful because it starts with the base rate probabilities of outcomes (disorder prevalence), then revises the probabilities based on new information.

As an example, Figure 2.4 shows the base rate probability for different diagnostic judgements that may be made when using assessments in schools. The overwhelmingly most likely categorical decision to be made from a random evaluation of any child in school is "normal." Thus, any information suggesting a different category would have to be overwhelmingly informative to overcome this large base rate. Unfortunately, no such information exists. Fortunately, Bayes's theorem provides a method to resolve this issue.

Bayes's theory is a method of revising probabilities based on data. As a simple example, suppose the base rate of classifications for a group of disabilities frequently made in children are as shown in Figure 2.4. Furthermore, suppose the probability of classification for each of the disabilities is related to IQ differentially. For example, the probability of not having a clinical condition is linearly associated with IQ. Children with learning disabilities on average may have an average or slightly below-average IQ, as do children with ADHD. Children with mild mental

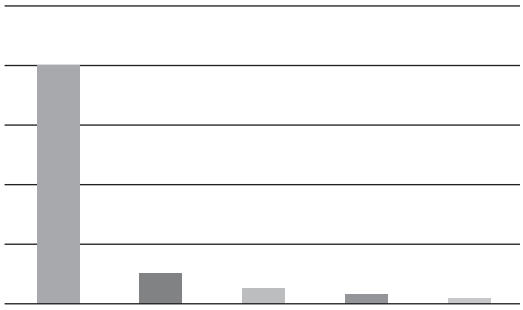


Figure 2.4 Initial probability for categorical judgements based on base rates.

retardation (MMR) have lower IQs and by definition, typically two standard deviations. Children with pervasive developmental disorders like autism may have on average low IQs, but children within this classification may also have large standard deviations.

Suppose that, during the assessment process, an IQ score was obtained and resulted in a score of 75. Further suppose that the probability of being normal given an IQ of 73 was .20, and the probability of not being normal given an IQ of 73 was .80. That is, 80% of children with an IQ of 73 are found to have some clinical condition and are judged "not normal" or "developmentally atypical." However, there is about a 20% chance of finding children who test this low on an IQ test but do not exhibit any atypical developmental features or any other impairment. What is the value in changing the probabilities of determining a child is normal based on this information?

Bayes's theorem states that the probability of having a condition (C) given the data (D) is equal to the probability of the data given that the hypothesis is true (sensitivity), multiplied by the base rate, then divided by a normalizing factor that includes test specificity. Here the values are:

P(D|C) = .20 (probability of getting 73 given normal, sensitivity)

P(C) = .80 (base rate of normal)

P(D|~C) = .80 (probability of getting test score given NOT normal, specificity)

$$P(C|D) = P(D|C)*P(C) / P(D|C) * P(C) + P(D|\text{ not } C)*(1-P(C)$$

If these numbers are entered into Bayesian formula, then the probability of being normal goes from .80 (base rate) to:

$$P(C|D) = .20* .80 / .20* .80 + .80*(1- .80)$$

$$P(C|D) = .50$$

Suppose further that it was known that the probability of having MMR, given a test score of 73, was .80 and a specificity .20. That is, 80% of children with an IQ of 73 may also be shown to have low adaptive behavior, family history of MMR, very severe academic deficits that progressively drop by grade, etc. How would this information change the likelihood of MMR?

Using the same procedure as before, only changing the base rate to 3% (prevalence of MMR), the new probability is .11. Disregarding the effect of the other classification, the new graph revised where the probability of "normal" goes from .80 to .50 and the probability of MMR goes from .03 to .11 (Figure 2.5).

Now, additional information, such as "adaptive behavior," which has its own sensitivity and specificity with normal and MMR, can be added to further change the likelihood of different categories. Similarly, background information such as gender, ethnicity, or age, could be added to influence the results (see Franklin & Krueger, 2003, for more complex examples using Baysian networks). Eventually, multiple sources of information can be "integrated" to inform a categorical judgement. This procedure directly addresses the base rate problem as well as other problems that have plagued clinical inference (McFall & Treat, 1999). The base rate problem is overcome by the accumulation of highly sensitive and specific information. Additionally, it is proposed that this process "simulates" what good clinicians do when they "integrate" or "holistically" appraise test data within the assessment process. Additional implications will be discussed at the end of the chapter.

Prescriptive Action

The purpose of clarifying classification decisions is not just to provide a better "label," but rather to reduce the uncertainty of options in classification schemes, which in turn provides ready access to research on interventions (Kamphaus, Reynolds, & Imperato-McCammon, 1999). That is, classification or diagnosis supports prescriptive actions. Tests

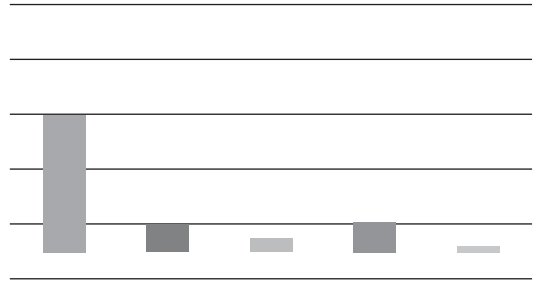


Figure 2.5 Probability revision given IQ scores.

are used in assessment to provide information that reduces uncertainty in decision-making, which leads to a judgement, which in turn leads to a prescriptive action.

The term *prescriptive action* is used here to represent the fact that assessment is not conducted for the purpose of getting scores on tests. Judgements and conclusions by themselves are useless unless such judgements guide future actions. The term *treatment* is not used but, rather, is considered a type of prescriptive action. Not all prescriptive actions must be physical acts; they can also be "states of knowledge." A clinician may do an assessment and make a judgement that a patient's memory is impaired, it has decreased over time, and this decline indicates dementia. In some cases the prescriptive action may be to inform client so the client can make necessary arrangements. In other cases it may be a referral for medication, assisted living, etc. As such, the term *prescriptive action* is used to indicate the actions that were taken, or belief states that were changed, as a result of the assessment judgement. Prescriptive action is a mediator between assessment and social outcome. Additionally, assessment for the purpose of writing reports that includes recommendations does not fully specify that recommendations are prescriptive actions, although often only indirectly related to assessment data. The results of assessment, and the interpretive process, should provide evidence to increase or decrease the probability of different hypotheses, which in turn lead to different prescriptive actions. As such, the pinnacle, or goal, of assessment is not test interpretation. The results of assessment, and the interpretive process, should test hypotheses that lead to different actions. The link between test interpretation, decision-making, prescriptive action, and outcome is rarely formulated as a unified model because often there is a high degree of contextual dependence in applied contexts. Neuropsychologists may conduct assessments to determine whether someone has suffered a brain injury, the nature of the injury, and the extent of functional loss. The type of judgement made depends on the prescriptive action, or purpose of the assessment. Suggesting a specialized intervention to remediate a cognitive processing weakness is of little value when the original purpose of the assessment was to determine whether the client is competent to stand trial! It is important that the prescribed action, judgement, and assessment process be in alignment. Although the contextual dependence of prescriptive action limits its specification because it may differ by context (and there are numerous contexts), it may still be specified in the abstract.

Figure 2.3 and the specification of judgement and prescriptive action as precursors to social outcome may help clarify some substantial problems in contemporary assessment literature. The problem is how to determine the utility of psychological assessment. Figure 2.3 also makes it clear that the utility of test in an assessment process cannot be directly determined by social outcome or benefit that results or does not result from the assessment. Assessment is several steps removed from beneficial client outcomes. Rather, the utility of tests depends on how they are used in a context to inform judgements that lead to different actions or outcomes.

Interventions are a type of prescriptive action that includes intentional manipulations to cause a change of some attribute or indicator in an intended direction. One difficulty in treatment is selecting an intervention from among numerous possible interventions. In school-based practice, numerous children having difficulty reading are prescribed phonological interventions, which are supported by research. Unfortunately, many of these children do not improve because they do not have problems in phonological processing, which a 10-minute test in phonological processing would have suggested. A child may do poorly in reading instruction, perhaps due to social-emotional problems like depression. Such a child may show improvement in reading skills as a result of reading intervention, although the underlying problem of depression remains and may affect future academic behavior. In such situations, what would be the value of administering a test that would have clarified the attributes of a child and would have in turn led to a better prescriptive action? Currently, there is no metric for determining this value. Similarly, there is no metric for reducing uncertainty in determining the underlying problem or selecting the appropriate intervention. Testing reduces the uncertainty in these possibilities. The use of testing to reduce the possibilities of error in defining the underlying problem of a child is not included in behavioral studies of treatment validity. Such studies often "assume" that a child's status is known (e.g., depression, reading problems, etc.) and then asks how would giving a test reduce depression or improve reading (see Fletcher, Lyon, Fuchs, & Barnes, 2007, as an example in reading). Testing provides information about the attributes of an individual that contributes to decision-making within an assessment process, which in turn contributes to interventions that influence outcomes.

Other factors also impact decisions, as do the actual prescriptive actions taken that are more

causally related to treatment outcomes. The problem here is analogous to that of measurement. There is a construct with natural attributes, and one must assign labels to it in order to study it. Namely, the process involved in clinical decision-making, prescriptive action, and social outcome must be pre-specified and structured as data. Testing provides information when test results reduce the uncertainty in decision-making possibilities. Testing need not reduce the probability of one category to certainty ($p = 1.0$) to be informative, but rather just change the distribution of possibilities (see previous example). Additionally, the process of "judgement" and "prescriptive action" need not be simply grouped under an umbrella of "interpretation" and assumed to be impenetrable to analysis. Measurement theory suggests a solution. Clarify the underlying attributes through theory, label them, and investigate.

Social Outcome

Of course, like judgements, prescriptive actions are not selected in a vacuum but rather linked to social utility. That is, a prescriptive action is selected because it is judged or predicted to result in some benefit. Traditionally, this has been framed as *consequential validity,* but as a line of validity evidence rather than, as indicated here, as a more central element of assessment. The reason why such social goals or outcome variables need greater representation in test development is because such goals provide feedback on how to construct the decision-making model. The decision-making model informs the type of validity evidence needed for a test, which in turn influences how a test is constructed, as demonstrated in the previous example of maximizing information value for decision-making thresholds.

Demonstrating how psychological assessment services provide utility in psychological outcomes has been a defining characteristic of contemporary psychological practice. Influenced by managed health care, evidence-based practice has focused on "outcomes" by which to evaluate psychological services (Maruish, 1994). Effectiveness in providing services is determined by the degree to which specified outcomes are obtained. The influence of this philosophy is vast, and an outcomes orientation has influenced everything in psychology, from standards in training, to insurance reimbursement from third-party payments.

The use of psychological tests has not escaped this scrutiny. Interestingly, there are conflicting opinions on the utility of assessment in impacting treatment outcomes. Some have suggested that assessment is of little to no value (Hayes et al., 1987), which is supported by many researchers with a behavioral orientation (Gresham & Witt, 1997; Reschly & Gresham, 1989). Similarly, some have suggested that outcomes should be the core aspect of test validity (Fuchs et al., 2002). Others have presented cogent arguments on the limitations of such an approach (Cone, 1989; Decker, 2008). Additionally, meta-analysis of more than 125 studies led to the conclusion that there is strong evidence for psychological test validity, that psychological test validity is comparable to that of many medical tests, assessment instruments provide unique information, and clinician decision-making is enhanced by the results of psychological tests (Meyer et al., 2001).

The misunderstanding inherent in approaches that dismiss the utility of psychological testing comes from a lack of specification in the application of psychological services. Namely, the role of decision-making is neglected. Administering Block Design from the Wechsler tests will not cause a desired outcome. However, results of such a test may provide information within the assessment process that requires clinician decisions to inform some course of action. Similarly, testing helps us record change as a result of intervention. Although traditional single-subject design methods are used, psychometric methods may also apply to interventions. The termed *intervention psychometrics* has been used to describe the application of psychometric theory to intervention methods (Decker, 2008).

Despite the inherent benefits in the focus on outcomes, there are some drawbacks. Perhaps the two most important are the two most general. First, the sole focus on any one thing inevitably leads to a neglect of other concepts. Second, singularity of focus often causes an oversimplification that creates a model unable to match the complexities of practical applications. Outcomes are important but perhaps no more so than methodologies determining service needs, adequate measurement representation of person-need, and adequate representation of the type of services matched to needs. Such measurement is needed if it can ever be determined that a particular configuration of matching needs to services through a diagnostic process creates a benefit beyond that which could be obtained through no diagnostic process or random matching.

An additional issue must be mentioned in the process of integrating data with social values. Similar to descriptions of assessment as top-down/bottom-up, or diagnostic/prescriptive, there are problems with describing assessment as driven by social values as a top-down process in assessment. Historic social

examples have shown the push of social values is not always just. The dichotomy of data as indicators of reality and social values as interpretive mechanisms has historically been a core theme to describe the relationship between science and religion. Science, as an attempt to describe the world as accurately as possible, and religion, as a prescriptive approach to how the world should be guided, have been at odds many times. Other examples can be given, but my guess is the reader gets the idea. Test validation, as such, may not be described as a methodical process involving reliability coefficients but may better been viewed as a "belief management" technique: that is, evidence is provided to support beliefs (i.e., clinical inferences), which in turn justify actions. *Validity* is a method of determining the degree of which beliefs concerning constructs can be "believed." However, given the current status of validity research, there is yet a procedure in which the quantification of beliefs can be attained. How much, or how many lines, of validity evidence are needed before one's action is selected over another? How are beliefs and actions to be connected? What if two contradictory belief systems are both supported by different lines of validity evidence? The Bayesian approach to hypothesis testing (previously presented) may serve as one technique to more explicitly represent clinician decision-making, which in turn helps make explicit the value of assessment. Currently, nothing in the current system of test validity exists to resolve these issues.

Figure 2.3 makes explicit judgement is linked to prescriptive action which is linked to social goals, which in turn are linked to theory. Table 2.4 provides different judgement, prescriptive action, and outcomes for different assessment contexts. Although it is difficult to define all possible values for each stage, it is possible to provide broad indictors for each stage of assessment.

Here the emphasis is on pre-specification of possible events at each stage. The events listed for these stages in Table 2.4 are simplified to the point of being irrelevant for the listed assessment applications, but more specific and sophisticated classification schemes exist for each of the areas (see Wodrich & Schmitt, 2006, for an example of school-based classification). Such models provide direct linkage of assessment to actions that may be taken as a result of testing. For example, an educator may solely be interested in identifying children who are at risk for reading problems. This implies a binary decision-making outcome ("at –risk" or "not at –risk"). The test used to make such a decision need not be a comprehensive measure that measures the entire range of reading capability. Rather, such a test need only maximize information at the decision-making threshold for making a decision as to whether a child is at risk or not at risk.

Conclusion

This chapter broadly reviewed measurement theory, scale development, testing, and assessment. The chapter was divided into two broad areas to represent distinct phases of testing involving test development and test application. The integrated role of testing with measurement and assessment components was demonstrated. Theories of measurement are reviewed with the use of the Item Response Theory, not only for the purpose of objective measurement but as a basic model to analyze how personal attributes interact with test stimuli. Interpretive phases of tests within an assessment process are described, which include decision-making, prescriptive action, and social outcomes. This

Table 2.4 Linking assessment and outcomes through judgment and actions

Assessment Purpose	Judgement	Prescriptive Action	Outcome
Disability	Disability present Disability absent	Not eligible Eligible	Educational modification
Forensic (competence to stand trial)	Competent Incompetent	Stand trial Do not stand trial	Social justice
Risk	At risk Not at risk	Protect No protection	Safety
Neuropsychological	Brain injury No brain injury	Remediation/ accommodation	Life adjustment
Intervention	Determine problem	Intervene on problem	Improvement

extension is based ambiguous concepts inherent in contemporary test theory. The interconnected "network" of concepts in testing contributes to the complexity of understanding testing, but nonetheless is necessary. Testing depends on measurement, which in turn depends on the theoretical basis of a construct. Assessment depends on testing, or typically, multiple tests for interpretation. Interpretation is a sub-part of a decision-making process in which some action, such as an intervention, is to be implemented. Interventions influence outcomes, which are evaluated by social goals and values.

Perhaps due to the complexities involved with testing, numerous misunderstandings have occurred that result not only in controversy in research but in misapplication of tests in society. Furthermore, the historic difficulties of not clarifying "interpretive" issues in testing have led to variations in the application of psychological testing, with some of the variability extending into the misapplication of practice. One need only look at the historic use of IQ measures as an example. Despite the fact the measures of intelligence are perhaps the greatest successful application of psychology, the negative connotations that surround lower IQ have created a negative impression on society, and it is doubtful that the term *IQ* will become vindicated. Additionally, the disconnection between how a test is developed and how it is used has led to criticisms involving the treatment utility of tests. This issue was indirectly addressed in this chapter by providing a clarification of why practitioners perceive the value of tests but that value is not captured in research studies. The value of assessment in treatment is not a result of assessments directly causing a change in functional status. Rather, tests used in assessment reduce uncertainty in the decision-making process, which leads to prescriptive action that causes change in social outcomes. This provides an explanation for why assessments have not been adequately tested within a treatment validity paradigm, but evidence is still required to demonstrate the decision-making utility of assessment for particular applications. A Bayesian model is reviewed as a demonstration of how this may be accomplished.

The purpose of reviewing the Rasch model in detail was to demonstrate how qualitative data can be quantified and converted to a unit of measurement. Similarly, most diagnostic classification schemes, although categorical, can be placed along a dimensional continuum (e.g., symptom severity, number of symptoms). Additionally, the value of social outcomes can be rank ordered. Most would agree that full recovery or return to normal parameters of functioning is a more desired outcome for a client than simply being informed of the diagnosis, which in turn is more valuable than not knowing what the problem is at all. Providing a unitary metric of social outcomes to monitor treatment progress may be useful.

Messick, in a review of different perspectives of construct theory, concluded with:

> The use of constructs to develop intuitive systems of meaning for observed relationships appears to be a fruitful heuristic if buttressed by convergent and discriminant evidence to validate the interpretation of test scores as measures of constructs and to validate observed relationships as reflective of meaningful connections in the nomological system (p. 587).

The model presented here for integrating testing with measurement and assessment may similarly be viewed as a "fruitful heuristic" in clarifying the utility of psychological assessment.

Future Directions

There are several future implications for research and practice based on the model presented in this chapter. Test construction may benefit from more focus and clarification of the social outcomes specified by the theory that guides test development. Similarly, clarification of the information value of a test is needed, as well as increased focus on the theoretical analysis of the resulting decision-making. That is, what service is to be provided and what is its benefit.

Another implication is that the assessment field needs to develop better metrics of "information." Such metrics exist but are not a part of mainstream psychometrics. The study of information was most formally begun by Claude Shannon (Shannon, 1948). The intended applications of the study of information were in the digital transmission of communication channels. Numerous attempts have been made to apply information theory to topics in psychology, with only a few successes (Luce, 2003). In contemporary research, its most important applications have come from statistics. For purposes of this chapter, the importance of information is that information can be formally measured. Central to its conceptualization is the statistical probability of an event to determine the likelihood of an actual event. Essentially, information theory quantifies statistically rare events as more informative. Entropy is maximized when a system of variables is completely random. As events become more orderly, entropy decreases.

Finally, another future implication of the model presented here is to provide "scale" value to prescriptive action and social outcomes. Just as measurement requires a representation of an attribute, unified models of assessment need better representations of judgement, prescriptive actions, and social outcomes. Representing attributes of these stages would enhance the investigation of how these processes are involved in assessment and would make clear how they contribute to social outcomes. Until such processes are made explicit, they will continually be viewed as a "black box" and either held in high esteem by some or disregarded by others.

Future Directions

1. How could commercial test developers assist clinicians who would want to use Bayesian models of diagnostic decision-making?

2. Is it possible to develop an abstract clinician decision-making model that fits most situations in which psychological tests are used?

3. How could it be determined that a more context-specific decision-making model is better than a general, all-purpose decision-making model?

4. What are the different values that the variable "social benefit" may take?

Acknowledgment

The author would like to thank Dr. Catherina Chang for reviewing this chapter.

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Bechtoldt, H. (1959). Construct validity: A critique. *American Psychologist*, *14*, 619–629.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*, 27–41.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, D. T., & Fiske, S. T. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Campbell, N. R. (1920). *Physics, the elements.* Cambridge, UK: Cambridge University Press.

Cone, J. D. (1989). Is there utility for treatment utility? *American Psychologist*, *44*(9), pp. 1241–1242.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Decker, S. L. (2008). Intervention psychometrics: Using norm-referenced methods for treatment planning and monitoring. *Assessment for Effective Interventions*, *34*(1), 52–61.

Elwood, R. (1993). Clinical discrimination and neuropsychological data. *The Clinical Neuropsychologist*, *7*(2), 224–233.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–186.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. *Test design: developments in psychology and psychometrics,* 195–218.

Embretson, S. E. (1999). *New rules of measurement: What every psychologist and educator should know.* Mahwah: NJ: Lawrence Erlbaum Associates.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics. *American Psychologist*, *61*(1), 50–55.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fiske, S. T., & Pavelchak, M. (1986). Category-based versus piecemeal-based affective responses: Development in schema-triggered affect. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation cognition: Foundations of social behavior* (Vol. 1; pp. 167–202). New York: Guilford Press.

Fletcher, J. M., Lyon, R. G., Fuchs, L. S., & Barnes, M. A. (2007). *Learning disabilities: From identification to intervention*. New York: The Guilford Press.

Franklin, R. D., & Krueger, J. (2003). Bayesian inference and belief networks. In R. D. Franklin (Ed.), *Prediction in forensic and neuropsychology: Sound statistical practices* (pp. 65–87). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Fuchs, L. S., Fuchs, D., & Speece, D. L. (2002). Treatment validity as a unifying construct for identifying learning disabilities. *Learning Disability Quarterly*, *25*, 33–45.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, *87*(3), 564–567.

Garb, H. N. (1998). *Studying the clinician: Judgement research and psychological assessment*. Washington, DC: American Psychological Association.

Garb, H. N., & Schramke, C. J. (1996). Judgement research and neuropsychological assessment: A narrative review and meta-analysis. *Psychological Bulletin*, *120*(1), 140–153.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you.* New York: Simon & Schuster.

Gresham, F. M. (2002). Responsiveness to intervention: An alternative approach to the identification of learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 467–519). Mahwah, NJ: Lawrence Erlbaum.

Gresham, F. M., & Witt, J. C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. *School Psychology Quarterly*, *12*(3), 249–267.

Guttman, L. (1977). What is not what in statistics. *The Statistician*, *26*, 81–107.

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist*, *42*, 963–974.

Kamphaus, R. W., & Campbell, J. M. (2006). *Psychodiagnostic assessment of children: Dimensional and categorical approaches*. Hoboken, NJ: John Wiley & Sons.

Kamphaus, R. W., Reynolds, C. R., & Imperato-McCammon, C. (1999). Roles of diagnosis and classification in school psychology. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (3rd ed.; pp. 292–306). Hoboken, NJ, US: John Wiley & Sons Inc.

Kazdin, A. E. (2003). *Research design in clinical psychology*. Boston: Allyn & Bacon.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Vol. 1: Additive and polynomial representations.* New York: Academic Press.

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, *18*(1), 89–109.

Lewicka, M. (1988). On objective and subjective anchoring of cognitive acts: How behavioral valence modifies reasoning schemata. In: W. J. Baker, L. P. Mos, H. V. Rappard, & H. J. Stamm (Eds.), *Recent trends in theoretical psychology.* New York: Springer-Verlag, 285–301.

Lewicka, M. (1997). Is hate wiser than love? Cognitive and emotional utilities in decision making. In R. Ranyard, W. R. Crozier & O. Svenson (Eds.), *Decision making: Cognitive models and explanations* (pp. 90–108). New York: Routledge.

Lord, F. (1946). On the statistical treatment of football numbers. *American Psychologist*, *8*(750–751).

Luce, D. (2003). Whatever happened to information theory in psychology? *Review of General Psychology*, *7*(2), 183–188.

Luce, R. D., & Suppes, P. (2001). Representational measurement theory. In J. Wixted & H. Pashler (Eds.), *The Stevens handbook of experimental psychology* (3rd ed.; Vol. 4; pp. 1–41). Hoboken: John Wiley & Sons.

Maruish, M. E. (1994). Introduction. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 3–21.). Hillsdale, NJ: Lawrence Erlbaum Associates.

McFall, R. M., & Townsend, J. T. (1998). Foundations of psychological assessment: Implications for cognitive assessment in clinical science. *Psychological Assessment*, *10*(4), 316–330.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology*, *50*, 215–241.

Meehl, P. E., & Rosen, A. (1955). Antecedents probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–216.

Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist, 35*(11), 1012.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . & Read, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist, 56*(2), 128.

Michell, J. (1986). Measurement scales and statistics: A class of paradigms. *Psychological Bulletin*, *100*(3), 398–407.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Oxford, England: Nielsen & Lydiche.

Reckase, M. D. (2000). Scaling techniques. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (3rd ed.; pp. 43–64). Oxford, England: Elsevier Science Ltd.

Reschly, D., & Grimes, J. (2002). Best practices in intellectual assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (Vol. 4; pp. 763–773). Washington, DC: National Association of School Psychologists.

Reschly, D. J., & Gresham, F. M. (1989). Current neuropsychological diagnosis of learning problems: A leap of faith. In *Handbook of clinical neuropsychology* (pp. 503–519). New York: Plenum.

Reynolds, C. R. (1986). Measurement and assessment of childhood exceptionality. In I. B. Weiner, R. T. Brown, & C. R. Reynolds (Eds.), *Wiley series on personality processes. Psychological perspectives on childhood exceptionality: A handbook* (pp. 65–87). New York: Wiley-Interscience.

Reynolds, C. R., Lowe, P. A., & Saenz, A. L. (1999). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (3rd ed.; pp. 549–546.). New York: Wiley.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego: Jerome M. Sattler, Publisher.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations.* San Diego: Jerome M. Sattler, Publisher.

Schwager, K. W. (1991). The representational theory of measurement: An assessment. *Psychological Bulletin*, *110*(3), 618–626.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*, 379–423.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, *161*, 849–856.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, *96*, 394–401.

Wodrich, D. L., & Schmitt, A. J. (2006). *Patterns of learning disabilities*. New York: The Guilford Press.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca: Riverside Publishing.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

CHAPTER

3 Measurement and Statistical Issues in Child Assessment Research

Matthew R. Reynolds *and* Timothy Z. Keith

Abstract

This chapter focuses on measurement and statistical issues in child psychological assessment research. Topics with worked examples include multiple regression, confirmatory factor analysis, the Schmid-Leiman transformation, measurement invariance, and MIMIC models. Comparisons are made between simultaneous and sequential regression, higher-order and hierarchical factor models, and multiple-group mean and covariance structure analysis and MIMIC models. The chapter also discusses issues such as dealing with missing data, formative versus reflective measurement, and categorical versus continuous latent variables.

Key Words: hierarchical models, measurement invariance, higher-order confirmatory factor analysis, MIMIC models, multiple regression

Introduction

Research in psychological assessment permeates the practice of psychology. Psychological assessment relies on psychological measurement research, which in turn relies on psychological theory. Psychological assessment research itself is a broad topic. Here we will skip statistical and measurement basics because these topics are well explicated elsewhere (e.g., McDonald, 1999). The chapter will begin with a discussion of multiple regression, an increasingly popular method that is not always well understood. Following multiple regression, a variety of topics such as dealing with missing data, confirmatory factor analysis, and measurement invariance will be reviewed. Worked examples will be provided for some of them.

Some of the issues discussed in this chapter may be considered advanced, though they have been around for a few decades. Modern computing, however, makes for easy implementation of these procedures. In fact, we believe some of these advances should be part of standard practice in psychological assessment research. It is our intention to keep the presentation style as non-technical as possible with the goal of raising awareness; applied examples will demonstrate interesting and important questions that may be asked and answered using these methods. The references provide much more technical detail and should be consulted by readers who are interested in learning more. A theme of this chapter is that researchers should design research that is consistent with theory, and use methods to critically test those theories. To do so, it is essential to have good theories to draw from, and to have tools useful to test them. We believe the topics covered in this chapter include some of those tools.

Multiple Regression

Most readers will be familiar with multiple regression, a popular analytical tool that allows researchers to answer questions about the effects of presumed causes on presumed effects. Two popular approaches, simultaneous and sequential regression, will be compared and contrasted in an example.

The two approaches are sometimes either treated as entirely different methodologies or are applied rigidly according to a cookbook set of rules. In the example it will be demonstrated that the statistical processes underlying the approaches are not different, and that the differences between the two are often found in interpretation only. Understanding the similarities and differences of the two approaches is useful so that the appropriate approach can be applied to specific research questions.

Data from the Early Childhood Longitudinal Study–Kindergarten (ECLS-K), a large-scale, publicly available dataset, were used in this example. Four variables were used to explain the science achievement of fifth-grade students: sex of the student (Sex, dummy coded so that boys = 0 and girls = 1), first-grade reading ability (Prior Reading), self-perceived competence in all academic areas (Perceived Competence), and teacher ratings of children's approach to learning (Learning Approach), which includes behavior such as attention and organization skills. The ECLS-K includes these variables among a multitude of others.[1] The sample for this example included 1,027 children.

There are three common approaches to multiple regression: simultaneous, sequential, and stepwise. Stepwise regression will not be illustrated here. The method is used only for predictive purposes, not explanatory (we will discuss these two purposes in more detail later on), and we do not encourage its use. There are numerous reasons that this atheoretical approach should be avoided, including the fact that it capitalizes on chance findings due to random sampling fluctuations and because it does not require a researcher to think (see Keith, 2006; Thompson, 1995).

Simultaneous Regression

Simultaneous regression (also known as *standard multiple regression* or *forced-entry multiple regression*) is commonly used in explanatory research. Simultaneous regression produces estimates of the direct (unique) effect of the explanatory variables on the outcome variable. Specifically, correlations among the explanatory variables are accounted for so that the unique effects of the explanatory variables are estimated after the effects of the other variables have been removed. The method is useful for comparing the relative influences of variables on a single outcome variable of interest. All of the explanatory variables are entered or "forced" into the regression equation simultaneously. It is typical that R^2 (i.e., the proportion of the outcome variance explained by the optimal linear combination of predictors) and standardized (β) and unstandardized regression (b) coefficients are interpreted in simultaneous regression.

In this example, Science Achievement was regressed on Sex, Learning Approach, Prior Reading, and Perceived Competence. The linear combination of these variables explained 37% of the variation in Science Achievement (R^2 = .37, F [4, 1022] = 149.03, p < .01). When the other variables were held constant, Sex (b = −3.98, β = −.21, p < .001), Learning Approach (b = 2.47, β = .17, p < .001), and Prior Reading (b = .54, β = .52, p < .001) each had statistically significant effects on Science Achievement. The effect of Perceived Competence (b = .19, β = .01, p = .61) was not statistically significant when the other variables were statistically controlled. A qualitative comparison of the standardized effects shows that Prior Reading (β = .52) was the most important influence on subsequent Science Achievement.

Sequential Regression

In sequential regression, the explanatory variables are not forced into the equation at once; rather, they are entered sequentially in what are often referred to as *blocks* (this type of regression is also often referred to as *hierarchical regression*). The order of entry has important interpretative implications and should be based on a researcher's knowledge or beliefs about causal order. In this example, Sex was entered in the first block, Prior Reading was entered in the second block, and Learning Approach and Perceived Competence were entered simultaneously in the third block.

Sex was entered in the first block because it has time precedence over the other explanatory variables. For example, reading ability in the first grade does not explain a child's sex, but a child's sex may have important implications for first-grade reading ability. Perhaps one could *predict* a child's sex by including first-grade reading in a prediction equation, but the interest here is in *explanation*. Moreover, the prediction of a child's sex based on reading ability is uninteresting and does not make sense.

In block two, Prior Reading was entered. Prior reading is likely to influence fifth-grade science achievement because students who are better at reading will read more and build their stored knowledge base. It may also affect perceived academic competence and learning approaches in fifth grade, which are developed from prior experiences. Learning Approach and Perceived Competence were entered

together in the third and final block, based on the belief that both of the variables combined add to the explanation of Science Achievement. They were entered last because lack of organizational skills and inattentiveness probably interfere with learning. Moreover, perceived academic competence is a general construct, most likely acquired over years of schooling, and thus this perceived competence should influence engagement and performance in specific academic areas like science.

Proper entry is critical in sequential regression, so it is worth considering further. It is plausible that how much a student knows in science influences teacher ratings of that student's approach to learning. If a student lacks knowledge in science, then that student may appear inattentive and unorganized in science class. The ratings are based on general academics, however, and not science, so the original order makes sense. The important point is that researchers must carefully consider order of entry and must also be prepared to defend their decisions (and defend them much more rigorously, within a theoretical framework, than we have done here).

Given this emphasis on the order of entry in sequential regression, our decision to enter Learning Approach and Perceived Competence together in one block may seem curious. Such a decision may suggest that the researchers are unsure of the proper causal sequence of the variables, or alternatively, that they believe that the variables assess related, overlapping constructs, and are interested in the effect of that overarching construct. Researchers should examine these kinds of decisions, or non-decisions, because they have important interpretative consequences. Finally, researchers should be prepared to defend their reasoning for including variables and omitting potential common causes in a regression. That is what solid research is about, and the omission of important common causes renders interpretation of the regression coefficients invalid.

For the sequential regression, Sex was entered into the equation first. ΔR^2 was used to determine if there was a statistically significant improvement in the proportion of variance explained in Science Achievement after Sex was included. ΔR^2 (.016) was statistically significant (F [1, 1025] = 17.09, $p < .01$). Sex improves the explanation of Science Achievement above that of having no explanatory variables in the equation. Although ΔR^2 is most commonly interpreted in sequential regression, some researchers also interpret the coefficient associated with each variable as it is added to the equation. For the current example, these are $b = -2.44$ and $\beta = -.13$, and the b suggests that, on average, girls score 2.44 points lower on the Science test than do boys (the negative coefficient means that girls, coded 1, score lower than boys, coded 0). *If* variables are entered in the proper order, these coefficients represent the *total* effect of Sex on Science Achievement, and this effect is different from the direct effect of Sex obtained from the simultaneous results. (We will return to this issue later.)

Second, Prior Reading was added to the equation, resulting in a statistically significant $\Delta R^2 = .32$ (F [1, 1024] = 506.10, $p < .01$). Explanation of Science Achievement was improved beyond the proportion of variance explained by Sex alone. The effect of Prior Reading ($b = .60$, $\beta = .57$) was large. This effect represents the *total* effect of prior Reading on Fifth-Grade Science Achievement, both directly and possibly indirectly through the soon to be added variables Learning Approach and Perceived Competence. Someone unaware of what the regression coefficients in sequential regression represent might be confused by the coefficients related to Sex produced at this step. The coefficients for Sex have changed ($b = -3.31$, $\beta = -.17$). Does this mean that Sex is more important than it was previously? We will address this issue in more detail below.

Last, Learning Approach and Perceived Competence were added as a block, resulting in a $\Delta R^2 = .028$ (F [2, 1022] = 22.27 $p < .01$) that was statistically significant. The addition of these two variables, in combination, improves the explanation of individual differences in Science Achievement. Regression coefficients estimated for Sex ($b = -3.98$, $\beta = -.21$, $p < .001$), Prior Reading ($b = .54$, $\beta = .52$, $p < .001$), Learning Approach ($b = 2.47$, $\beta = .17$, $p < .001$), and Perceived Competence ($b = .19$, $\beta = .01$, $p = .61$) in this step were identical to those obtained in the simultaneous regression. We now have several sets of coefficients that could be interpreted from the sequential regression. If we are interested in interpreting the effects, which are appropriate, and which should we interpret? Perhaps this interpretation is best explained by comparing the results with the results from the simultaneous regression.

Simultaneous and Sequential Regression: A Comparison

Path diagrams will be used to help compare simultaneous and sequential regression. A path model of the simultaneous regression is shown in Figure 3.1. In the diagram, the rectangles represent observed variables; the arrows, or paths,

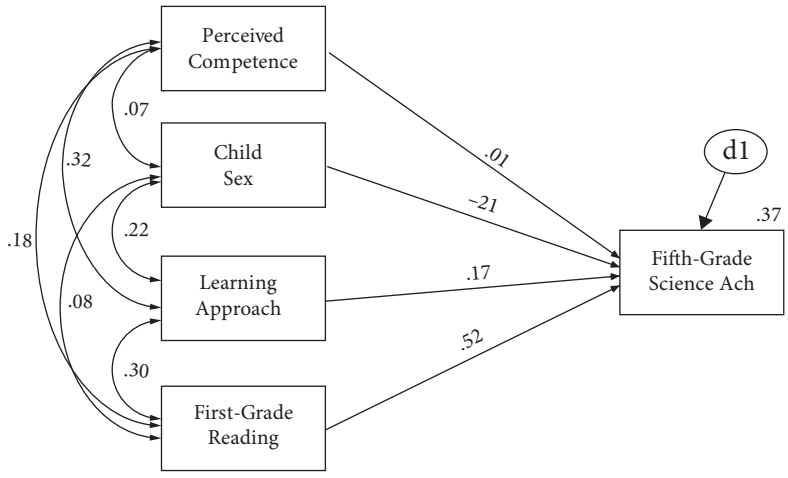


Figure 3.1 Simultaneous Regression in Path Form.

show a directed relation between the variables; the double-headed arrows represent a non-directive relation (i.e., correlation); and the oval represents a disturbance, commonly referred to as a *residual variance* in regression. In structural equation modeling, ovals typically represent latent, or unmeasured, variables. In this example, the oval represents all influences on the corresponding measured variables other than those shown in the model; these influences are not measured or modeled, and may include measurement error, nonlinear effects, random unknown influences, and all other unknown influences on the outcome (Arbuckle & Wothke, 1999; Bollen, 2002).

In the simultaneous regression shown in Figure 3.1, the explanatory variables correlate with each other, and each explanatory variable has a path connecting it directly to the outcome (Science Achievement), representing the presumed effect of these variables on science achievement. These effects are direct. Because the interrelations among the variables are controlled (by allowing them to correlate), these effects are also referred to as *unique effects*. Whatever is not explained by the linear effects of the explanatory variables is captured in the disturbance.

Compare this path model to the sequential regression path models shown in the left side of Figure 3.2. In the sequential regression there are three steps corresponding to what happened at each block of variable entry. Sex is entered first (Figure 3.2, Block 1), Prior Reading second (Figure 3.2, Block 2), and Learning Approach and Perceived Competence third (Figure 3.2, Block 3). An obvious difference between simultaneous (Figure 3.1) and sequential regression (on the left in Figure 3.2) is that some of the non-directed arrows are now directed. The variables on the left side of Figure 3.2 show an order from left to right, and that ordering reflects the causal assumptions made in the sequential regression and justified, albeit rather weakly, earlier in this chapter.

Figure 3.2 shows two types of path models, with the disturbances not included. The models on the left illustrate the causal reasoning underlying our sequential regression coefficients. The path models on the right demonstrate the coefficients associated with each step that are produced in the output. Although most analyses using sequential regression focus on R^2 and ΔR^2 interpretations, we will focus on the interpretation of coefficients from the regression because these effects are often confused. Note, however, that the R^2 values are the same regardless of whether the example on the left or right is used (these values are shown on the top right of the Science Achievement outcome variable in each Figure). It is instructive to uncover what is happening during a sequential regression, and this becomes clear with a focus on the coefficients estimated at each block.

Starting on the left in Figure 3.2, Block 1 shows Science Achievement regressed on Sex. The regression coefficient ($\beta = -.13$) is interpreted as the total effect of Sex on Science Achievement (generally, we would interpret the unstandardized coefficient when focusing on a dummy variable, but the standardized coefficients are used in subsequent blocks, and so will be used in this first block. See Keith, 2006, for guidelines for interpreting standardized versus unstandardized coefficients, and Table 3.1 for the unstandardized coefficients).

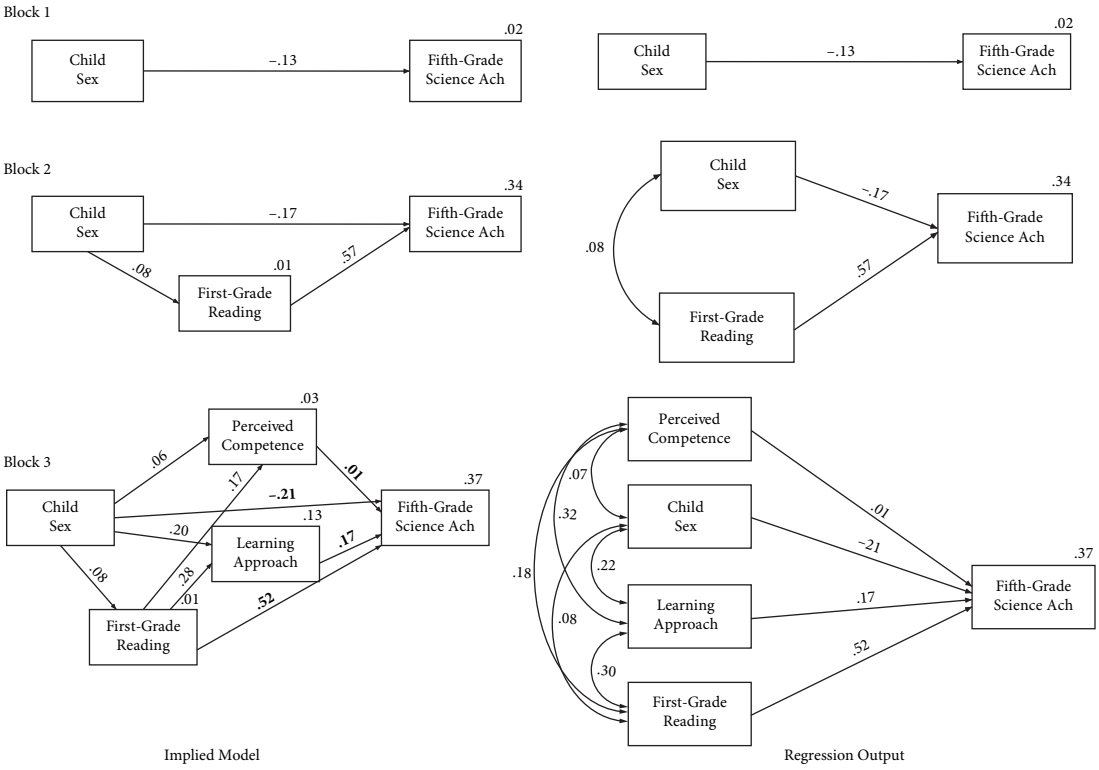


Figure 3.2 Sequential Regression Comparison in Path Forms.

In Block 2, Prior Reading was added. Shown in the second model on the left, Sex had a direct effect on Science Achievement and on Prior Reading, and thus through Prior Reading, an *indirect* effect on Science Achievement. On the right hand side of Figure 3.2 in Block 2 is the model that was *actually* run in Block 2, and the regression coefficients produced in the computer output. It may look like a simultaneous regression, and indeed it is, but with only two predictors. The standardized effect (β = –.13) associated with Sex in Block 1 was different from the standardized coefficient (β = –.17) in Block 2. The total effect from Block 1 (β = –.13) is now split into the direct effect (β = –.17) and the indirect effect, with the indirect effect equalling the path from Sex to Prior Reading times the path from Prior Reading to Science Achievement (.08 × .57 ~.04). The indirect

Table 3.1 Comparison of Direct Effects from Simultaneous Regression and Total Effects from Sequential Regression

Variable	Direct effects obtained in simultaneous regression		Total effects obtained in sequential regression	
	$b(SE_b)$	β	$b(SE_b)$	β
Sex	–3.98(.49)	–.21	–2.44(.59)[a]	–.13
Prior Reading	.54(.03)	.52	.60(.03)[b]	.57
Learning Approach	2.47(.40)	.17	2.47(.40)[c]	.17
Perceived Competence	.19(.38)	.01	.19(.38)[c]	.01

Note: [a] From Block 1; [b] From Block 2; [c] From Block 3. Note that the effects for Learning Approach and Perceived Competence are the same across models because these variables were entered in the last block of the sequential regression.

effect is not calculated in the regression output, but this indirect effect is easily obtained by subtracting the direct effect (–.17) in Block 2 from the total effect (–.13) in Block 1 = .04. To answer the question posed earlier, the apparent effect of Sex does increase from the first to the second model because part of the *total* effect of Sex is explained by Prior Reading, and in this case, the indirect effect is positive, while the direct effect is negative. That is, girls have higher prior reading scores. The effect of Prior Reading on Science Achievement is also estimated (i.e., $\beta = .57$). If the causal order is correct, this effect represents the total effect of Prior Reading on Science Achievement. In Table 3.1, the direct effects interpreted in a simultaneous regression and total effects from the sequential regression are shown for comparison.

Last, Perceived Competence and Learning Approach were entered. On the left in Figure 3.2, Block 3, the estimates actually produced in the regression output in Block 3 are bolded. The estimates are identical to the direct effects obtained in a simultaneous regression. And of course a simultaneous regression is exactly what is shown on the right in Block 3. Calculations could be used to estimate the indirect effects. But if a researcher was really interested in all of these effects, this procedure can easily be performed in a structural equation modeling (SEM) program so that the direct, indirect, and total effects are all calculated (and statistical significance can be tested). The model in the SEM program would probably be specified to match the model in the left of Figure 3.2, Block 3.

We hope this illustration allowed the reader to make some mental connections between the two approaches. Why go through the trouble of illustrating these similarities and differences? First, many researchers are interested in the *unique* effects of variables on some outcome. That is, researchers are often interested in the effect of the explanatory variable of interest on the outcome, *controlling for other variables* in the model. They often use sequential regression and estimate the unique effect by adding the variable in the last block. It should be clear now, however, that these effects are easily captured in either simultaneous or sequential regression. The multiple blocks in a sequential regression are not required if this is the interest, even though sequential regression is often used by researchers for this purpose. Second, if researchers have a causal ordering in mind and they want to use sequential regression, it is important that they understand the nature of the coefficients they are interpreting. In fact, drawing out a path diagram as shown on the left of Figure 3.2 would be beneficial so it is clear what types of effects are obtained. We urge both users and consumers of sequential regression research to routinely draw the models underlying their and others' regressions. Of course, if one is capable of drawing a model, it may be easier to simply analyze the model via a structural equation modeling program!

Summary of Multiple Regression

Keith (2006) outlined additional similarities and differences between the two regression approaches, and a few will be mentioned here. There are many similarities, and even the differences are not necessarily true differences, but rather are differences in rigidly applied conventional interpretations. Note however that ΔR^2 is generally used as a test of statistical significance and interpreted in sequential regression.[2] It is common, however, to see regression coefficients also reported and interpreted in sequential regression. R^2 and the statistical significance of the regression coefficients are generally interpreted in simultaneous regressions. R^2 in simultaneous regression is identical to the R^2 obtained in the final step in sequential regression when all of the explanatory variables are included. When coefficients are interpreted, however, it should be noted that sequential regression is focused on total effects, and simultaneous regression is focused on direct effects. Simultaneous regression also allows comparisons of the relative (direct) effects using standardized coefficients, and can typically be used to answer questions researchers use sequential regression to answer. And lastly, sequential regression might be considered when testing for moderators or for curves in the regression plane, but only if the researcher is interested in an overall test of an interaction effect or several interaction effects in a block. (This issue will be discussed more in the section on moderation.)

Simultaneous and sequential regression may be used for either explanation or prediction. In explanation, the regression coefficients represent the effects of the presumed causes on the outcome variable of interest, given the adequacy of the model. Prediction equations can also be obtained so that optimal linear combinations of variables can be used to predict an outcome. In our experience, most researchers are interested in explanation even though they may pretend that they are only interested in prediction. One typical scenario

(and we encourage the reader to do a quick literature search to find a multitude of examples) is for authors to discuss prediction in the introduction and results, and then switch to explanation when the findings are discussed. This is the research version of a bait and switch! The researcher may not even know that a switch has taken place, but any time a researcher makes a statement along the lines of "this research suggests that increases in variable x would lead to increases in variable y," he or she has made an explanatory interpretation. Researchers should ask themselves whether the purpose of their research is *really* prediction or whether it is really explanation before they begin the process (Keith, 2006). To thine own self be true!

Lastly, although we have yet to note *explicitly* that researchers should use the method to match the purpose of their research, we are doing so here. *A priori* conceptual models are associated with structural equation modeling, but it should be obvious that such models are similarly important in regression. Therefore, researchers need to decide what type of regression, or combination of regressions, will be most consistent with their theoretical models.

Mediation

There are plenty of excellent sources on mediation, so this introduction will be brief (Mackinnon, 2008; Shrout & Bulger, 2002; see also Kristopher Preacher's website: www.quantpsy.org). Mediation occurs when a variable that is between the presumed cause and outcome partially or fully explains the effect of the presumed cause on the outcome. A test of mediation is generally considered a test of the indirect effect of one variable through another variable. Although sequential regression may be used to get an idea about or sometimes test mediation, tests of mediation are probably better off performed in structural equation modeling programs. The study of mediating variables is important because these variables provide an understanding of change mechanisms; for example, an understanding of how treatment effects arise. They are especially interesting because they help explain how outcomes come about.

Moderation

Multiple regression may be used to test for interaction effects, or what is commonly referred to as *moderation*. Moderation is commonly tested via sequential regression. In child assessment research, moderation is often used to test predictive bias or invariance. For example, do scores from a reading fluency measure predict reading comprehension equally well for boys and girls (see Kranzler, Miller, & Jordan, 1999)? That is, does sex moderate the relation between reading fluency and reading comprehension? To test for an interaction using multiple regression (i.e., moderated multiple regression), first, a new variable is created as the cross-product of the two variables of interest (e.g., sex multiplied reading fluency scores). Centering any continuous variables prior to creating the cross product is also often used to improve interpretation (Aiken & West, 1991; Keith, 2006), so for example the reading fluency scores would be centered. Next, the main effects (e.g., sex and reading fluency scores centered) are entered in the first block; the cross-product (sex times reading fluency scores centered) is entered in the second block. This cross-product, or interaction term, is added to determine if the interaction term adds unique information to the explanation of the outcome variable. If it adds to the explanation, then it may be said that the effect of one variable (reading fluency) on the outcome (reading comprehension) *depends* on or is moderated by the other variable (sex). Especially with a single cross-product, this analysis could also be performed in simultaneous regression, with the statistical significance of the unstandardized coefficient used as a test of significance for the cross product. The use of sequential regression, however, allows an omnibus test for multiple cross products or the calculation of an effect size (viz., ΔR^2) for the interaction term (Turrisi &Jaccard, 2003, p. 86). See Keith (2006) for more examples of using regression to test for moderation.

Missing Data

Missing data are a perennial concern for researchers. Advances in statistical theory in recent years, along with excellent and accessible reviews of missing data assumptions and techniques, have substantially improved our knowledge of how to handle missing data (e.g., Graham, 2009; Schafer & Graham, 2002; Wothke, 2000). In fact, rather than just *dealing* with missing data, implementing planned "missingness" into research designs may be a cost-effective, efficient method of collecting data (McArdle, 1994). We will provide a brief explanation of missing data assumptions and techniques, but we encourage the reader to refer to some of the excellent sources for more in-depth and informed coverage (Enders, 2010; Graham, 2009; Schafer & Graham, 2002; Wothke, 2000).

There are three general mechanisms assumed to underlie missing data: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR; Little & Rubin, 1987; Rubin, 1976). MCAR requires the assumption that missing data do not differ from those that are non-missing. Say a researcher was interested in studying the effects of IQ and motivation on achievement. After the data were collected, the researcher noticed that scores from the motivation variable were missing for several cases. If data were MCAR, then the missingness of motivation cases was unrelated to motivation scores themselves, as well as to the IQ and achievement scores. The MCAR assumption is required for researchers to use the common deletion methods of handling missing data (i.e., pairwise and listwise deletion). If the assumption is met, then the biggest concern about deleting cases should be a loss of statistical power. If the assumption is not met, then parameter estimates, such as means and regression coefficients, and standard errors of those coefficients, may be inaccurate.

MAR, the second assumption, implies data are missing at random. Indications of why data are missing, however, may be found in the other variables included in the dataset. For example, if motivation scores were missing, the missingness can at least be explained partially by IQ scores, achievement scores, or both. That is, it may be that higher-achieving individuals were more likely to answer the motivation questions. If data are MAR, modern methods such as maximum likelihood estimation and multiple imputation may be used to obtain unbiased estimates. When data are MAR, but the deletion methods are used in analysis, parameter estimates are likely to be biased in that they over- or underestimate the population values, and this bias will probably to be difficult to detect.

The third possibility is that data are missing not at random, or MNAR. This type of missingness presents a problem. The missing values depend on something not measured in the dataset, and the reasons for their absence are unknown or unmeasured.

Knowledge of these three underlying mechanisms provides a framework from which a researcher can work. The good news is that it is typical for the mechanism to be at least somewhat understood. Moreover, when a researcher errs in making the assumption of MAR (when MNAR is really the case), the effects on the estimates in the model may be minimal (Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997).[3] Finally, even if MCAR is assumed by the researcher, the methods discussed below are better to use than the outdated deletion and mean substitution methods because all of the cases can be used in the analysis.

Outdated Methods

Listwise deletion, pairwise deletion, and mean substitution are examples of commonly used, but outdated methods. Deletion methods are a simple way of handling missing data because cases are simply dropped from the analysis. Deletion, however, results in fewer participants and will result in biased estimates if the MCAR assumption is not met. Mean substitution, like it implies, involves substituting a mean for missing scores, and is a simple, outdated, and potentially hazardous approach to handling missing data. Although all cases are included in the analysis when mean substitution is used, this procedure should be avoided because it produces biased estimates. Because the same value (the mean) is substituted repeatedly for the missing values of that variable, the variance will be reduced, as will the relation of that variable with the other variables in the model (Wothke, 2000). Other outdated methods include regression-based and hot deck imputations, but these approaches also suffer from limitations. Rather than discussing these methods further, we recommend the application of modern methods and will focus on those (Schafer & Graham, 2002).

Modern Methods

There are a few modern model–based approaches to deal with missing data, including the expectation maximization algorithm (which uses a maximum likelihood approach) and multiple-group structural equation modeling, but here we will discuss two popular and relatively easy-to-implement methods: maximum likelihood (ML) and multiple imputation (MI). These model-based methods require the less stringent MAR assumption compared to the outdated methods that require MCAR.

ML is the first model-based method. Space precludes a detailed description of the procedure, but a few important points can be made (see Wothke, 2000). First, ML estimation does not impute individual values; rather, the parameter estimates are obtained from using all of the available information in the observed data. Second, in large-sample statistical theory it is well known that ML produces consistent estimates that reflect the population values when data are multivariate normal. Third, ML estimation (with missing data) is available in structural equation

modeling (SEM) software, and its implementation does not require additional work for the researcher. In fact, multiple regression models can be analyzed in SEM programs, making it easy to implement ML methods when data are missing. Last, and perhaps most important, ML results in similar or, more likely, more consistent and less-biased estimates than those obtained after performing deletion and mean imputation; these differences may be dramatic when the data are MAR (Wothke, 2000).

Multiple imputation (MI) is another model-based method. Like ML, the statistical theory of MI is established. In MI, rather than imputing one value for each missing value in the dataset, a set of values representing plausible values is imputed for each missing datum, creating several new datasets with different sets of these new plausible values. Analyses are conducted on each dataset, like they would be with a complete dataset, and the results are pooled. Valid statistical inferences can thus be made, as the results incorporate the uncertainty due to the missing data (Graham, 2009). Like ML, MI generally assumes multivariate normal data, although it seems to also handle multivariate non-normal data fairly well. Given a large sample size, the estimates from ML or MI should be similar. Many statistical programs now include programs for dealing with MI, making it fairly simple to implement.

Researchers should recognize that missing data are not something to be ignored, but something that should be dealt with thoughtfully. Trying to understand the mechanisms that underlie missingness can assist in a better understanding of the data that are available as well as those missing. MI and ML are two fairly simple ways to deal with missing data, even when large amounts of data are missing. Given the relative ease of implementation, these methods should be considered the standard since they outperform outdated procedures, allowing researchers to use all of their data.

Factor Analysis

Factor analysis is an invaluable tool for understanding latent constructs and evaluating validity, and is commonly used to evaluate psychological assessment (measurement) instruments. The purpose of factor analysis is to uncover latent psychological attributes that account for correlations among observed variables. Quite simply, factor analysis is useful for understanding whether an instrument measures what it is supposed to measure. Although commonplace and useful in assessment research, factor analysis and other complex methods cannot make up for lack of relevant theory, common sense, knowledge-base, and carefulness of a researcher. Factor analysis may be misused and abused, intentionally or unintentionally.

The two main types of factor analysis are confirmatory (CFA) and exploratory (EFA) factor analysis. We will focus on CFA in this chapter, with only a few comments on EFA. EFA is older, growing out of Spearman's early twentieth-century explorations of the nature of intelligence (Spearman, 1927). With EFA, researchers choose the method (e.g., principal factors, maximum likelihood), the criteria for selecting factors (e.g., eigenvalues greater than one, *a priori* knowledge), the criteria for meaningful loadings, and the rotation method, and then interpret the results. Each step requires judgement, and multiple factor solutions are often examined. If done well—by researchers who are careful in developing the measures, and who apply combinations of criteria, use good judgement, and have knowledge of the relevant literature—EFA can be an invaluable tool in uncovering latent variables that explain relations among observed variables. But, EFA may also be abused. It is not unusual to see researchers put little thought into the theory that guides measurement; gather data; use inflexible criteria for factor extraction and rotation; and interpret their findings as if they were revealed truth. The judgement required to be good at EFA is a feature, not a design flaw! For more information about EFA, readers should refer to other, excellent sources (e.g., Preacher & MacCallum, 2003; Wolfle, 1940).

One other topic worth mentioning concerning EFA is the distinction between factor analysis and principal components analysis (PCA). A component obtained in a PCA is different from a factor obtained in a factor analysis. A component is a composite variable. Factors are latent variables. Most psychological attributes are conceptualized as latent variables, not composites, and these attributes should be invariant across the different instruments designed to measure them. One of the long-standing critiques of PCA is that the components (i.e., composites) are not psychologically meaningful (Wolfle, 1940). Factor analysis is thus the appropriate tool to use in latent variable research, not PCA.

Second, the procedures are used for different purposes. PCA, a descriptive procedure, was developed for data reduction and to maximize the variance explained in observed variables. Factor analysis, a model-based procedure, was designed to uncover psychologically meaningful latent variables that explain the correlations among observed variables. Factor analysis thus analyzes the common variance,

separating it from the unique variance. Unique and common variances are not separated in PCA. These distinctions have not stopped researchers from substituting PCA for factor analysis. As some have noted, perhaps this is because even some popular statistical programs do not differentiate the two (see Borsboom, 2006, for a discussion). Although space precludes further discussion of this issue, there are other excellent treatments of the topic (e.g., Preacher & MacCallum, 2003; Widaman, 2007; Wolfle, 1940). For a demonstration of potential different findings related to the use of PCA—and outdated missing data methods—in applied research, see Keith, Reynolds, Patel, and Ridley (2008).

Confirmatory Factor Analysis

CFA requires a researcher to specify the number of factors and the pattern of zero and free factor loadings *a priori*. CFA is commonly used in psychological assessment research to address questions related the measurement of psychological constructs and construct validity. We will work through an example to demonstrate the usefulness of CFA in establishing construct validity in an individually administered intelligence test. Throughout the example, we will describe and deal with various issues that may arise when conducting CFA.

During the last 20 years there has been a shift in the development of intelligence measurement instruments; many developers now rely on underlying theory during the developmental phase. The shift represents a major advancement that has not only informed measurement, but has likewise informed research and theory (Keith & Reynolds, 2010). The most popular theory, or perhaps better described as a *taxonomy*, underlying the development of these instruments is the Cattell-Horn-Carroll (CHC) theory of intelligence (McGrew, 2009), a theory that combines Cattell-Horn's G*f*-G*c* theory (Horn & Noll, 1997) and John Carroll's three-stratum theory (Carroll, 1993).

The Kaufman Assessment Battery for Children–II (KABC-II; Kaufman & Kaufman, 2004) is an example of a popular measure of child and adolescent intelligence in which theory was used during the developmental phase. In fact, the KABC-II may be interpreted using either CHC theory or Luria's information processing theory. In our CFA examples, we will use the norming data from the KABC-II to evaluate the measurement structure of the test. The CFA models will be consistent with the scoring structure using the CHC theory interpretation only. The scoring structure for the KABC-II battery includes index scores for five CHC broad abilities and a general ability referred to as the Fluid-Crystallized Index (FCI), as well as a few tests to supplement the broad ability indexes. The five broad CHC index scores include G*c* (Knowledge), G*v* (Visual Processing), G*f* (Fluid Reasoning), G*lr* (Long-Term Retrieval), and G*sm* (Short-Term Memory). G*c* is measured with three subtests, G*v* with four, G*f* with two, G*lr* with four, and G*sm* with three. We should note that the standard battery has fewer subtests, and some subtests were included as supplemental tests. The supplemental tests were used in our CFA to maximize the information available.

The data used in this example were age-standardized scores obtained from adolescents who ranged in age from 15 to 18. The sample included 578 participants. There were missing values for a few of the cases. The MCAR assumption was tenable. Rather than deleting cases, however, we chose to include *all* of the cases in the analyses using ML in Amos (Arbuckle, 2006) to handle missing data. Untimed scores were substituted for timed scores because in previous research the timed scoring procedure has been shown to introduce construct-irrelevant variance (Reynolds, Keith, Fine, Fisher, & Low, 2007). First-order and higher-order CFA models were estimated, and these are described below.

FIRST-ORDER MODELS

Specification

A first-order CFA model with five factors representing the five broad-ability factor indexes is shown in Figure 3.3. Essentially, we are interested in answering this question: Does the hypothesized latent structure underlying the observed data match the KABC-II measurement (scoring) structure? We address this question empirically by explicitly matching our factor model to the five broad-ability indexes. In Figure 3.3, the ovals represent latent variables; rectangles represent the observed variables; directed arrows represent directed effects; and nondirected arrows represent correlations/covariances.

Each factor is indicated only by the specific subtests that make up that broad index (Figure 3.3). Relevant theory (Jensen, 1998) and the use of an overall test score would also suggest that the factors should be correlated; therefore, the first-order model allows for intercorrelations among the factors rather than specifying them as independent of each other. The G*lr* measurements each included a delayed recall version of the original test. The residual variances associated with the first measurement and corresponding delayed measurement were specified to correlate freely (e.g., Rebus with Rebus

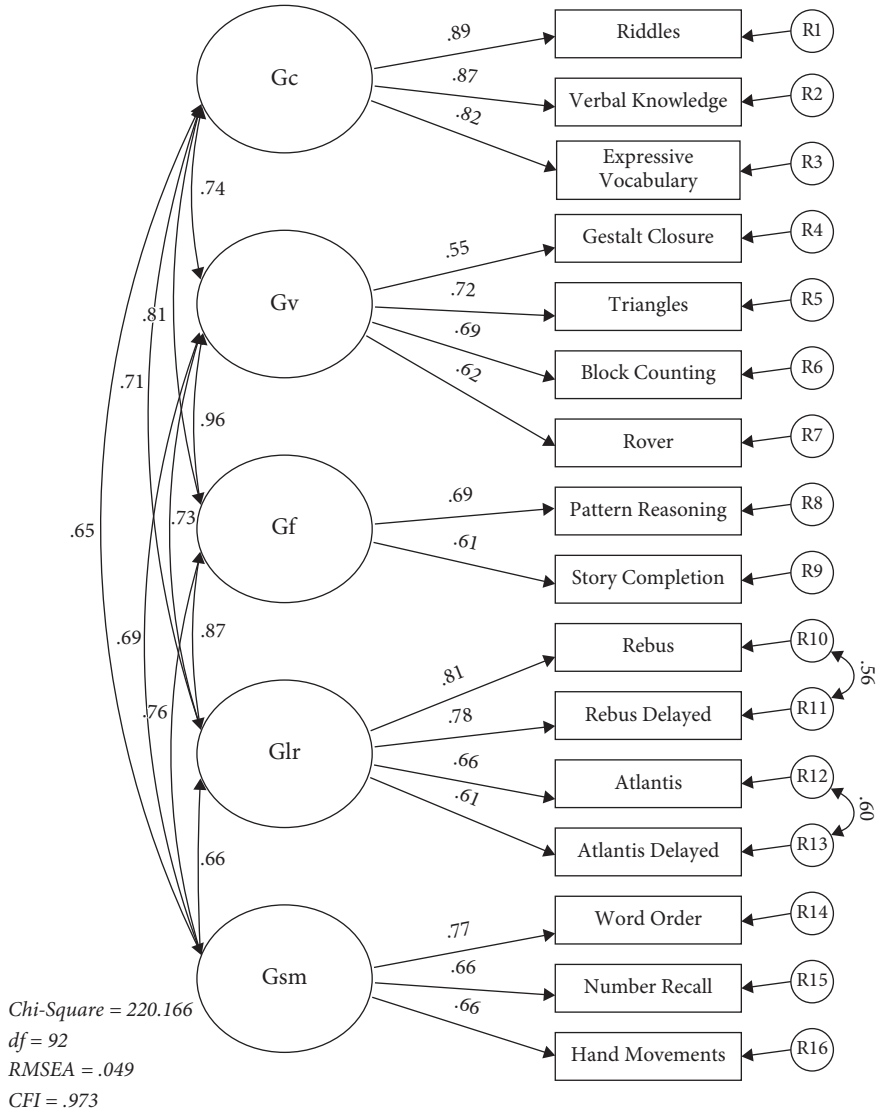


Figure 3.3 KABC-II First-Order Factor Structure with Standardized Loadings.

Delayed). These correlated specific factors represent overlap between the tests above and beyond what is explained by the G*lr* factor. Although our input model is not presented, the residual variance paths and one loading per factor were fixed to one so that the scales were properly set and the model was properly identified.

Model Evaluation

Indexes have been developed to assist researchers in evaluating fit. (More detailed explanation of these indexes is given elsewhere [e.g., Marsh, Hau, & Grayson, 2005].) For this example, model fit was evaluated with the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) and comparative fit index (CFI; Bentler, 1990), with values below .05 and above .95 indicating good fit, respectively. In addition, chi-squared (χ^2) was used to evaluate the fit of single models, and change in chi-squared ($\Delta\chi^2$) was used to evaluate competing nested models (i.e., models that can be derived by constraining additional parameters in a model). Chi-squared demonstrates excessive power to detect model misfit in large sample sizes, but, in general, the lower the χ^2 value relative to *df*, the better.

Results

The fit indexes for this measurement model indicated model fit was acceptable: χ^2 (92) = 220.17, RMSEA = .05, CFI = .97. The model with

standardized factor loadings is shown in Figure 3.3. The factor loadings were all substantial.[4] Like the directed paths in path analysis, one can interpret these loadings as regression coefficients. For example, the .88 standardized effect of G*c* on Riddles suggests that a one standard deviation increase in latent G*c* would result in a .88 standard deviation increase in a Riddles score.

The results support the interpretation of the broad-ability indexes on the KABC-II. It is not uncommon, however, for some subtests to measure more than one latent broad ability. Such subtests are often described as being *factorially complex* (McDonald, 1999). For example, to perform well on complex memory tasks requiring multiple steps, a person may employ a novel cognitive strategy to reduce the memory load, therefore reducing the memory requirement for successful performance. Novel problem–solving ability is associated with G*f*, and working memory and G*f* typically correlate strongly. The Hand Movements subtest on the KABC-II is an example of test that requires relatively complex memory; it is thus plausible that people high in G*f* could reduce the memory load of the task via their novel problem–solving abilities. To test this hypothesis, we loosened the strict assumption that all subtests measure only one factor and allowed Hand Movements to indicate G*f* and G*sm*, or "cross-load." This model fit the data well: χ^2 (91) = 180.81, RMSEA = .04, CFI = .98. Moreover, the improvement in model fit was statistically significant, as indicated by $\Delta\chi^2$ (1) = 39.36, $p < .01$. When allowed to load on both factors, Hand Movements had a standardized loading of .40 on the G*f* factor and .30 on the G*sm* factor.

There are a few salient points related to this finding. First, Hand Movements is a supplemental test and not part of the standard battery. Perhaps the authors were not confident enough that this indicator reflected G*sm*, and it was not included in the standard battery for this reason. Therefore, the finding does not invalidate the measurement of G*sm* using the broad index. Second, the finding provides some initial evidence that Hand Movements may measure more than G*sm*, or that it is factorially complex. Third, when such *post hoc* modifications are made, there are always increased risks for sample-specific findings that may or may not be important.

Resolving what latent cognitive abilities Hand Movements measures will be left up to future research. One excellent method that could be used to investigate further what it measures is cross-battery factor analysis (CB-FA); that is, a factor analysis of Hand Movements with G*sm* and G*f* tests from other intelligence batteries. Cross-battery factor analyses (and cross-battery confirmatory factor analysis, CB-CFA, in particular) across measurement instruments is an extremely useful method used to understand what tests measure (Keith & Reynolds, 2010, 2012). On a related note, it is not uncommon to see researchers factor analyze only the standard tests in a battery, even when both standard and supplemental tests are available; or, alternatively, to conduct two analyses, one including all tests and one including only those from the standard battery. We generally discourage this approach. Except under rare circumstances (e.g., a poorly designed or theoretically murky test), more measures will generally lead to a deeper understanding of the underlying constructs. In the present example, Hand Movements may be less desirable because it is factorially complex, but its inclusion, and its theoretically predictable cross-loading, also supports the validity of the underlying constructs. That is, Hand Movements can be understood as requiring novel reasoning as well as short-term memory. The fact that it shows substantial cross-loadings on two such factors supports those factors as indeed representing G*f* and G*sm*, respectively. The alternative, analyzing fewer measures, has the potential to mislead; when fewer tests are analyzed, factors are more likely to represent narrower abilities, and are more likely not to appear at all. When understanding the constructs underlying the test is the purpose, more is almost always better.

HIGHER-ORDER MODELS

Specification

In addition to the five CHC broad abilities, the KABC-II provides an index of a general mental ability, the FCI index. The next step is to match the analytic model (technically now a structural model because the covariance among the first-order factors were structured) with the overall scoring structure of the test. We consider the higher-order model, such as the one shown in Figure 3.4, the most appropriate. Typically, general intelligence (*g*) is considered to influence performance on all measures of cognitive ability. The nature of *g* cannot be understood by surface characteristics of the items or tests designed to measure it, however, as tests that look completely different on the surface often have similar loadings on a *g* factor (Jensen, 1998). Instead, *g* is conceptualized at a higher level of abstraction than the broad abilities, which *are* typically defined by surface characteristics of the measurement instruments. Therefore, we believe that the higher-order model most accurately

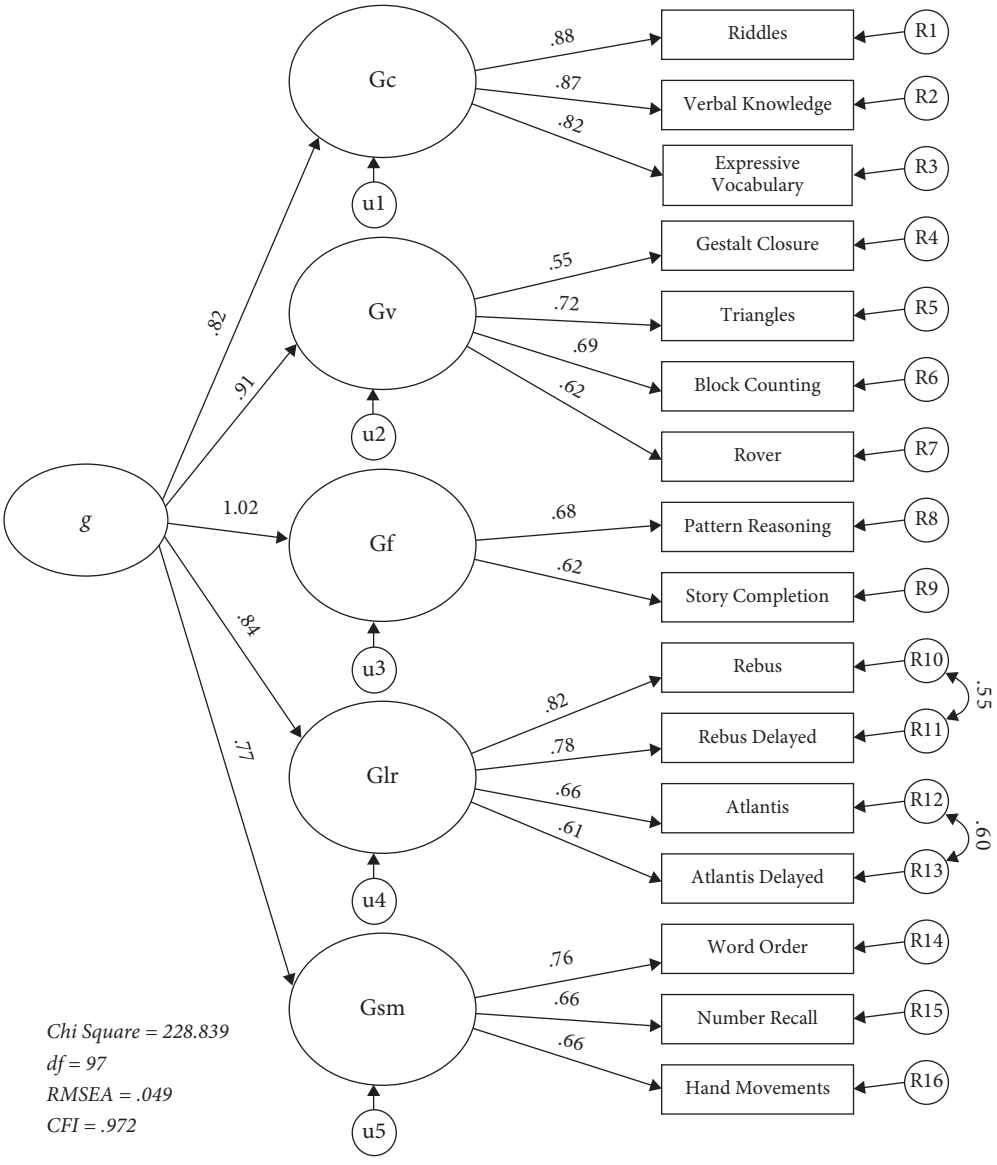


Figure 3.4 KABC-II Second-Order Factor Structure with Standardized Loadings.

mirrors current conceptions of human cognitive abilities (see Carroll, 1993; Jensen, 1998).

There are a few interesting things to note about the higher-order model shown in Figure 3.4. First, the second-order factor, *g*, in part, accounts for the covariance among the first-order factors. This conceptualization provides a more restricted and parsimonious account of the data than does the first-order model. In the higher-order model, there are five loadings on the *g* factor, while in the first-order model there were 10 correlations among the factors. Second, the *g* factor is indicated by the five latent variables and not the observed variables. It is a latent variable indicated by latent variables; *g* is at a higher order of abstraction. In addition, *g* is considered to be more general than the broad abilities because it influences performance on all tests, albeit indirectly through those broad abilities. There are no "direct" effects of *g* on the subtests. The effect of *g* is completely mediated by the broad ability factors. Direct effects may be included,[5] but the representation in Figure 3.4 is both parsimonious and theoretically consistent with contemporary theory (Carroll, 1993; Jensen, 1998). Last, in Figure 3.4, notice the ovals, labeled with "u's," with arrows directed at the first-order factors. These uniquenesses, or disturbances, represent the variance left unexplained by *g*. These disturbances are interesting because they represent the unique aspects of the

broad abilities. That is, they represent unique variance only, not unique (or specific) and error variance, as do the residuals for the subtests. They reference the first-order factors, and the factors are perfectly reliable (cleansed of error), unlike the subtests.

Results

The higher-order model fit well: χ^2 (97) = 228.83, RMSEA = .05, CFI = .97. The fit of the model along with relevant theory would indicate that the higher-order model was a plausible model for these data. However, there was also an oddity in the standardized factor loadings: The factor loading of G*f* on *g* was 1.02 (Figure 3.4). How can a standardized loading be greater than one? It is possible, like in regression (Jöreskog, 1999), although such a result is almost always worth investigating further. Although not shown in Figure 3.4, in addition to the loading of 1.02, the unique variance (u3) for G*f* was not statistically significantly different from zero. These two pieces of information suggest that G*f* and *g* may not be statistically distinguishable, or that they are correlated perfectly. Interestingly, some have posited that G*f* and *g* are identical (Gustafsson, 1984). An identical G*f* and *g* is a theoretical question, however, because perfectly correlated variables need not be identical constructs. Nonetheless, by fixing the G*f* unique variance (u3) to zero, rerunning the model, and then evaluating whether the model fit worse based on $\Delta\chi^2$, the *statistical* equivalence of the two variables in this sample could be tested. We ran such an analysis. The model with the u3 fixed to zero fit the data well: χ^2 (98) = 229.27, RMSEA = .05, CFI = .97. The $\Delta\chi^2$ (1) was 0.44 (p = .51), and was not statistically significant, suggesting that G*f* and *g* are *statistically* equivalent, a finding not uncommon for higher-order analyses of intelligence data (Keith & Reynolds, 2012).

Higher-Order Models and the Schmid-Leiman Transformation

As already noted, in the higher-order model shown in Figure 3.4, *g* only affected the subtests indirectly, via the first-order factors. Said differently, the broad abilities completely mediate the effect of *g* on the subtests. Thus it is possible to calculate the total effect of *g* on each of the subtests in order to get some sense of the loading of each subtest (indirectly) on *g*. It would also be possible to compare this loading on *g* to the subtest's loadings on the broad abilities to get some sense of the relative effect of *g* versus the broad ability. The factor loadings of each subtest on its broad ability and on *g* are shown in Table 3.2. This may feel like cheating in some sense of the word, however, because for both loadings the effect of the broad abilities on the subtests is used. So, for example, the loading of Word Order on G*sm* is .76, whereas the loading of Word Order on *g* is .76 × .77 = .59. Given that all of *g* loadings go through the first-order factors, these loadings are constrained by the loading of the first-order factor on *g*.

Table 3.2 KABC-II Loadings on the First-Order Factors (see Figure 3.4 for the First-Order Factor Names) and the Second-Order *g* Factor. The final column shows the residualized first-order factor loadings, with the effect of *g* removed.

Subtest	First-Order	*g*	Residualized First-Order
Riddles	.885	.728	.503
Verbal Knowledge	.868	.714	.494
Expressive Vocabulary	.824	.677	.469
Gestalt Closure	.555	.504	.232
Triangles	.724	.658	.303
Block Counting	.685	.622	.288
Rover	.617	.560	.258
Pattern Reasoning	.693	.693	.000
Story Completion	.627	.627	.000
Rebus	.817	.688	.440
Rebus Delayed	.783	.660	.421
Atlantis	.657	.554	.352
Atlantis Delayed	.610	.514	.328
Word Order	.764	.590	.486
Number Recall	.656	.507	.417
Hand Movements	.662	.511	.421

If the double use of the first-order loadings makes you feel uncomfortable, an alternative would be to ask: What is the residual effect of the broad abilities after *g* is taken into account? One way to calculate these residual effects is to square the *g* loadings (to obtain the variance accounted for by *g*) and subtract these from the R^2 for each subtest (a statistic available in any SEM program). The resulting value would represent the variance explained uniquely by the broad abilities, after accounting for the variance explained by *g*. The square root of that

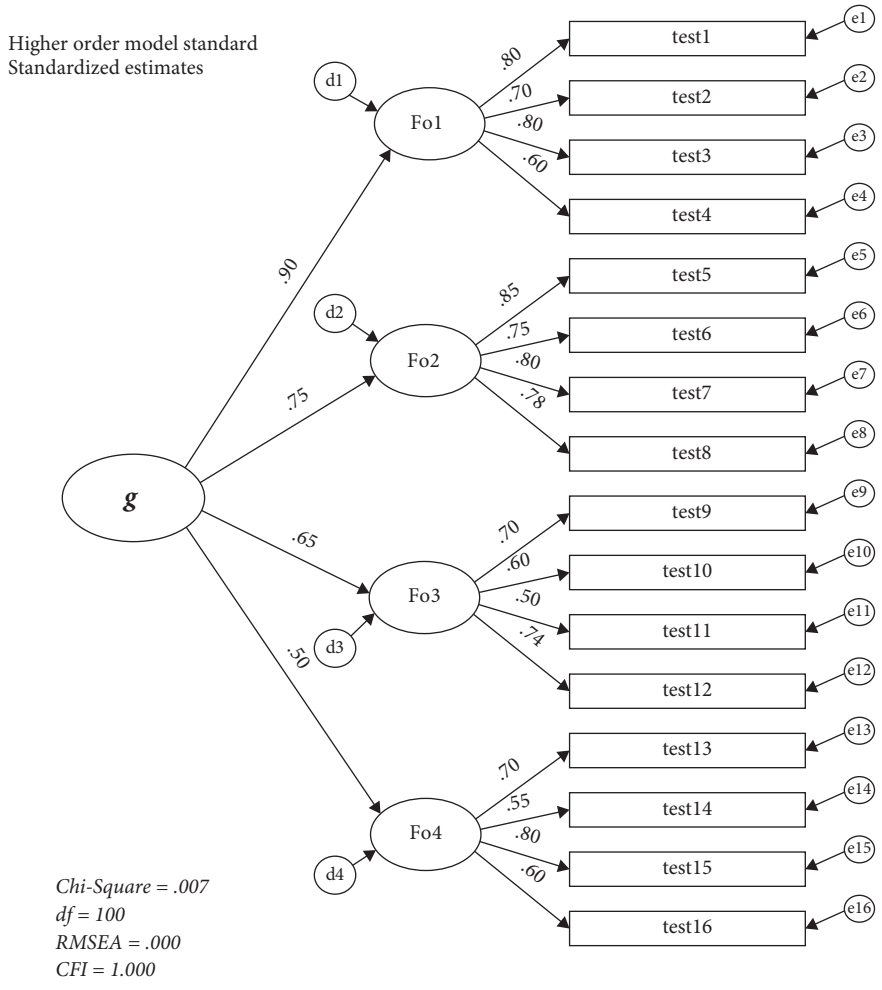


Figure 3.5 Higher-Order Model Results for the Simulated Data. The model fits nearly perfectly because the model shown was used to generate the data.

unique variance would then represent the unique loading of each subtest on the broad abilities after *g* is taken into account. These values are also shown in Table 3.2 in the last column on the right.

To further illustrate these and subsequent points, we will switch to simulated data. Figure 3.5 shows a straightforward factor model of 16 tests measuring four first-order factors and *g*, a higher-order factor. The model and the data are designed to be consistent with findings from analyses of intelligence test data. (The model fits these data perfectly, or nearly perfectly, because the model shown was used to simulate a matrix, which was then used in the analysis.) The first factor (Fo1) is most similar to *g*, and there is variability among the tests in how well they measure each broad ability. Figure 3.5 shows the loadings of each test on the broad abilities, and the first column of numbers in Table 3.3 shows the loadings of each test on *g*. The second column of numbers in Table 3.3 shows the residualized loading of each test on the corresponding broad ability, or the unique effect of each broad ability on their subtest indicators, after accounting for *g*.

Discussion of the unique effect of the broad abilities suggests another way to calculate these effects. Figure 3.6 shows a slight variation of the higher-order model. In the initial figure, the disturbances of the first-order factors were scaled by constraining the path from the disturbance to the factor to 1.0 (what Kline, 2011, calls "unit loading identification" [ULI]). In Figure 3.6, an alternative method was used to scale the disturbances: The variances of the disturbances were set to 1, and the paths from the disturbances to the factors were estimated (unit variance identification, or UVI). With

Table 3.3 Loadings on the Higher-Order *g* Factor Versus the Residualized First-Order Factor Loadings, Calculated with Two Methods, for the Simulated Data

	g	First Order $\sqrt{R^2 - g^2}$	First Order *ul* × *fol*
Test 1	.720	.349	.349
Test 2	.630	.305	.305
Test 3	.720	.349	.349
Test 4	.540	.262	.262
Test 5	.638	.562	.562
Test 6	.563	.496	.496
Test 7	.600	.529	.529
Test 8	.585	.516	.516
Test 9	.455	.532	.532
Test 10	.390	.456	.456
Test 11	.325	.380	.380
Test 12	.481	.562	.562
Test 13	.350	.606	.606
Test 14	.275	.476	.476
Test 15	.400	.693	.693
Test 16	.300	.520	.520

Note: First-order loadings represent the first-order factor effect on the test, with effects of *g* removed and calculated with two different methods (see text for explanation).

this specification, it is possible to calculate the indirect effects from the disturbances (the *unique variances* of the first-order factors) to the tests. So, for example, the effect of d4 on test 16 is .866 x .600 = .5196, or .520. Again, this is the unique effect of the first-order factor, with the effect of *g* removed. These values are shown in the final column of Table 3.3; they are the same as those shown in the previous column.

This calculation of the unique, or residualized, effects of the first-order factors, after accounting for the second-order factor, is analogous to the common Schmid-Leiman procedure in exploratory factor analysis. Table 3.4 shows the Schmid-Leiman transformation for these same data. The solution is based on an exploratory principal factors analysis of the simulated data used for Figures 3.5 and 3.6 (and Table 3.4), with extraction and promax rotation of four factors. As can be seen by comparing Table 3.4 with Table 3.3, the estimates from this exploratory analysis are quite close to those from the confirmatory analysis. (It should be noted that the ordering of factors was changed; that is, what is labeled as "Factor 1" in the table actually came out as "Factor 2" in the EFA). Again, the table comparisons show that residualizing the first-order factor loadings from a higher-order model is methodologically equivalent to a Schmid-Leiman transformation.

Several points are worth mentioning about these procedures. First, they go by a variety of names. Here we have referred to this as a *residualization of the first-order factor loadings*, accounting for the second. Others may refer to this as an *orthogonalization* (e.g., Watkins, Wilson, Kotz, Carbone, & Babula, 2006) because the first-order factors have been made *orthogonal* (uncorrelated with) the second-order factor. This concept is well illustrated in Figure 3.6, where the unique factors are uncorrelated with the second-order *g* factor. It would also be correct to refer to these loadings as the *g loadings* and the unique effects of the first-order factors. Some writers may simply refer to these as *first* and *second-order factor loadings*, apparently not recognizing that the first-order loadings are with *g* statistically controlled.

Effects versus Proportion of Explained Variance

Readers may wonder why we focused on factor loadings rather than variances. After all, one method used to calculate the factor loadings did so by converting the *g* loadings to variances. We believe the focus on factor loadings rather than variances is appropriate for several reasons. First, factor loadings are the original metric. They are readily interpretable as effects; that is, the effect of *g*, or the broad abilities, on the tests. Second, the focus on factor loadings also makes this procedure easily interpretable as a Schmid-Leiman transformation. Finally, because variances focus on the original metric squared, they provide misleading estimates of the relative importance of the factors (Darlington, 1990; Keith, 2006, Chap. 5).

Total versus Unique Effects

A final point concerning this residualization is the reminder that the tables show the loadings of *g*, and the loadings of the first-order factors *with g controlled*, or the total effect of *g* and the unique

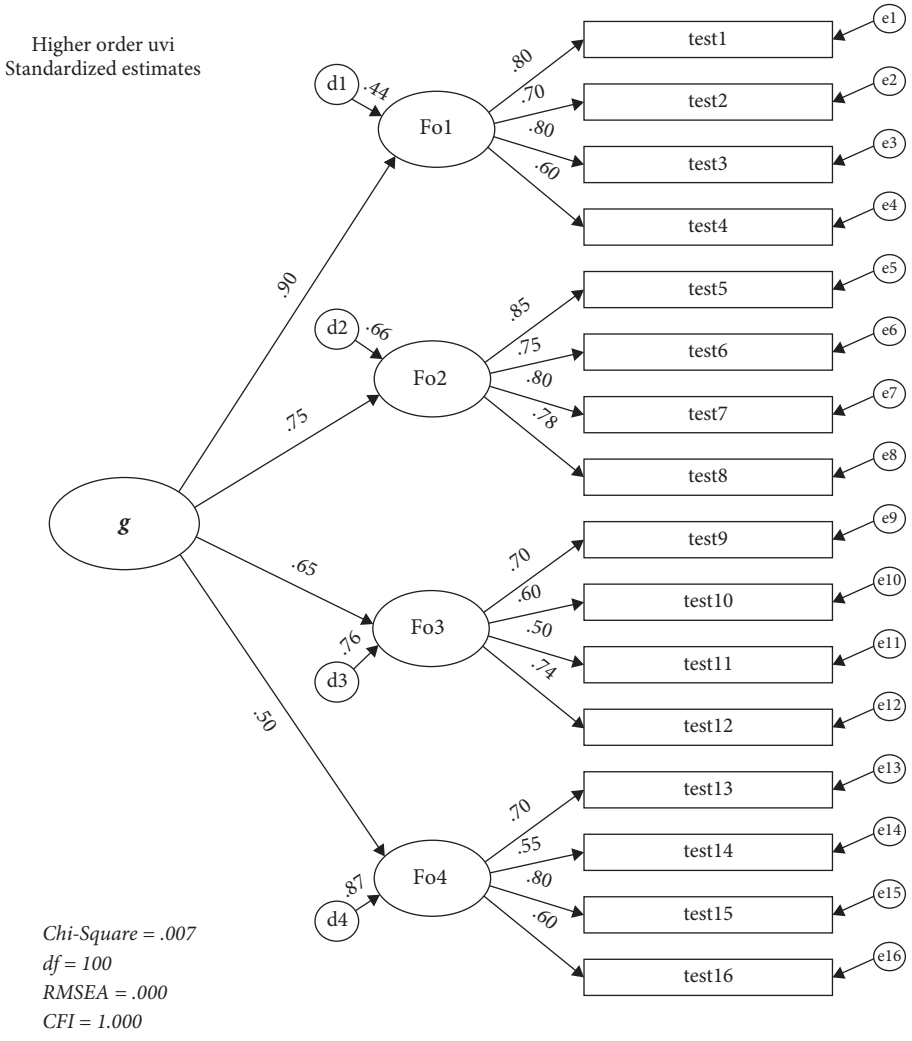


Figure 3.6 Higher-Order Model Estimated Using Unit Variance Identification. Note the standardized paths from the disturbances to the first-order factors.

effect of the first-order factors. As such, the technique gives interpretive predominance to *g,* essentially a tacit, Spearman-like notion that *g* is most important. There is nothing wrong with this interpretation as long as researchers and readers understand it. Readers who believe that first-order factors should be given interpretive predominance (a Thurstone-like notion) could reasonably argue for the opposite of this procedure: the interpretation of the first-order factor loadings versus the unique effect of *g, while controlling for the first-order factors.* Because the strict higher-order model has *g* affecting the subtests only through the broad abilities, for this approach the first-order factor loadings (i.e., the "First-Order" in Table 3.2) represent the effect of the broad abilities on the subtests, but the effect of *g* on the subtests would all be equal to zero (because there are no direct effects of *g* on the broad abilities)! Thus we recommend reporting results of this transformation, but also reporting the original, un-residualized, first-order factor loadings.

An Alternative to the Higher-Order Hierarchical Model

The higher-order model is the most common method of estimating both broad (e.g., G*f*, G*c*) and general (*g*) abilities in the same model. Another type of hierarchical model is often referred to as the *nested-factors* or *bi-factor model*. In this type of hierarchical model, the general and broad factors are at the same level; an example using the simulated model is shown in Figure 3.7. Note that methodologists have used different names to refer to such models. The higher-order model is sometimes called