



GARY GROTH-MARNAT
A. JORDAN WRIGHT

HANDBOOK OF
**PSYCHOLOGICAL
ASSESSMENT**

SIXTH EDITION

WILEY

HANDBOOK OF PSYCHOLOGICAL ASSESSMENT

HANDBOOK OF PSYCHOLOGICAL ASSESSMENT

Sixth Edition

Gary Groth-Marnat and A. Jordan Wright

WILEY

Copyright © 2016 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, or on the Web at www.copyright.com. Requests to the publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, 201-748-6011, fax 201-748-6008, or online at www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Readers should be aware that Internet Web sites offered as citations and/or sources for further information may have changed or disappeared between the time this was written and when it is read.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in rendering professional services. If legal, accounting, medical, psychological or any other expert assistance is required, the services of a competent professional should be sought.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 800-956-7739, outside the U.S. at 317-572-3986, or fax 317-572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Names: Groth-Marnat, Gary, author. | Wright, A. Jordan, author.

Title: Handbook of psychological assessment / Gary Groth-Marnat and A. Jordan Wright.

Description: Sixth edition. | Hoboken, New Jersey : John Wiley & Sons, Inc., [2016] | Includes index.

Identifiers: LCCN 2015042889 (print) | LCCN 2016000499 (ebook) | ISBN 9781118960646 (cloth) | ISBN 9781118960684 (pdf) | ISBN 9781118960653 (epub)

Subjects: LCSH: Psychological tests. | Personality assessment.

Classification: LCC BF176 .G76 2016 (print) | LCC BF176 (ebook) | DDC 150.28/7—dc23

LC record available at <http://lcn.loc.gov/2015042889>

Cover Image: © 501room/Shutterstock

Cover Design: Wiley

Printed in the United States of America

SIXTH EDITION

HB Printing 10 9 8 7 6 5 4 3 2 1

Contents

Preface	xi
Chapter 1 Introduction	1
Organization of the Handbook	1
Role of the Clinician	3
Patterns of Test Usage in Clinical Assessment	6
Evaluating Psychological Tests	10
Validity in Clinical Practice	23
Clinical Judgment	26
Phases in Clinical Assessment	32
Recommended Reading	37
Chapter 2 Context of Clinical Assessment	39
Types of Referral Settings	39
Ethical Practice of Assessment	50
Assessing Diverse Groups	59
Selecting Psychological Tests	67
Computer-Assisted Assessment	72
Recommended Reading	74
Chapter 3 The Assessment Interview	77
History and Development	78
Issues Related to Reliability and Validity	82
Assets and Limitations	84
The Assessment Interview and Case History	86
Mental Status Evaluation	93
Interpreting Interview Data	99
Structured Interviews	100
Recommended Reading	111

Chapter 4	Behavioral Assessment	113
	History and Development	115
	Issues Related to Reliability and Validity	117
	Assets and Limitations	120
	Strategies of Behavioral Assessment	121
	Recommended Reading	136
Chapter 5	Wechsler Intelligence Scales	139
	Testing of Intelligence: Pros and Cons	139
	History and Development	142
	Reliability and Validity	147
	Assets and Limitations	151
	Use with Diverse Groups	153
	Meaning of IQ Scores	156
	Cautions and Guidelines in Administration	159
	WAIS-IV/WISC-V Successive-Level Interpretation Procedure	161
	Wechsler Indexes and Subtests	173
	Assessing Brain Damage	196
	Assessing Additional Special Populations	204
	Short Forms	209
	Recommended Reading	213
Chapter 6	Wechsler Memory Scales	215
	History and Development	216
	Reliability and Validity	222
	Assets and Limitations	224
	Use with Diverse Groups	226
	Interpretation Procedure	227
	Interpreting Patterns of Index Scores	230
	Comparing Scores on the WAIS-IV and the WMS-IV	238
	Additional Considerations: Malingering and Evaluating Change	241
	Recommended Reading	242
Chapter 7	Minnesota Multiphasic Personality Inventory	243
	History and Development	246
	Reliability and Validity	251
	Assets and Limitations	255
	Use with Diverse Groups	259

Administration	262
MMPI-2 Interpretation Procedure	263
MMPI-2 Computerized Interpretation	272
MMPI-2 Validity Scales	273
MMPI-2 Clinical Scales	279
MMPI-2 2-Point Codes	302
MMPI-2 Content Scales	328
MMPI-A Content Scales	331
MMPI-2 Harris-Lingoes and <i>Si</i> Subscales	334
MMPI-2 Critical Items	337
MMPI-2 and MMPI-A Supplementary Scales	338
MMPI-2-RF Interpretation Procedure	341
MMPI-2-RF Validity Scales	346
MMPI-2-RF Higher-Order Scales	353
MMPI-2-RF Restructured Clinical Scales	355
MMPI-2-RF Specific Problem Scales	361
Interest Scales	366
MMPI-2-RF Personality Psychopathology Five Scales	366
Recommended Reading	370
Chapter 8 Personality Assessment Inventory	371
History and Development	373
Reliability and Validity	373
Assets and Limitations	375
Use with Diverse Groups	377
Interpretation Procedure	377
Validity Scales	380
Clinical Scales	384
Treatment Scales	404
Interpersonal Scales	408
Additional Clusters	410
Critical Items	415
Recommended Reading	416
Chapter 9 Millon Clinical Multiaxial Inventory	417
History and Development	419
Reliability and Validity	424
Assets and Limitations	427

Use with Diverse Groups	431
Interpretation Procedure	432
Validity Scales	436
Clinical Personality Patterns	439
Severe Personality Pathology	476
Clinical Syndromes	486
Severe Syndromes	488
Recommended Reading	488
Chapter 10 NEO Personality Inventory	489
History and Development	490
Reliability and Validity	491
Assets and Limitations	493
Use with Diverse Groups	494
Interpretation Procedure	495
Recommended Reading	512
Chapter 11 The Rorschach	513
History and Development	514
Reliability and Validity	518
Assets and Limitations	523
Use with Diverse Groups	527
Comprehensive System: Administration	528
Comprehensive System: Coding	531
Comprehensive System: Scoring the Structural Summary	536
Comprehensive System: Interpretation	539
R-PAS: Administration	581
R-PAS: Coding	584
R-PAS: Scoring the Structural Summary	592
R-PAS: Interpretation	593
Recommended Reading	614
Chapter 12 Screening for Neuropsychological Impairment	615
History and Development	617
Interviewing for Brain Impairment	620
Domains of Neuropsychological Functioning	624
Bender Visual-Motor Gestalt Test, Second Edition	633

Repeatability Battery for the Assessment of Neuropsychological Status Update	645
Recommended Reading	652
Chapter 13 Brief Instruments for Treatment Planning, Monitoring, and Outcome Assessment	653
Selecting Brief Instruments	654
Symptom Checklist-90-R and Brief Symptom Inventory	655
Beck Depression Inventory–II	662
State Trait Anxiety Inventory	666
Recommended Reading	670
Chapter 14 Treatment Planning and Clinical Decision Making	671
Development and Approaches to Treatment Planning	672
Intervention Options	677
Clinical Decision Making	679
Case Formulation	680
Understanding the Problems	683
Understanding Problem Context	692
Treatment-Specific Client Characteristics	696
The Systematic Treatment Selection (STS)/Innerlife Approach	703
Recommended Reading	705
Chapter 15 The Psychological Report	707
General Guidelines	708
Feedback	728
Format for a Psychological Report	730
Sample Reports	743
Recommended Reading	772
Appendix A Test Publishers/Distributors	773
Appendix B Testing Organizations	777
References	779
Author Index	857
Subject Index	895

Preface

Thank you so much for your support in buying and reading this book. Our intention has been to create a resource that will cover the A to Z of assessment. In other words, our aim has been to provide guidance that includes larger issues on assessment as well as specific stages in the assessment process, from clarifying the referral question through writing up the report and providing feedback and consulting with your referral sources and clients. We hope it brings clarity, practical guidelines, insights, and useful strategies to your work. Feedback on the previous editions assures us that this is often the case. This fact makes it worth the many long hours hidden away inside a small room incubating ideas and reading, writing, revising, and editing.

As with the previous editions, we have tried to integrate the best of science with the best of practice. Necessarily, psychological assessment involves technical knowledge. But in presenting this technical knowledge, we have tried to isolate, extract, and summarize in as clear a manner as possible the core information that is required for practitioners to function competently. At the same time, assessment is also about the very human side of understanding, helping, and making decisions about people. We hope we have been able to comfortably blend this technical (science) side with the human. An assessment that does not have at least some heart to it is cold and falls short in understanding the experience of the client. To keep in touch with the practitioner/human side of assessment, we have continually maintained active assessment practices in which we have tried to stay close to and interact with the ongoing personal and professional challenges of practitioners. We hope that within and between the sentences in the book, our active involvement with the world of practice is apparent.

It has been seven years since the previous (fifth) edition was published. During that time, much has changed but much has remained the same. The big tests that professional psychologists use most frequently are somewhat different, and this is reflected in changes to this sixth edition. This includes eliminating the chapter on the California Psychological Inventory and replacing it with the more widely used NEO Personality Inventory—3. While both focus on normal personality traits, the NEO is based on the strongly empirically supported five-factor model of personality. Additionally, this edition has eliminated the chapter on the Thematic Apperception Test. Although its use in clinical practice is unclear (anecdotally it seems still to be relatively widely used), the test itself has suffered from the lack of consensus on a coding and scoring protocol and a subsequent lack of strong, consistent empirical support. In place of this chapter, we have included a chapter on the Personality Assessment Inventory, which has gained both strong empirical support and wide clinical popularity.

In addition to these major changes in tests covered, there are important changes within other chapters. The chapter on the Wechsler Intelligence Scales includes

updated information on the newly developed Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V). Additionally, two chapters have been significantly expanded because each test has two alternate forms, both of which currently are in wide use. Specifically, the chapter on the Minnesota Multiphasic Personality Inventory includes information on both the MMPI-2 and the MMPI-2–Restructured Form (MMPI-2-RF), and the chapter on the Rorschach includes information on both the Comprehensive System and the Rorschach Performance Assessment System (R-PAS). For both of these tests, the next 10 or so years may see the field lean toward one or the other form of the tests, but at the moment the field is split, and both versions of both tests are widely used. Finally, we have worked to strengthen the sections on “Use with Diverse Groups,” which reflect the more extensive use of assessment for a wide variety of populations and the importance of competently and sensitively working with diverse populations.

There are also many smaller changes throughout this sixth edition. It has been fully updated with new research in the field. There has also been greater emphasis on making assessment more user friendly and consumer oriented. This is reflected in suggestions for using everyday language in reports, connecting interpretations to actual client behavior, strategies for wording interpretations in a manner likely to enhance client growth, and the importance of collaborating with clients. The treatment planning and clinical decision making chapter has been completely updated, and the psychological report writing chapter has been updated to include current thinking of the American Psychological Association and the Society for Personality Assessment about proficiency in personality assessment. We hope that these changes will provide readers with the best, most current, and most practical of what can be available in assessment.

The development of the *Handbook of Psychological Assessment* has been a group effort. It started many years ago with ideas and cowriting with Gary Groth-Marnat’s colleague Dorothy Morena. We wanted to develop a resource that would assist students with all phases of psychological assessment. Our sincere thanks to her. A series of editors at John Wiley & Sons have been invaluable, including Herb Reich, Jennifer Simon, Tracey Belmont, Lisa Gebo, Peggy Alexander, and Marquita Flemming. We have very much enjoyed and appreciated our relationship with Wiley; not only have we been treated as respected authors, but they have also welcomed us into the Wiley “family.” Colleagues who have provided valuable input include Steve Smith, Larry Beutler, Steve Finn, Alan Kaufman, Dawn Flanagan, Greg Meyer, Joni Mihura, Aasha Foster, and the invaluable and nonstop list of articles from the Kenneth Pope website and listerv. Seth Grossman, C. J. Thompson, and their colleagues at Pearson Assessment were extremely helpful and generous in supplying us with advance information on the MCMI-IV. Finally, much of our professional work is devoted toward helping students to achieve the best of what they are capable of. In return, working with them has inevitably helped us refine this sixth edition. Finally, we would like to dedicate the sixth edition to Gary’s parents, Barbara and Rudy, in memoriam, as well as to Jordan’s husband, Matt, and daughter, Millie, for their unwavering support.

Gary Groth-Marnat and A. Jordan Wright
July 28, 2015

HANDBOOK OF PSYCHOLOGICAL ASSESSMENT

INTRODUCTION

The *Handbook of Psychological Assessment* is designed to develop a high level of practitioner competence by providing relevant, practical research, and theoretical information. It can serve as both a reference and an instructional guide. As a reference book, it aids in test selection and the development of a large number and variety of interpretive hypotheses. As an instructional text, it provides students with the basic tools for conducting an integrated psychological assessment. The significant and overriding emphasis in this book is on assessing areas that are of practical use in evaluating individuals in a clinical context. It is applied in its orientation, and, for the most part, theoretical discussion has been kept to a minimum. Many books written on psychological testing and the courses organized around these books focus primarily on test theory, with a brief overview of a large number of tests. In contrast, the intent of this book is to focus on the actual processes that practitioners go through during assessment. We begin with such issues as role clarification and evaluation of the referral question and end with treatment planning and the actual preparation of the report itself.

One of the crucial skills that we hope readers of this text will develop, or at least have enhanced, is a realistic appreciation of the assets and limitations of assessment. This includes an appraisal of psychological assessment as a general strategy as well as an awareness of the assets and limitations of specific instruments and procedures. A primary limitation of assessment lies in the incorrect handling of the data, which are not integrated in the context of other sources of information (behavioral observations, history, other test scores). Also, the results are not presented in a way that helps solve the unique problems clients or referral sources are confronting. To counter these limitations, the text continually provides practitioners with guidelines for integrating and presenting the data in as useful a manner as possible. The text is thus not only a book on test interpretation (although this is an important component) but on test integration within the wider context of assessment. As a result, psychologists should be able to create reports that are accurate, effective, concise, and highly valued by the persons who receive them.

ORGANIZATION OF THE HANDBOOK

The central organizational plan for the *Handbook of Psychological Assessment* replicates the sequence practitioners follow when performing an evaluation. They are initially concerned with clarifying their roles, ensuring that they understand all the implications of the referral question, deciding which procedures would be most appropriate for the assessment, and reminding themselves of the potential problems

associated with clinical judgment (this chapter). They also need to understand the context in which they will conduct the assessment. This understanding includes appreciating the issues, concerns, terminology, and likely roles of the persons from these contexts. Practitioners also must follow clear ethical guidelines, know how to work with persons from diverse backgrounds, and recognize issues related to computer-assisted assessment and the ways that the preceding factors might influence their selection of procedures (see Chapter 2).

Once practitioners have fully understood the preliminary issues discussed in this chapter and Chapter 2, they must select different strategies of assessment. The three major strategies are interviewing, observing behavior, and psychological testing. An interview is likely to occur during the initial phases of assessment and is also essential in interpreting test scores and understanding behavioral observations (see Chapter 3). The assessment of actual behaviors might also be undertaken (see Chapter 4). Behavioral assessment might be either an end in itself or an adjunct to testing. It might involve a variety of strategies, such as the measurement of overt behaviors, cognitions, alterations in physiology, or relevant measures from self-report inventories.

The middle part of the book (Chapters 5 through 13) provides a general overview of the most frequently used tests. Each chapter begins with an introduction to the test in the form of a discussion of its history and development, current evaluation, and procedures for administration, as well as use with diverse populations. The main portions of these chapters provide a guide for interpretation, which includes such areas as the meaning of different scales, significant relations between scales, frequent trends, and the meaning of unusually high or low scores. When appropriate, there are additional subsections. For example, Chapter 5, “Wechsler Intelligence Scales,” includes additional sections on the meaning of IQ scores, estimating premorbid IQ, and assessing special populations. Likewise, several chapters include alternative procedures for using the tests, such as Chapter 7, “Minnesota Multiphasic Personality Inventory,” which includes procedures for both the MMPI-2 and the MMPI-2-RF, and Chapter 11, “The Rorschach,” which includes both the Comprehensive System and the R-PAS versions of the Rorschach. Chapter 12, “Screening for Neuropsychological Impairment,” varies somewhat from the preceding format in that it is more a compendium and interpretive guide to some of the most frequently used short neuropsychological tests. It also includes a section on special considerations in conducting a neuropsychological interview. This organization reflects the current emphasis on and strategies for assessing patients with possible neuropsychological dysfunction.

Several of the chapters on psychological tests are quite long, particularly those for the Wechsler intelligence scales, the Minnesota Multiphasic Personality Inventory, and the Rorschach. These chapters include extensive summaries of a wide variety of interpretive hypotheses intended for reference purposes when practitioners must generate interpretive hypotheses based on specific test scores. To gain initial familiarity with the tests, we recommend that practitioners or students carefully read the initial sections (history and development, psychometric properties, etc.) and then skim through the interpretation sections more quickly. Doing this provides the reader with a basic familiarity with the procedures and types of data obtainable from the tests. As practical test work progresses, clinicians can then study the interpretive hypotheses in greater depth and gradually develop more extensive knowledge of the scales and their interpretation.

Based primarily on current frequency of use, these tests are covered in this text: Wechsler intelligence scales (WAIS-IV/WISC-V), Wechsler Memory Scales (WMS-IV), Minnesota Multiphasic Personality Inventory (MMPI-2 and MMPI-2-RF), Millon Clinical Multiaxial Inventory (MCMI-IV), Personality Assessment Inventory (PAI), NEO Personality Inventory–3 (NEO-PI-3), Bender Visual Motor Gestalt Test–II, Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), and the Rorschach (Comprehensive System and R-PAS; Camara, Nathan, & Puente, 2000; C. Piotrowski & Zalewski, 1993; Rabin, Barr, & Burton, 2005; Watkins, 1991; Watkins, Campbell, Nieberding, & Hallmark, 1995). The NEO-PI-3 was selected because of the importance of including a broad-based inventory of normal functioning, along with its excellent technical development and relatively large research base. We have also included Chapter 13 focusing on the most frequently used brief, symptom-focused inventories because of the increasing importance of monitoring treatment progress and outcome in a cost- and time-efficient managed care environment (Eisman et al., 2000; C. Piotrowski, 1999). The preceding instruments represent the core assessment devices used by most practitioners.

Finally, the clinician must generate relevant treatment recommendations and integrate the assessment results into a psychological report. Chapter 14 provides a systematic approach for working with assessment results to develop practical, evidence-based treatment recommendations. Chapter 15 presents guidelines for report writing, a report format, and four sample reports representative of the four most common types of referral settings: medical setting, legal context, educational context, and psychological clinic. Thus, the chapters follow a logical sequence and provide useful, concise, and practical knowledge.

ROLE OF THE CLINICIAN

The central role of clinicians conducting assessments should be to answer specific questions and make clear, specific, and reasonable recommendations to help improve functioning. To fulfill this role, clinicians must integrate a wide range of data and bring into focus diverse areas of knowledge. Thus, they are not merely administering and scoring tests. A useful distinction to highlight this point is the contrast between a psychometrist and a clinician conducting psychological assessment (Maloney & Ward, 1976; Matarazzo, 1990). Psychometrists tend to use tests merely to obtain data, and their task is often perceived as emphasizing the clerical and technical aspects of testing. Their approach is primarily data oriented, and the end product is often a series of traits or ability descriptions. These descriptions are typically unrelated to the person's overall context and do not address unique problems the person may be facing. In contrast, psychological assessment attempts to evaluate an individual in a problem situation so that the information derived from the assessment can somehow help with the problem. Tests are only one method of gathering data, and the test scores are not end products but merely means of generating hypotheses. Psychological assessment, then, places data in a wide perspective, with its focus being problem solving and decision making.

The distinction between psychometric testing and psychological assessment can be better understood and the ideal role of the clinician more clearly defined by briefly

elaborating on the historical and methodological reasons for the development of the psychometric approach. When psychological tests were originally developed, group measurements of intelligence met with early and noteworthy success, especially in military and industrial settings where individual interviewing and case histories were too expensive and time consuming. An advantage of the data-oriented intelligence tests was that they appeared to be objective, which would reduce possible interviewer bias. More important, they were quite successful in producing a relatively high number of true positives when used for classification purposes. Their predictions were generally accurate and usable. However, these facts created the early expectation that all assessments could be performed using the same method and would provide a similar level of accuracy and usefulness. Later assessment strategies often tried to imitate the methods of earlier intelligence tests for variables such as personality and psychiatric diagnosis.

A further development consistent with the psychometric approach was the strategy of using a “test battery.” It was reasoned that if a single test could produce accurate descriptions of an ability or trait, administering a series of tests could create a total picture of the person. The goal, then, was to develop a global yet definitive description for the person using purely objective methods. This goal encouraged the idea that the tool (psychological test) was the best process for achieving the goal, rather than being merely one technique in the overall assessment procedure. Behind this approach were the concepts of *individual differences* and *trait psychology*. These concepts assume that one of the best ways to describe the differences among individuals is to measure their strengths and weaknesses with respect to various traits. Thus, the clearest approach to the study of personality involved developing a relevant taxonomy of traits and then creating tests to measure those traits. Again, there was an emphasis on the tools as primary, with a deemphasis on the input of the clinician. These trends created a bias toward administration and clerical skills. In this context, the psychometrist requires little, if any, clinical expertise other than administering, scoring, and interpreting tests. According to such a view, the most preferred tests would be highly standardized and ideally machine-scored so that the normed scores, rather than the psychometrist, provide the interpretation.

The objective psychometric approach is most appropriately applicable to ability tests such as those measuring intelligence or mechanical skills. Its usefulness decreases, however, when users attempt to assess personality traits such as dependence, authoritarianism, or anxiety. Personality variables are far more complex and, therefore, need to be validated in the context of history, behavioral observations, and interpersonal relationships. For example, a moderately elevated score on a scale measuring high energy level takes on an entirely different meaning for a high-functioning physician than for an individual with a history of mood disorders and associated work and interpersonal difficulties. When the purely objective psychometric approach is used for the evaluation of problems in living (coping more effectively, resolving interpersonal relationships, etc.), its usefulness is questionable. Scores need to be connected to each other and to the context in which they emerge.

Psychological assessment is most useful in the understanding and evaluation of personality and in elucidating the likely underlying causes of problems in living. These issues involve a particular problem situation having to do with a specific individual. The central role of the clinician performing psychological assessment is that of an

expert in human behavior who must deal with complex processes and understand test scores in the context of a person's life. The clinician must have knowledge concerning problem areas and, on the basis of this knowledge, form a general idea regarding behaviors to observe and areas in which to collect relevant data. Doing this involves an awareness and appreciation of multiple causation, interactional influences, and multiple relationships. As Woody (1980) stated, "Clinical assessment is individually oriented, but it always considers social existence; the objective is usually to help the person solve problems."

In addition to an awareness of the role suggested by psychological assessment, clinicians should be familiar with core knowledge related to measurement and clinical practice. This includes descriptive statistics, reliability (and measurement error), validity (and the meaning of test scores), normative interpretation, selection of appropriate tests, administration procedures, variables related to diversity (ethnicity, race, age, gender, culture, etc.), testing individuals with disabilities, and an appropriate amount of supervised experience (Turner, DeMers, Fox, & Reed, 2001). Persons performing psychological assessment should also have basic knowledge related to the demands, types of referral questions, and expectations of various contexts—particularly employment, education, vocational/career, health care (psychological, psychiatric, medical), and forensic. Furthermore, clinicians should know the main interpretive hypotheses in psychological testing and be able to identify, sift through, and evaluate a series of hypotheses to determine which are most relevant and accurate. Rather than merely knowing the labels and definitions for various types of anxiety or thought disorders, for example, clinicians should also have in-depth operational criteria for them. As another example, the concept of intelligence, as represented by the IQ score, can sometimes appear misleadingly straightforward. Intelligence test scores can be complex, though, involving a variety of cognitive abilities, the influence of cultural factors, varying performance under different conditions, and issues related to the nature of intelligence. Unless clinicians are familiar with these areas, they are not adequately prepared to handle IQ data.

The above knowledge should be integrated with relevant general coursework, including abnormal psychology, the psychology of adjustment, theories of personality, clinical neuropsychology, psychotherapy, and basic case management. A problem in many training programs is that, although students frequently have knowledge of abnormal psychology, personality theory, and test construction, they usually have insufficient training to integrate their knowledge into the interpretation of test results. Their training focuses on developing competency in administration and scoring rather than on knowledge relating to what they are testing.

The approach in this book is consistent with that of psychological assessment: Clinicians should be not only knowledgeable about traditional content areas in psychology and the various contexts of assessment but also able to integrate the test data into a relevant description of the person. This description, although focusing on the individual, should take into account the complexity of his or her social environment, personal history, and behavioral observations. Yet the goal is not merely to describe the person but rather to develop relevant answers to specific questions and present clear, specific, and reasonable recommendations that aid in problem solving and facilitate decision making.

PATTERNS OF TEST USAGE IN CLINICAL ASSESSMENT

Psychological assessment is crucial to the definition, training, and practice of professional psychology. Although the data are old, Watkins et al. (1995) found that fully 91% of all practicing psychologists engage in assessment, and 64% of all nonacademic advertisements listed assessment as an important prerequisite (Kinder, 1994). Assessment skills are also strong prerequisites for internships and postdoctoral training. The theory and instruments of assessment can be considered the very foundation of clinical investigation, applied research, and program evaluation. In many ways, psychological assessment is professional psychology's unique contribution to the wider arena of clinical practice. The early professional psychologists even defined themselves largely in the context of their role as psychological testers. Practicing psychologists spend 10% to 25% of their time conducting psychological assessment (Camara et al., 2000; Watkins, 1991; Watkins et al., 1995).

Although assessment has always been a core, defining feature of professional psychology, the patterns of use and relative importance of assessment have changed with time. During the 1940s and 1950s, psychological testing was frequently the single most important activity of professional psychologists. In contrast, the past 60 years have seen psychologists become involved in a far wider diversity of activities. Lubin and his colleagues (Lubin, Larsen, & Matarazzo, 1984; Lubin, Larsen, Matarazzo, & Seever, 1985, 1986) found that the average time spent performing assessment across five treatment settings was 44% in 1959, 29% in 1969, and only 22% in 1982. The average time spent in 1982 performing assessments in the five different settings ranged from 14% in counseling centers to 31% in psychiatric hospitals (Lubin et al., 1984, 1985, 1986). Camara et al. (2000) found that the vast majority of professional psychologists (81%) spend 0 to 4 hours a week conducting formal assessment, 15% spend 5 to 20 hours a week, and 4% spend more than 20 hours. It is expected that over the last 20 years, the time spent doing assessment has likely decreased even further. The gradual decrease in the total time spent in assessment is due in part to the widening role of psychologists. Whereas in the 1940s and 1950s a practicing psychologist was almost synonymous with a tester, professional psychologists currently are increasingly involved in administration, consultation, organizational development, and many areas of direct treatment (Bamgbose, Smith, Jesse, & Groth-Marnat, 1980; Groth-Marnat, 1988; Groth-Marnat & Edkins, 1996). Decline in testing has also been attributed to disillusionment with the testing process based on criticisms about the reliability and validity of many assessment devices (Garb, Wood, Nezworski, Grove, & Stejskal, 2001; Wood, Lilienfeld, Garb, & Nezworski, 2000; Ziskin & Faust, 2008) and reductions in reimbursement (Cashel, 2002). In addition, psychological assessment has come to include a wide variety of activities beyond merely the administration and interpretation of traditional tests. These include conducting structured and unstructured interviews, behavioral observations in natural settings, observations of interpersonal interactions, neuropsychological assessment, behavioral assessment, and using assessment findings as part of the overall therapeutic process (Finn, 2007; Garb, 2007).

The relative popularity of different traditional psychological tests has been surveyed since 1935 in many settings, such as academic institutions, psychiatric hospitals, counseling centers, Veterans Administration centers, institutions for those with developmental disabilities, private practice, and various memberships and

professional organizations. Surveys (somewhat dated) of test usage have usually found that the 10 most frequently used tests are the Wechsler intelligence scales, Minnesota Multiphasic Personality Inventory, Rorschach, Bender Visual Motor Gestalt Test, Thematic Apperception Test, projective drawings (Human Figure Drawing, House-Tree-Person), Wechsler Memory Scale, Beck Depression Inventory, Millon Clinical Multiaxial Inventories, and California Psychological Inventory (Camara et al., 2000; Kamphaus, Petoskey, & Rowe, 2000; Lubin et al., 1985; C. Piotrowski & Zalewski, 1993; Watkins, 1991; Watkins et al., 1995). The pattern for the 10 most popular tests has remained quite stable since 1969, except that the ranking of Human Figure Drawings dropped (Camara et al., 2000). It is expected that some newer measures, especially the Personality Assessment Inventory, would be ranked quite highly in use. However, no recent surveys of test usage have been published. The pattern of test usage varies somewhat across different studies and varies considerably from setting to setting. Schools and centers for those with intellectual disabilities emphasize tests of intellectual abilities, such as the WISC-V and behavior rating scales; counseling centers are more likely to use vocational interest inventories; and psychiatric settings emphasize tests assessing level of pathology, such as the MMPI or MCMI.

One clear change in testing practices has been a relative decrease in the use and status of projective techniques (Groth-Marnat, 2000b; C. Piotrowski, 1999). Criticisms have been wide ranging but have centered on overly complex scoring systems, questionable norms, subjectivity of scoring, poor predictive utility, and inadequate or even nonexistent validity (Garb, 2005a; Garb et al., 2001; D. N. Miller, 2007; Pruitt, Smith, Thelen, & Lubin, 1985; D. Smith & Dumont, 1995). Further criticisms include the extensive time required to effectively learn the techniques, heavy reliance of projective techniques on psychoanalytic theory, and the greater time and cost efficiency of alternative objective tests. These criticisms have usually occurred from within the academic community, where the techniques are used less and less for research purposes (C. Piotrowski, 1999; C. Piotrowski & Zalewski, 1993; Watkins, 1991). As a result of these criticisms, there has been a slight but still noteworthy reduction in the use of the standard projective tests in professional practice (Archer, Buffington-Vollum, Stredny, & Handel, 2006; Camara et al., 2000; Kamphaus et al., 2000; C. Piotrowski, 1999). Although there has been a reduction, the Rorschach and Thematic Apperception Test (TAT) continue to have a strong foothold in clinical practice. This can be attributed to lack of time available for practitioners to learn new techniques, expectations that students in internships know how to use them, unavailability of other practical alternatives, and the fact that practitioners usually give more weight to clinical experience than to empirical evidence. This suggests distance between the quantitative, theoretical world of the academic and the practical, problem-oriented world of the practitioner. In fact, assessment practices in many professional settings seem to have little relationship to the number of research studies done on assessment tools, attitudes by academic faculty, or the psychometric quality of the test (Garb, Wood, Lilienfeld, & Nezworski, 2002). In contrast to the continued use of projective instruments in adult clinical settings, psychologists in child settings are likely to rely more on behavior rating scales (e.g., Child Behavior Checklist) than projective tests (Cashel, 2002; Kamphaus et al., 2000; D. N. Miller, 2007).

The earliest form of assessment was through clinical interview. Clinicians like Freud, Jung, and Adler used unstructured interaction to obtain information regarding history, diagnosis, and underlying structure of personality. Later clinicians organized

interviews using outlines of the areas that should be discussed. During the 1960s and 1970s, much criticism was directed toward the interview, leading many psychologists to perceive interviews as unreliable and lacking empirical validation. Tests, in many ways, were designed to counter the subjectivity and bias of interview techniques. During the 1980s and 1990s, a wide variety of structured interview techniques gained popularity and have often been found to be reliable and valid indicators of a client's level of functioning. Structured interviews such as the Diagnostic Interview Schedule (DIS; Robins, Helzer, Cottler, & Goldring, 1989), Structured Clinical Interview for the DSM (SCID; Spitzer, Williams, & Gibbon, 1987), and Renard Diagnostic Interview (Helzer, Robins, Croughan, & Welner, 1981) are often given preference over psychological tests. These interviews, however, are very different from the traditional unstructured approaches. They have the advantage of being psychometrically sound even though they might lack important elements of rapport, idiographic richness, and flexibility that characterize less structured interactions (Garb, 2007; R. Rogers, 2001).

A further trend has been the development of neuropsychological assessment (see Groth-Marnat, 2000a; Lezak, Howieson, Bigler, & Tranel, 2012). The discipline is a synthesis between behavioral neurology and psychometrics and was created from a need to answer questions such as the nature of a person's organic deficits, severity of deficits, localization, and differentiating between functional versus organic impairment. The pathognomonic sign approach and the psychometric approaches are two clear traditions that have developed in the discipline. Clinicians relying primarily on a pathognomonic sign approach are more likely to interpret specific behaviors such as perseverations or weaknesses on one side of the body, which are highly indicative of the presence and nature of organic impairments. These clinicians tend to rely on the tradition of assessment associated with Luria (Bauer, 2000; Luria, 1973) and base their interview design and tests on a flexible method of testing possible hypotheses for different types of impairment. In contrast, the more quantitative tradition represented by Reitan and his colleagues (Reitan & Wolfson, 1993; Russell, 2000) is more likely to rely on critical cutoff scores, which distinguish between normal persons and those with brain damage. Reitan and Wolfson (1985, 1993) have recommended using an impairment index, which is the proportion of brain-sensitive tests that fall into the brain-damaged range. In actual practice, most clinical neuropsychologists are more likely to combine the psychometric and pathognomonic sign approaches (Rabin, Barr, & Burton, 2005). The two major neuropsychological test batteries are the Luria-Nebraska Neuropsychological Battery (Golden, Purisch, & Hammeke, 1985) and the Halstead Reitan Neuropsychological Test Battery (Reitan & Wolfson, 1993). A typical neuropsychological battery might include tests specifically designed to assess organic impairment along with tests such as the MMPI, Wechsler intelligence scales, and the Wide Range Achievement Test (WRAT-4). As a result, extensive research over the past 15 to 20 years has been directed toward developing a greater understanding of how the older and more traditional tests relate to different types and levels of cerebral dysfunction.

During the 1960s and 1970s, behavior therapy was increasingly used and accepted. Initially, behavior therapists were concerned with an idiographic approach to the functional analysis of behavior. As their techniques became more sophisticated, formalized methods of behavioral assessment began to arise. These techniques arose in part from

dissatisfaction with the methods of diagnosis of the second edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM-II*; American Psychiatric Association, 1968), as well as from a need to have assessment relate more directly to treatment and its outcomes. There was also a desire to be more accountable for documenting behavior change over time. For example, if behaviors related to anxiety decreased after therapy, the therapist should be able to demonstrate that the treatment had been successful. Behavioral assessment could involve measurements of movements (behavioral checklists, behavioral analysis), physiological responses (galvanic skin response [GSR], electromyograph [EMG]) or self-reports (self-monitoring, Symptom Checklist-90-R, assertiveness scales). Whereas the early behavioral assessment techniques showed little concern with the psychometric properties of their instruments, there has been an increasing push to have them meet adequate levels of reliability and validity (First, Frances, Widiger, Pincus, & Davis, 1992; Follette & Hayes, 1992). Despite the many formalized techniques of behavioral assessment, many behavior therapists feel that an unstructured, idiographic approach is most appropriate.

Traditional means of assessment, then, have decreased because of an overall increase in other activities of psychologists and an expansion in the definition of assessment. Currently, a psychologist doing assessment might include such techniques as interviewing, administering, and interpreting traditional psychological tests (MMPI-2/MMPI-A/MMPI-2-RF, WAIS-IV, etc.), naturalistic observations, neuropsychological assessment, and behavioral assessment. In addition, professional psychologists might be required to assess areas that were not given much emphasis before the 1980s: personality disorders (borderline personality, narcissism), stress and coping (life changes, burnout, existing coping resources), hypnotic responsiveness, psychological health, adaptation to new cultures, changes associated with increasing modernization, and strengths (related to positive psychology movements). Additional areas might include family systems interactions, relation between a person and his or her environment (social climate, social supports), cognitive processes related to behavior disorders, and level of personal control and self-efficacy. All these require clinicians to be continually aware of new and more specific assessment devices and to maintain flexibility in the approaches they take.

The future of psychological assessment will probably be most influenced by the trends toward computerized assessment, adaptation to managed health care, and distance health care delivery (Groth-Marnat, 2000b, 2009; Kay, 2007). Computerized assessment is likely to enhance efficiency through rapid scoring, complex decision rules, reduction in client–practitioner contact, novel presentation of stimuli (i.e., virtual reality), and generation of interpretive hypotheses (Lichtenberger, 2006). Future assessments are also likely to tailor the presentation of items based on the client's previous responses (Forbey & Ben-Porath, 2007). Unnecessary items will not be given, with one result being that a larger amount of information will be obtained through the presentation of relatively fewer items. This time efficiency is in part stimulated by the cost-savings policies of managed care, which require psychologists to demonstrate the cost-effectiveness of their services (Groth-Marnat, 1999; Groth-Marnat & Edkins, 1996). In assessment, this means linking assessment with treatment planning. Thus, psychological reports of the future are likely to need to link client dynamics directly to recommendations and treatment options. Whereas considerable evidence supports the

cost-effectiveness of using psychological tests in organizational contexts, health care needs to demonstrate that assessment can increase the speed of treatment as well as optimize treatment outcome (Blount et al., 2007; Groth-Marnat, 1999; Groth-Marnat, Roberts, & Beutler, 2001; Lambert & Hawkins, 2004; Yates & Taub, 2003).

A further challenge and area for development is the role distance health will play in assessment (Leigh & Zaylor, 2000; M. J. Murphy, Levant, Hall, & Glueckauf, 2007). Distance assessment as a means in and of itself is likely to become important. Professional psychologists may be required to change their traditional face-to-face role to one of developing and monitoring new applications as well as consulting/collaborating with clients regarding the results of assessments derived from the computer.

EVALUATING PSYCHOLOGICAL TESTS

Before using a psychological test, clinicians should investigate and understand the theoretical orientation of the test, practical considerations, the appropriateness of the standardization sample, and the adequacy of its psychometric properties (reliability and validity). Often, helpful descriptions and reviews that relate to these issues can be found in the test manuals as well as past and future editions of the *Mental Measurements Yearbook* (Carlson, Geisinger, & Jonson, 2014); *Tests in Print* (L. L. Murphy, Geisinger, Carlson, & Spies, 2011); *Tests: A Comprehensive Reference for Assessment in Psychology, Education, and Business* (Maddox, 2003); and *Measures for Clinical Practice: A Sourcebook* (Fischer & Corcoran, 2007). Reviews can also be found in assessment-related journals, such as the *Journal of Personality Assessment*, the *Journal of Psychoeducational Assessment*, and *Educational and Psychological Measurement*. Table 1.1 outlines the more important questions that should be answered. Each issue outlined in this table is discussed further. The discussion reflects a practical focus on problems that clinicians using psychological tests are likely to confront. It is not intended to provide a comprehensive coverage of test theory and construction; if a more detailed treatment is required, the reader is referred to one of the many texts on psychological testing (e.g., Aiken & Groth-Marnat, 2006; R. M. Kaplan & Saccuzzo, 2005).

Theoretical Orientation

Before clinicians can effectively evaluate whether a test is appropriate, they must understand its theoretical orientation. Clinicians should research the construct that the test is supposed to measure and then examine how the test approaches this construct. This information can usually be found in the test manual. If for any reason the information in the manual is insufficient, clinicians should seek it elsewhere. Clinicians can often obtain additional useful information regarding the construct being measured by carefully studying the individual test items. Usually the manual provides an individual analysis of the items, which can help the potential test user evaluate whether they are relevant to the trait being measured.

Practical Considerations

A number of practical issues relate more to the context and manner in which the test is used than to its construction. First, tests vary in terms of the level of education

Table 1.1 Evaluating a Psychological Test

Theoretical Orientation

1. Do you adequately understand the theoretical construct the test is supposed to be measuring?
2. Do the test items correspond to the theoretical description of the construct?

Practical Considerations

1. If reading is required by the examinee, does his or her ability match the level required by the test?
2. How appropriate is the length of the test?

Standardization

1. Is the population to be tested similar to the population the test was standardized on?
2. Was the size of the standardization sample adequate?
3. Have specialized subgroup norms been established?
4. How adequately do the instructions permit standardized administration?

Reliability

1. Are reliability estimates sufficiently high (generally around .90 for clinical decision making and around .70 for research purposes)?
2. What implications do the relative stability of the trait, the method of estimating reliability, and the test format have on reliability?

Validity

1. What criteria and procedures were used to validate the test?
 2. Will the test produce accurate measurements in the context and for the purpose for which you would like to use it?
-

(especially reading skill) that examinees must have to understand them adequately. The examinee must be able to read, comprehend, and respond appropriately to the test. Second, some tests are too long, which can lead to a loss of rapport with or extensive frustration on the part of the examinee. Administering short forms of the test may reduce these problems, provided these forms have been properly developed and are interpreted with appropriate caution. Finally, clinicians have to assess the extent to which they need training to administer and interpret the instrument. If further training is necessary, a plan must be developed for acquiring this training.

Standardization

Another central issue relates to the adequacy of norms (see Cicchetti, 1994). Each test has norms that reflect the distribution of scores by a standardization sample. The basis on which individual test scores have meaning relates directly to the similarity between the individual being tested and the sample. If a similarity exists between the group or individual being tested and the standardization sample, adequate comparisons can be made. For example, if the test was standardized on white American college students between the ages of 18 and 22, useful comparisons can be made for college students in

that racial and age bracket (if we assume that the test is otherwise sufficiently reliable and valid). The more dissimilar the person is from this standardization group (e.g., different national group, over 70 years of age), the less useful the test is for evaluation. The examiner may need to consult the literature to determine whether research that followed the publication of the test manual has developed norms for different groups. This is particularly important for tests such as the MMPI and the Rorschach, for which norms for various cross-national populations have been published.

Three major questions that relate to the adequacy of norms must be answered. The first is whether the standardization group includes representation from the population on which the examiner would like to use the test. The test manual should include sufficient information to determine the representativeness of the standardization sample. If this information is insufficient or in any way incomplete, it greatly reduces the degree of confidence with which clinicians can use the test. The ideal and current practice is to use stratified random sampling. However, because this can be an extremely costly and time-consuming procedure, many tests do not meet this standard. The second question is whether the standardization group is large enough. If the group is too small, the results may not give stable estimates because of too much random fluctuation. Finally, a test may have specialized subgroup norms as well as broad national norms. Knowledge relating to subgroup norms gives examiners greater flexibility and confidence if they are using the test with similar subgroup populations (see Dana, 2005). This is particularly important when subgroups produce sets of scores that are significantly different from the normal standardization group. These subgroups can be based on factors such as ethnicity, sex, geographic location, age, level of education, socioeconomic status, urban versus rural environment, or even diagnostic history. Knowledge of each of these subgroup norms allows for a more appropriate and meaningful interpretation of scores.

Standardization can also refer to administration procedures. A well-constructed test should have clear instructions that permit examiners to give the test in a manner similar to that of other examiners and also similar to themselves from one testing session and the next. Research has demonstrated that varying the instructions between one administration and the next can alter the types and quality of responses the examinee gives, thereby compromising the test's reliability. Standardization of administration should refer not only to consistent administration procedures but also to ensuring adequate lighting, quiet, no interruptions, and good rapport.

Reliability

The reliability of a test refers to its degree of stability, consistency, and predictability. It addresses the extent to which scores obtained by a person are or would be the same if the person is reexamined by the same test on different occasions. Underlying the concept of reliability is the possible range of error, or error of measurement, of a single score. This is an estimate of the range of possible random fluctuation that can be expected in an individual's score. Because psychological constructs cannot be measured directly (e.g., through measuring a level in blood), test scores are at best an approximation of these constructs, and thus error is always present in the system. It may arise from such factors as a misreading of the items, poor administration procedures, or the

changing mood of the client. If there is a large degree of error, the examiner cannot place a great deal of confidence in an individual's scores. The goal of a test constructor is to reduce, as much as possible, the degree of measurement error. If this error reduction is achieved, the difference between one score and another for a measured characteristic is more likely to result from some true difference than from some chance fluctuation.

Two main issues relate to the degree of error in a test. The first is the inevitable, natural variation in human performance. Typically variability is less for measurements of ability than for those of personality and state of being. Whereas ability variables (intelligence, mechanical aptitude, etc.) may show gradual changes resulting from growth and development, many personality traits and states of being are much more highly dependent on factors such as mood. This is particularly true in the case of a characteristic such as anxiety. The practical significance of this in evaluating a test is that certain factors outside the test itself can serve to reduce the reliability that the test can realistically be expected to achieve. Thus, an examiner should generally expect higher reliabilities for an intelligence test than for a test measuring a personality variable such as anxiety. It is the examiner's responsibility to know what is being measured, especially the degree of variability to be expected in the measured trait.

The second important issue relating to reliability is that psychological testing methods are necessarily imprecise. For the hard sciences, researchers can make direct measurements, such as the concentration of a chemical solution, the relative weight of one organism compared with another, or the strength of radiation. In contrast, many constructs in psychology are often measured indirectly. For example, intelligence cannot be perceived directly; it must be inferred by measuring behavior that has been defined as being intelligent. Variability relating to these inferences is likely to produce a certain degree of error resulting from the lack of precision in defining and observing inner psychological constructs. Variability in measurement also occurs simply because people have true (not because of test error) fluctuations in performance between one testing session and the next. Whereas it is impossible to control for the natural variability in human performance, adequate test construction can attempt to reduce the imprecision that is a function of the test itself. Natural human variability and test imprecision make the task of measurement extremely difficult. Although some error in testing is inevitable, the goal of test construction is to keep testing errors within reasonably accepted limits. A high measure of reliability is generally .80 or more, but the variable being measured also changes the expected strength of the statistic. Likewise, the method of determining reliability alters the relative strength of the statistic. Ideally, clinicians should hope for reliability statistics of .90 or higher in tests that are used to make decisions about individuals, whereas a reliability of .70 or more is generally adequate for research purposes.

The purpose of reliability is to estimate the degree of test variance caused by error. The four primary methods of obtaining reliability involve determining (1) the extent to which the test produces consistent results upon retesting (test-retest), (2) the relative accuracy of a test at a given time (alternate forms), (3) the internal consistency of the items (split-half and coefficient alpha), and (4) the degree of agreement between two examiners (interscorer). Another way to summarize this is that reliability can be time to time (test-retest), form to form (alternate forms), item to item (split-half/coefficient

alpha), or scorer to scorer (interscorer). Although these are the main types of reliability, there is a fifth type, the Kuder-Richardson; like the split-half and coefficient alpha, it is a measurement of the internal consistency of the test items. However, because this method is considered appropriate only for tests that are relatively pure measures of a single variable, it is not covered in this book.

Test-Retest Reliability

Test-retest reliability is determined by administering the test and then repeating it on a second occasion. The reliability coefficient is calculated by correlating the scores obtained by the same person on the two different administrations. The degree of correlation between the two scores indicates the extent to which the test scores can be generalized from one situation to the next. If the correlations are high, the results are less likely to be caused by random fluctuations in the condition of the examinee or the testing environment. Thus, when the test is being used in actual practice, the examiner can be relatively confident that differences in scores are the result of an actual change in the trait being measured rather than error.

A number of factors must be considered in assessing the appropriateness of test-retest reliability. One is the potential for practice and memory of a test taken on one occasion to affect performance on a second occasion, termed practice effect. Some tasks can simply improve between one administration and the next because of practice. This is a particular problem for speeded and memory tests, such as those found on the Coding and Arithmetic subtests of the WAIS-IV. Another factor to consider is that the interval between administrations, which can affect reliability. A test manual should specify the time interval, as well as any likely significant life changes that the examinees may have experienced, such as counseling, career changes, or psychotherapy. For example, tests of preschool intelligence often give reasonably high correlations if the second administration is within several months of the first one. However, correlations with later childhood or adult IQ are generally low because of innumerable, unavoidable intervening life changes. Additional sources of variation may be the result of random, short-term fluctuations in the examinee or of variations in the testing conditions. In general, test-retest reliability is the preferred method only if the variable being measured is relatively stable. If the variable is highly changeable (e.g., anxiety), this method is usually not adequate.

Alternate Forms

The alternate forms method avoids many of the problems encountered with test-retest reliability. The logic behind alternate forms is that, if the trait is measured several times on the same individual by using parallel forms of the test, the different measurements should produce similar results. The degree of similarity between the scores represents the reliability coefficient of the test. As in the test-retest method, the interval between administrations should always be included in the manual, as well as a description of any likely significant intervening life experiences. If the second administration is given immediately after the first, the resulting reliability is more a measure of the correlation between forms and not across occasions. Correlations determined by tests given with a wide time interval, such as two months or more, provide a measure of both the relation between forms and the degree of temporal stability.