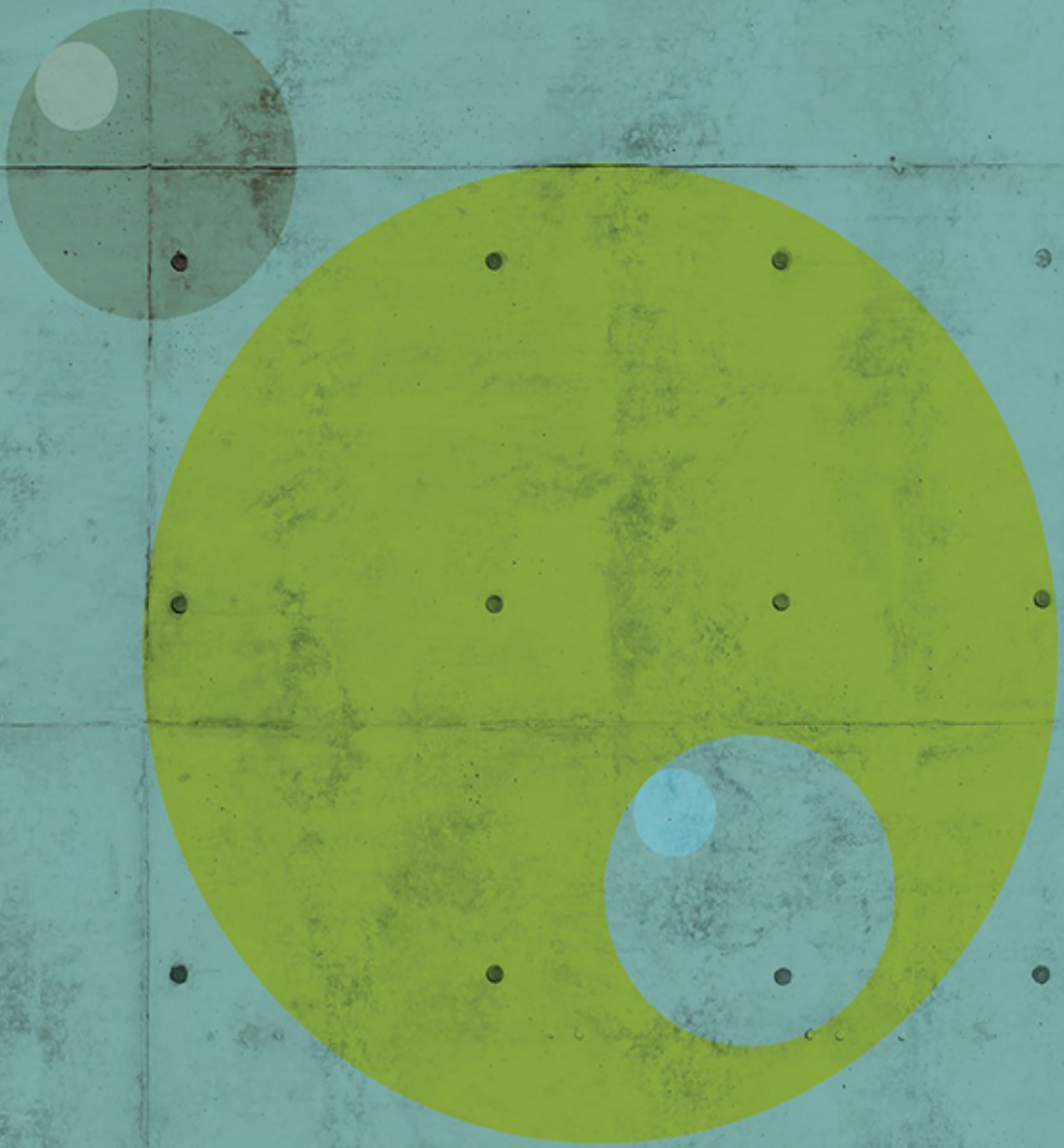
PSYCHOLOGICAL TESTING AND ASSESSMENT

THIRD EDITION

DAVID SHUM,
JOHN O'GORMAN,
PETER CREED,
& BRETT MYORS



OXFORD

PSYCHOLOGICAL TESTING AND ASSESSMENT

THIRD EDITION

PSYCHOLOGICAL TESTING AND ASSESSMENT

THIRD EDITION

DAVID SHUM, JOHN O. GORMAN, P

ETERCREED&BRETTMYORS

OXFORD
UNIVERSITY PRESS
AUSTRALIA & NEW ZEALAND

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

Published in Australia by
Oxford University Press
253 Normanby Road, South Melbourne, Victoria 3205, Australia

First edition published 2006

Second edition published 2013

National Library of Australia Cataloguing-in-Publication entry
Creator: Shum, David H. K. author.
Title: Psychological testing and assessment / David Shum, John O'Gorman, Peter Creed, Brett Myors.
Edition: Third edition

ISBN: 9780190305208 (paperback)

Notes: Includes bibliographical references and index.

Subjects: Psychological tests. Behavioral assessment Psychology–Methodology.

Other Creators/Contributors: O'Gorman, John, author. Creed, Peter, author. Myors, Brett, author.

Edited by Pete Cruttenden
Cover credit by Shutterstock
Typeset by Newgen KnowledgeWorks Pvt. Ltd., Chennai, India
Proofread by Liz Filleul
Indexed by Jeanne Rudd
Printed in China by Leo Paper Products Ltd.

CONTENTS

Chapter 1: Psychological Tests: What Are They and Why Do We Need Them?

Introduction
A brief history of psychological testing
Psychological tests: why do we need them?
Psychological tests: definitions, advantages and limitations

Chapter 2: Psychological Testing and Assessment: Processes, Best Practice and Ethics

Introduction
Psychological testing versus psychological assessment
Areas of application
Types of psychological tests
Processes and best practices in psychological testing
Ethics
Accommodating the differently abled
Cultural differences, testing and assessment

PART 2 Methodological and Technical Principles of Psychological Testing

Chapter 3: Test Scores and Norms

Introduction
Interpreting test scores
Transforming scores for norm referencing
Standard scores and transformed scores based on them
Percentiles and transformed scores based on them
Relationships among the transformed scores
Other methods of scoring
Norms

Chapter 4: Reliability

Introduction
The meaning of reliability
The domain-sampling model
Calculating reliability coefficients
Extending the domain-sampling model
Some special issues

Chapter 5: Validity

Introduction
The meaning of validity
Content validity
Predictive validity
Construct validity
Factor analysis

Chapter 7: Intelligence

Introduction
The concept of intelligence
Binet's revolution
Spearman and 'g'
Terman and the Stanford-Binet Intelligence Scale
Wechsler scales
Thurstone and multiple mental abilities
Guilford: A different structure of intelligence
Vernon's hierarchical view of intelligence
Cattell's two-factor theory of intelligence
Cattell, Horn and Carroll extend the '*Gf-Gc*' model of intelligence
The Cattell-Horn-Carroll (CHC) model of intelligence
The CHC model and modern tests of intelligence
A developmental conception of intelligence
An information-processing view of intelligence
Gardner and multiple intelligences
Sternberg's triarchic theory of intelligence
So, 'what is intelligence?'
Aptitude versus achievement tests
Group (rather than individual) testing
Group differences in intelligence

Chapter 8: Personality

Introduction
The psychoanalytic approach
The interpersonal approach
The personological approach
The multivariate (trait) approach
The empirical approach
The social-cognitive approach
Positive psychology
Eclectic approaches

PART 4 Areas of Professional Application

Chapter 9: Clinical and Mental Health Testing and Assessment

Introduction
Clarifying the referral question
Case history data
Clinical interview
Mental status examination
Psychological tests
Psychological report

Chapter 10: Organisational Testing and Assessment

Introduction
Performance appraisal
Theories of performance
Personnel selection
Selection as a social process
Some tests used in selection
Work attitudes
Vocational interests
Strong Interest Inventory

Chapter 11: Neuropsychological Testing and Assessment

Introduction
What is clinical neuropsychology?
A brief history of neuropsychological assessment
What is neuropsychological assessment?
Purposes and procedures of neuropsychological assessment
Neuropsychological functions commonly assessed

Chapter 12: Forensic Psychological Testing and Assessment

Introduction
Forensic psychology and forensic psychological testing and assessment
Settings of forensic assessment
Differences between forensic and therapeutic assessment
Psychological tests and assessment techniques commonly used in forensic assessment
Limitations of forensic assessment

Chapter 14: The Future of Testing and Assessment

Introduction
Construct development
Technical and methodological developments
Contextual changes

Answers to Exercises
Technical Appendix
Glossary
References
Index

FIGURES

TABLES

PREFACE

In 2006 when we published *Psychological Testing and Assessment*, we wanted to provide our readers with a comprehensive introduction to the theory, research and best practice in this important field. In addition, we wanted to present not just the psychological tests but also the principles and practice of this professional speciality within an Australian context, using relevant examples and discussing local professional issues and controversies. This was trying to address the gap that no Australian textbook on psychological testing had been published. At that time, we were not envisaging publishing a revised edition of the book. Thanks to support from our readers and critical and constructive feedback from our colleagues (particularly, Dr Ian Price, University of New England and Dr Kate Jacobs, Monash University), we were fortunate to have the opportunity to produce a second edition in 2013 and a third edition in 2017.

In this third edition, we have taken into consideration feedback provided by users and colleagues and made a number of major changes and improvements to our book. In particular, we have added more discussion about assessing Aboriginal and Torres Strait Islander peoples, people with disabilities, and people with diverse cultural backgrounds; we reorganised Chapters 3–6 by moving some of the more technical details on psychometrics to a Technical Appendix; we added more description of psychological tests that are included by the Psychology Board of Australia in its national examination; we expanded Chapter 14 (The Future of Testing and Assessment); and we moved vocational assessment from Chapter 13 (Educational Testing and Assessment) to Chapter 10 (Organisational Testing and Assessment). Other changes include: updating the supplementary materials, adding margin notes for key terms in each chapter, adding new boxes to highlight issues in some of the chapters, updating tests where there have been new editions, and updating references.

ACKNOWLEDGMENTS

We thank Debra James and Melpo Christofi at Oxford University Press for their support and encouragement in the preparation of the third edition of our book. We also thank our copyeditor Pete Cruttenden for his efficient and professional input. Many thanks go to the following contributors for sharing and updating their practitioner profiles in Part 4 of the book: Professor Amanda Gordon, Dr Elizabeth Allworth, Dr Jan Ewing, Dr Danielle Schumack and Associate Professor Tim Hannan. Finally, we acknowledge the contribution of our research assistant, Candice Bowman, for helping us to research and prepare materials to revise the book.

The authors and the publisher also wish to thank the following copyright holders for reproduction of their materials:

Code of Ethics of the Australian Psychological Society (2007), 39; Getty Images/Time Life Pictures, 8; Profile excerpted from the MMPI(R)-2 (Minnesota Multiphasic Personality Inventory (R)-2) Manual for Administration, Scoring and Interpretation, Revised Edition. Copyright © 2001 by the Regents of the University of Minnesota. Used by permission of the University Minnesota Press. All rights reserved. 'MMPI' and 'Minnesota Multiphasic Personality Inventory' and trademarks owned by the Regents of the University of Minnesota. 205 Stanford University Libraries, 244; National Archives of Australia, 12; Extract 'The Dirty Dozen: A concise measure of the Dark Triad,' Psychological Assessment, 22(2), 420–32, by P.K. Jonason & D. G. Webser, 2010, 174; Shutterstock, front cover, 25 (both), 169, 322, 324, 331; State Library of Queensland, 11.

Every effort has been made to trace the original source of copyright material contained in this book. The publisher will be pleased to hear from copyright holders to rectify any errors or omissions.

PART 1 THE CONTEXT OF PSYCHOLOGICAL TESTING AND ASSESSMENT

1 Psychological Tests: What Are They and Why Do We Need Them?

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. explain how psychological tests have developed over time
2. define what a psychological test is and explain its defining characteristics
3. explain how psychological tests are better than other means used to assist people to understand behaviour and make decisions
4. explain the advantages and limitations of psychological tests.

KEY TERMS

criterion-referenced test
norm-referenced test
objective procedure
psychological test
psychometric properties
test obsolescence

Setting the scene

- Ambulance Victoria will introduce pre-employment psychological testing for new graduates to identify and intervene early with paramedics who might be at risk of suicide, as the suicide rate for paramedics is much higher than that for other workers. (*The Age*, 11 August 2015)
- The Medical Board of Australia released new guidelines that require children who want to undergo cosmetic surgery, but who don't have a medical justification, to complete mandatory psychological assessment. (*Sunday Mail*, 1 April 2012)
- A leading psychologist called for mandatory psychological testing of young drivers to ensure that their brains are 'mature' enough to be granted driving licences. (*Herald Sun*, 14 February 2010)
- Staff in Australian Football League (AFL) clubs used results of neuropsychological tests to determine when players who had suffered concussion should play for the team again. (*Herald Sun*, 22 July 2003)

Introduction

The development and application of **psychological tests** is considered one of the major achievements of psychologists in the last century (O'Gorman, 2007; Zimbardo, 2004, 2006). The news items above illustrate some of the ways psychological tests have been applied in our society. For the most part, tests are used to assist in promoting self-understanding or making decisions by providing more accurate and detailed information about human behaviour than is available without them. Psychological tests are also important tools for conducting psychological research. In this book our focus is on the former rather than the latter application.

psychological test
an objective procedure for sampling and quantifying human behaviour to make inferences about a particular psychological construct or constructs using standardised stimuli and methods of administration and scoring

The ability to select, administer, score and interpret psychological tests is considered a core competency skill for professional psychologists (Australian Psychology Accreditation Council, 2010; Psychology Board of Australia, 2016a). In Australia, assessment is one of the four content domains (the other domains are ethics, interventions, and communication) of the National Psychology Examination administered by the Psychology Board and it is one of the content areas required for psychology course accreditation. Thus, the teaching of this competency is typically included as one or more subjects for undergraduate and postgraduate psychology courses in Australia and other countries.

What are psychological tests and what are their defining characteristics? Who developed the first psychological test and how has psychological testing progressed over time? What are the advantages of using psychological tests to promote understanding and to assist decision-making processes about people in our society? What are the advantages and limitations of psychological tests? These are the topics of the first chapter of this book.

A brief history of psychological testing

The history of psychological testing has been well documented by DuBois (1970). O'Neil (1987), Keats and Keats (1988) and Ord (1977) have provided accounts of relevant developments in Australia. The following section draws freely on these sources (Box 1.1 highlights some of these historical developments).

Box 1.1

Timeline of major developments in the history of psychological testing

1890	The term 'mental test' is first used by James McKeen Cattell
1905	Alfred Binet and Theodore Simon devise the first test of intelligence for use with children
1916	Lewis Terman publishes the Stanford-Binet test, based on the pioneering work of Binet and Simon
1917	Robert Yerkes leads the development of the Army Alpha and Beta tests for selection for military service in the USA
1917	Robert Woodworth devises the first self-report test of personality
1921	Hermann Rorschach, a Swiss psychiatrist and psychoanalyst, publishes *Psychodiagnostics* on the use of inkblots in evaluating personality
1927	The first version of the Strong Vocational Interest Blank is published
1938	Oscar Buros publishes the first compendium of psychological tests, the *Mental Measurements Yearbook*
1939	David Wechsler develops an individual test of adult intelligence
1942	The *Minnesota Multiphasic Personality Inventory* (MMPI) is published to assist the differential diagnosis of psychiatric disorders
1948	Henry Murray and colleagues publish *Assessment of Men*, and the

	term 'assessment' comes to replace mental testing as a description of work with psychological tests
1957	Raymond Cattell publishes on performance tests of motivation
1962	Computer interpretation of the MMPI is introduced
1968	Walter Mischel publishes his widely cited critique of personality assessment
1970	Computers are used for testing clients; computerised adaptive testing follows
1971	The Federal Court in the USA challenges testing for personnel selection
1985	Publication in the USA of the first edition of the *Standards for Educational and Psychological Testing*
1988	Jay Ziskin and David Faust challenge the use of psychological test results in court
1993	The American Psychological Association publishes guidelines for computer-based testing and interpretation
1993	John Carroll publishes *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*, in which he proposes his three-stratum theory of intelligence
1999	The second edition of the *Standards for Educational and Psychological Testing* is published
2001	Gregory Meyer and colleagues publish the results of a review of 125 earlier literature reviews indicating the value of psychological tests

Based on a more extensive timeline in Sundberg (1977)

Binet and the birth of psychological testing

The origins of psychological testing can be found in the public service examinations used by Chinese dynasties to select those who would work for them. These were large-scale exercises involving many applicants and several days of testing, which from the era of the Han dynasty (206 BCE–220 CE) involved written examinations (Bowman, 1989). Programs of testing were conducted from about 2000 BCE to the early years of the twentieth century when they were discontinued, at about the time the modern era of psychological testing was being introduced in the USA. A major impetus to this modern development of testing was the need to select men for military service when the USA entered the First World War without a standing army. There were, however, a number of precursors to this development, the most significant being the work

of Alfred Binet (1857–1911) and his colleagues in France in the late nineteenth and early twentieth centuries.

Binet was asked by the Office of Public Instruction in Paris to provide a method for objectively determining which children would benefit from special education. In responding to this request, Binet devised the first of the modern intelligence tests, using problems not unlike those covered in a normal school program. In the process, he proposed a method for quantifying intelligence in terms of the concept of mental age; that is, the child's standing among children of different chronological ages in terms of his or her cognitive capacity. For example, a child whose knowledge and problem-solving ability was similar to that of the average 10-year-old was described as having a mental age of 10 years. The child's chronological age might be in advance or behind that. Binet showed how a test of intelligence might be validated by comparing the test performance of older with younger children, or the performance of those considered bright by their teachers with those considered dull. Given our understanding of ability, older children should do better than younger children on a test purporting to be a test of intelligence, and bright children should perform better than dull children. Determining the appropriate content, finding a unit of measurement and specifying methods for validating tests of this sort were all significant achievements, with the result that Binet is often thought of as the originator of psychological testing. Binet himself might not have been entirely pleased with this honour, because he was more concerned with the remediation of difficulties than with the classification process that has preoccupied many who adopted his procedures.

The assumption implicit in Binet's work—that performance on a range of apparently different problems can be aggregated to yield an overall estimate of, in his terms, mental age—was examined by Charles Spearman (1863–1945) in the UK in a series of investigations that yielded the first theory of intelligence. This theory proposed that there was something common to all tests of cognitive abilities: g in Spearman's terms. This proposal was to be sharply criticised by a number of US researchers, chief among them Louis Thurstone (1887–1955).

The theoretical arguments did not deter a number of researchers from adapting Binet's test to the cultural milieu in which they worked. Henry Goddard (1866–1957) in the USA, Cyril Burt (1883–1971) in the UK and Gilbert Phillips (1900–1975) in Australia all developed versions of Binet's test, but it was Lewis Terman (1877–1956) at Stanford University who published the most ambitious version for use with English speakers. His test was appropriate for children aged from 3 years to 16 years. It was Terman's version, which he termed the Stanford-Binet, which was to dominate as a test of intelligence for individuals until David Wechsler (1896–1981) published a test for the individual assessment of adult intelligence in 1948.

Figure 1.1 Imperial examination in China

Binet's test and the adaptations of it depended heavily on tapping skills that were taught in school, which were dominated by verbal skills. A number of researchers saw the need for practical or performance tests of ability that did not depend on verbal skills or exposure to mainstream formal schooling. One of the earliest of these researchers was Stanley Porteus (1883–1972), who in 1915 reported the use of mazes for assessing comprehension and foresight. Porteus was born and educated in Australia, but spent most of his working life in the USA, first at the Vineland Institute in New Jersey and then at the University of Hawaii. He returned to Australia from time to time to study the abilities of Aboriginal Australians. His test required the test taker to trace with a pencil increasingly complex mazes while avoiding dead ends and not lifting the pencil from the paper. The test is still used by neuropsychologists in assessing executive functions. Porteus's work was the forerunner of the development in Australia of a number of tests of ability that are not dependent on access to English for their administration, the most notable of which was the Queensland Test by Donald McElwain (1915–2000) and George Kearney (1939–). The administrator of this test used mime to indicate task requirements. In New Zealand, tests of cognitive ability for Māori children were undertaken by Ross St George (see Ord, 1977).

Binet's test and its adaptations, and the early performance tests were individual tests of ability as they required administration to one person at a time. An individual test of intelligence was of little use when thousands of individuals had to be tested in a short space of time—the situation in the First World War.

Arthur Otis (1886–1964) in the USA and Cyril Burt (1883-1971) in the UK trialled a variety of group tests of intelligence, but the most convincing demonstration of their usefulness was to come from Clarence Yoakum (1879–1945) and Robert Yerkes (1876–1956) and their colleagues, who developed two group tests of general mental ability for use with recruits to the US armed services during the First World War. The Army Alpha test was developed for assessing the ability levels of those who could read and write, and the Army Beta test for those who were not literate. Although there is some dispute about how valuable the Army Alpha and Beta tests were to the war effort, they gave considerable impetus to psychological testing in the postwar period, and their basic structure was used subsequently by Wechsler when developing the Verbal and Performance subscales for his test of adult intelligence.

Wechsler developed his test for use in an adult inpatient psychiatric setting as an aid in differential diagnosis. A patient in this setting might present with symptoms of schizophrenia or alcoholism, or be of low general intelligence. Wechsler sought a test that would provide not just an overall assessment of intellectual level, but would also assist in identifying which possible diagnosis was the most likely. The use of the test for this purpose has been criticised, but it is clear that, as an individual test of general ability for adults, Wechsler's test was superior to the Stanford-Binet. Not only was the content more age appropriate, but Wechsler also replaced the mental age scoring method with the Deviation IQ method, which was based on earlier work by Godfrey and his team in Edinburgh (Vernon, 1979). The Deviation IQ method compared the performance of the individual with that of his or her age peers by dividing the difference between the individual's score and the mean for the peer group by the standard deviation of scores for the peer group. The idea was used in a subsequent revision of the Stanford-Binet (the LM revision) and continues to this day in both the Wechsler and the Binet tests.

Figure 1.2 Group testing of US army recruits during the First World War

Woodworth and the beginnings of personality testing

During the First World War, Robert Woodworth (1869–1962) developed the first self-report personality test. This was a screening test for psychological adjustment to the military situation and comprised short questions identified from textbooks of psychiatry and other expert sources. It was used as a screening test, with the endorsement of a certain number of items in a direction suggestive of psychopathology leading to further evaluation by a military psychiatrist. It was the forerunner of a number of self-report tests, the most notable being the Minnesota Multiphasic Personality Inventory (MMPI) developed by Starke Hathaway (1903–1984) and John McKinley (1891–1950) at Minnesota in 1942. This test was designed to discriminate between those without symptoms of mental illness ('normals') and patient groups with particular diagnoses. Items were sought that would yield two clear patterns of response: one characteristic of normals and the other characteristic of a particular patient group (e.g. patients diagnosed with schizophrenia). The same strategy ('empirical keying', as it came to be called) had been used by Edward Strong (1884–1963) in his development of a test of vocational interest in 1928, which provided a basis for occupational and

vocational guidance. The MMPI was long (566 items), heterogeneous in content, and sophisticated to the extent that it included four validity scales for the purpose of identifying various forms of untruthful responding by the test taker that could invalidate inferences drawn from the content scales.

These various tests of cognitive and personality functioning provided a modest but important adjunct to clinical judgment, the principal method of evaluation practised until that time. Just as physical medicine relied on various tests of physiological functioning (e.g. the X-ray or blood test) to aid the process of judgment, so the mental test became a supplement to the unaided diagnostic ability of the doctor or psychiatrist.

The various tests mentioned to this point are sometimes described as 'objective', meaning that the method of scoring is sufficiently straightforward for two or more scorers of the same test performance to agree closely on the final score. There is another category of tests (or techniques, as advocates prefer to call them) that involves a good deal of judgment in their scoring. These 'projective techniques' had their genesis in psychodynamic theorising. Freud's fundamental assumption of psychic determinism—that all mental events have a cause—was taken to mean that no behaviour is accidental but that it betrays the operation of unconscious motivational effects. With such a premise, Hermann Rorschach (1884–1922), a Swiss psychiatrist and follower of Jung's theory, developed a test that purported to identify the psychological types that Jung postulated. The test involved a series of blots created by pouring ink on a page and folding the page in half. Such a random process gave rise to meaningless designs that the patient was asked to make sense of. In so doing, as Henry Murray (1893–1988) was later to formulate in the projective hypothesis, test takers are obliged to draw on their own psychic resources and thus demonstrate something of the workings of their mind. Expertise was essential for interpretation and required careful study of the interpretative strategies of psychodynamic theory.

With the acceptance of projective techniques, the task of testing was raised from a technical routine activity to one requiring the exercise of considerable judgment. A new title was required for this, and Henry Murray provided it. Working at the Psychological Clinic at Harvard University in the 1930s, he and his colleagues set about an intensive study of forty-nine undergraduate students. The project ran for several years and gave rise to Murray's theory of personality and to a number of techniques and procedures for studying personality. One was a projective test called the Thematic Apperception Test (TAT), which he developed with Christiana Morgan (1897–1967), and which became the second most widely used projective technique after the Rorschach. The other was the diagnostic council, a case conference at which all staff involved with a particular participant in the project would provide information and interpretation. From discussion, a consensus view would emerge about the personality structure and dynamics of the individual. When the USA entered the Second World War,

Murray (with a number of other psychologists) joined the war effort. In Murray's case, it was in the Office of Strategic Services, the forerunner of the CIA, which was charged with the task of selecting and preparing volunteers for espionage activities. Murray used many of the techniques from his Harvard days, added situational tests to them, and relied on a form of the diagnostic council. This work was one of the forerunners of the assessment centre, which was to be used successfully by business organisations after the war for the selection and promotion of senior executives. This technique is still used widely today in organisational psychology. Murray reported this wartime work in a book titled *Assessment of Men* (Office of Strategic Services Assessment Staff, 1948). 'Assessment' was the term required for the high-level reasoning process involved in the application of psychological procedures to the individual case, and henceforth almost completely replaced the term 'mental testing'.

Psychological tests under attack

The late 1940s and 1950s represented the peak of psychological testing and assessment, particularly in the USA. One estimate by Goslin in 1963 was that by that date more than 200 million tests of intelligence alone were being administered annually in the USA (Vernon, 1979). A public reaction to this was brewing, however, and hard questions were being asked about the evidence base of the projective techniques, with the theorising of Freud and other psychodynamic theorists being questioned. In the public arena, there were several challenges to psychological testing. One was that it involved a serious invasion of privacy; for example, by some of the questions asked on the self-report tests of personality. A second was the concern about the homogenising effects on the workforce by using psychological tests for selection, with only a limited set of personality characteristics and abilities being acceptable to an organisation. Most damaging to the testing enterprise was the charge that psychological tests were discriminatory. Because black and Hispanic Americans were found to score, on average, lower on ability tests than white Americans, and because test scores were used for selection in a number of workplace and academic settings, psychological tests of this sort were considered to be denying access to many members of minority groups. The criticisms began in magazine articles and popular books, but were given forceful expression in state and federal courts and legislatures. The criticism and legal interventions were more muted outside of the USA, but the critique was by no means limited to that country.

One of the benefits of this critique of psychological testing and assessment was the recognition that psychological testing might be a value-neutral technology in itself, but its application is always in a social context in which outcomes are valued differently by different observers. The most dramatic

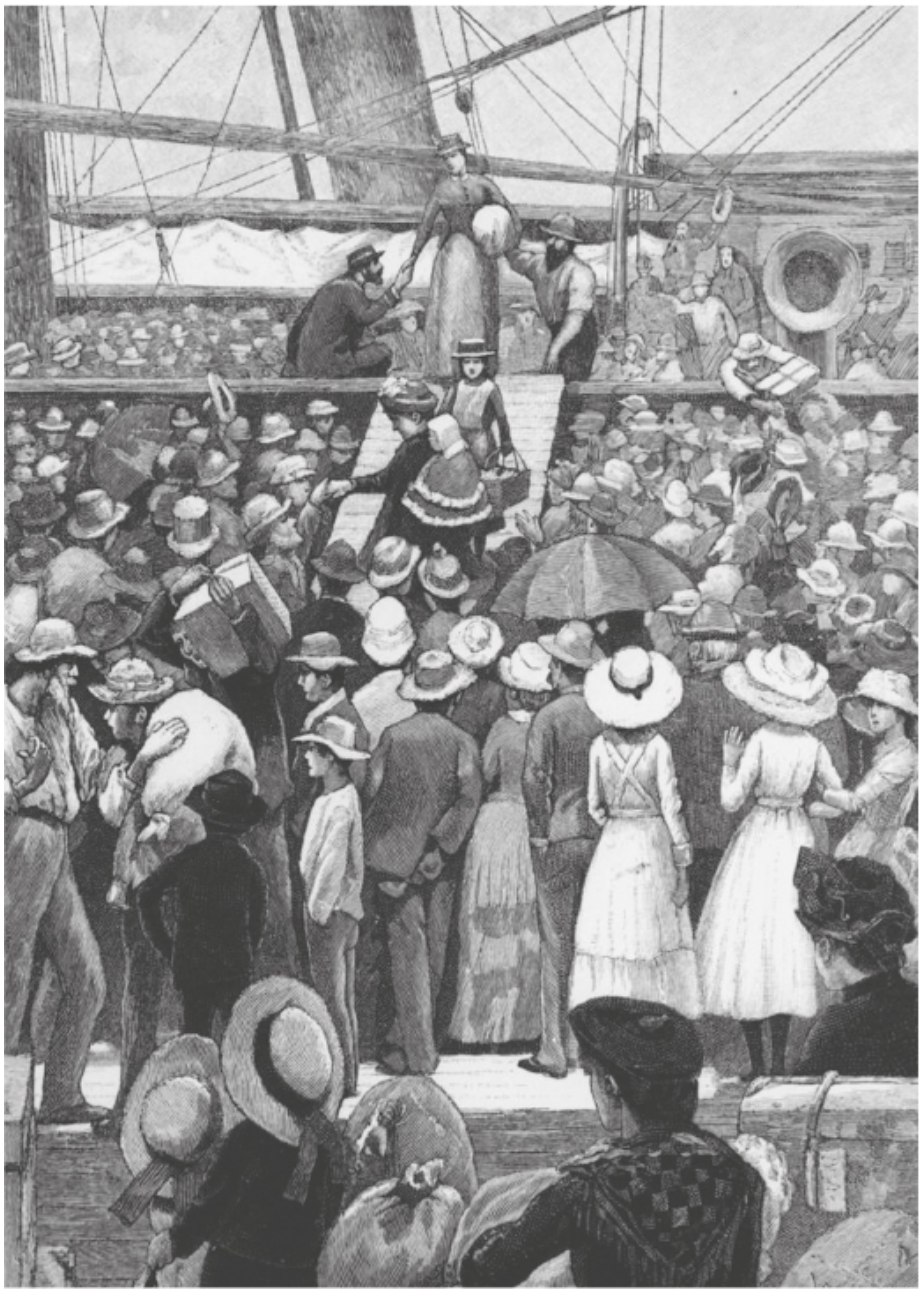
demonstration of this was the use of testing to enforce immigration policies that most of us today would recognise as manifestly unfair and unjust (see Box 1.2 for an example in the Australian context). The moral of the story is clear: test users need to appreciate the social context in which tests are used.

Box 1.2

Testing in the service of ideology

Immigration to the USA was restricted in the first part of the twentieth century by procedures aimed at preventing the entry of 'feebleminded' individuals from European countries who, it was thought, might adversely affect the gene pool or become a burden on the state (Richardson, 2003). Psychological testing formed a part of this process, which was supported by a social consensus on the dangers of unrestricted migration.

Figure 1.3 Drawing of migrants disembarking from a ship, circa 1885



In Australia, a similar social ideology prevailed, but psychological tests as such were not used in its service. Instead a dictation test was used to prevent entry by anyone judged to be undesirable, a judgment aided considerably by knowledge of the person's racial background (Commonwealth of Australia, 2000). The *Immigration Restriction Act 1901*(Cth), known popularly as the White Australia Policy, sought to maintain racial purity by preventing non-European migration and was part of Australia's culture for the first 50 years or more of the twentieth century.

The dictation test of some fifty words could be administered in any 'prescribed language', but in practice was in a European language and commonly

in English. The text was read to the migrant in the prescribed language and the migrant had to write the text in the same language. Some examples of the content of the dictation test used in 1925 are shown in Figure 1.4. The test could be applied many times and the likelihood of success when it was administered was low. In 1903, for example, 153 people were tested and only three passed. Although its use was directed principally at non-Europeans, it could be used with felons, and those with 'a loathsome or dangerous character'. A German migrant, who had served a prison sentence, was reported to have been given the dictation test in Greek, although he could speak German, English and French. The use of the dictation test as an entry permit to Australia was eventually abolished in 1958 with the introduction of the revised *Migration Act*.

Figure 1.4 Sample passages of the dictation test used in 1925

2

From 1st to 15th September, 1925. No.25/17.

The need for mental stillness, for quiet and balance, is obvious. People are too excited. Let us think how null and void our little revolutionary efforts are when tested by reality. Yet the fruitful results in our private lives and public efforts spring almost always from quiet reflection and mature contemplation.

———————

From 16th to 30th September, 1925. No.25/18.

The tiger is slightly shorter in the leg than the lion, but he is longer in the body. A well-nurtured male tiger weighs nearly a quarter of a ton. Every inch and every ounce of his terrible frame is perfect for the deadly business of the animal's daily life – for speed and certainly in killing.

———————

From 1st to 15th October, 1925. No.25/19.

Water as a liquid concerns us because our lives, like that of other living creatures, whether they be human, animal, or vegetable, from the biggest mammoth to the tiniest microbe, are dependent on water. Therefore, so far as we know, where there is no liquid water, there can be no life.

———————

From 16th to 31st October, 1925. No.25/20.

As nobody had ever been able to discover the actual history of the eel, people sought a miraculous explanation. They knew that the salmon comes up out of the sea to lay its eggs far up the river near its source, but the big eels were never found travelling in the rivers except towards the sea.

———————

Applied psychology had not begun in Australia when the dictation test was introduced, and many members of the profession would hope that if it had been established, the profession would have been a vociferous critic of such unfair testing procedures.

Testing in the computer age

By the 1950s, the major forms of psychological test had been developed for measurement of behavioural differences, and researchers such as Hans Eysenck (1916–1997) and Raymond Cattell (1905–1998) had begun work on developing performance measures of the personality and motivation domains similar to those developed in the cognitive domain. There were new tests published after that date, but they were refinements of the basic methods developed in the first half of the twentieth century. From the 1960s on, however, there were important developments in the use of computer technology to assist in psychological testing and assessment. The earliest use of computers was to reduce labour and the likelihood of error when manually scoring tests by allowing machine scoring of answer forms. Later, desktop computers were used to administer and score tests, and to store large amounts of data on test performance. It was a short step from here to computer interpretation of test results, and programs were written to provide descriptions of individual characteristics based on scores obtained using the tests. Not all psychologists (e.g. Matarazzo, 1986) considered this to be a positive development because of the danger of invalid interpretation in the hands of the novice.

The real power of the computer for psychological testing awaited developments in the theory of tests, and in particular the formulation of item-response theory (IRT; see Hulin, Drasgow & Parsons, 1983). Test developers had recognised from the earliest stages that single items were poor candidates for capturing psychologically interesting constructs, because variation among individuals in responding to them could be determined by a host of factors aside from the one of interest. By aggregating many items, however, the 'noise' associated with individual items could be submerged in the signal that each of them provided. Test theory developed to show why this was so and the implications of it. One implication was that a large number of items were usually required to determine any particular psychological characteristic. This implication was challenged by IRT, which showed how—by specifying in advance a particular statistical model for the test—more precise estimates could be obtained. When this method was linked to the processing speed of the computer, much shorter tests could be produced. Computerised adaptive testing (Weiss, 1983), as it came to be called, provided not only a considerable saving in time and effort for the test administrator but also, importantly, for the client.

As a practical example of the value of this development, consider the case of a young person in the 1950s who wishes to join the armed services. After completing the necessary paperwork, the applicant would need to wait until a group testing session for recruit selection was held (often a matter of months), set aside two to three hours to complete the tests, and then wait to find out if they had been successful. By the 1980s, with the advent of computerised adaptive testing, the potential recruit could attend the recruiting centre, complete the necessary paperwork, take the computer-based test on the spot or at a time of

their choosing, and in half an hour (or less) have the answer as to whether or not they were suitable. Rather than having to answer questions numbering in the hundreds, a dozen to twenty questions are now sufficient to give just as reliable an estimate of their abilities.

A further extension of the role of computing in psychological testing was ushered in by the arrival of the internet, as it became possible to administer tests to individuals at sites remote from the psychologist or test administrator. Although now a relatively simple procedure to implement, the technology raises salient issues for the security of test content and test results, and opens testing procedures to fraud in a way that had not existed previously with individual or group tests. No doubt these problems will be overcome in time, and information technology in all its forms, including its capacity to simulate environments, will push the technology of psychological testing in interesting and useful directions. We shall revisit the topic of computer and psychological testing and assessment in Chapter 14.

Continuing challenges to testing

The controversies and legal battles of the 1960s and 1970s over psychological testing taught the testing community how to accommodate many of the constraints placed on them—which were not always for the most sensible of reasons. The 1980s and 1990s brought fresh challenges for which these earlier accommodations were of no particular value. One challenge was the drive for cost containment in both the private and public sectors, exemplified, for example, in managed care in the USA, but also seen in most Western countries. In the health sector, the drive for cost containment led to a questioning of the time taken to administer and interpret psychological tests and their value for the cost involved. Psychologists had to begin to justify their procedures not in terms of their judgments or the judgments of other professionals as to their value, but in terms of their dollar value. Although there were attempts to do this in the organisational context by showing the dollar savings entailed in good selection practices using psychological tests (e.g. Schmidt et al., 1979; Vinchur, 2014), the task was far more difficult in the health-care context, and the response here was to stop using long tests or to substitute them with short forms of the tests with less validity.

The second challenge came with the increasing use of psychological assessments in determining personal injury and compensation cases in the courts. Psychological assessments and those who prepared them became caught up in the adversarial system that characterises courts that derive from the English legal tradition. Within this system, expert witnesses can expect to have to justify their conclusions quite precisely and to have their opinions attacked by the other

side. With outcomes involving large amounts of money, there is considerable incentive to find fault with testimony based on psychological assessment. Ziskin and Faust (1988) reviewed many of the procedures being used by psychologists and challenged the evidence that supported them. The response in this case was for psychologists to undertake more research to justify the procedures they used, or to discontinue procedures where evidence was lacking for its value, at the now quite high level of expert testimony required. The use of psychological tests and assessment in the legal area is the topic of Chapter 12.

The twentieth century saw a remarkable flowering of psychological tests. A period of sustained enthusiasm in the first half of the century was tempered by waves of public criticism of testing in the second half, but the enterprise was left on a very firm foundation, as the study by Meyer et al. (2001) demonstrated. These authors summarised the data of 125 previous studies on the validity of psychological tests and concluded that the evidence for validity was strong and compelling, and was comparable to that for the validity of medical tests.

Psychological tests: why do we need them?

In the previous section, we briefly reviewed the history of psychological testing. However, we have not directly explained why psychologists believe that psychological tests are better than other methods in assisting individuals to promote better understanding of human behaviour or to make decisions. For humans, the quest to understand ourselves and other people has a long history and the need to make decisions about people is not a new challenge for the human race. Human beings have always been fascinated by their own and others' behaviours. For example: why is this seemingly bright student underperforming in class? Why do I lack confidence in public speaking? Why is my memory not as good as it was 20 years ago? Similarly, every day people in our society are faced with the task of making decisions that are important and have long-term implications for individuals. Examples of such decisions include: Which university course should I pursue? Who should I appoint for this important position in my company? Does my client have a mental disorder? Should this patient return to work after her stroke? Traditionally, we have relied on a number of methods (e.g. tradition, supernatural forces, laws or logic) to assist us in these processes. For example, in ancient China, astrology and numerology were used to evaluate the compatibility between potential brides and grooms.

For the profession of psychology, personal judgment and clinical intuition have been used for a long time to assist psychologists to arrive at a decision or to understand behaviour. For example, psychologists who work in business organisations have made decisions about hiring individuals based on the face-to-face interview. Similarly, clinicians have used interviews to decide if someone is

suffering from mental illness or brain injury. It has been shown repeatedly, however, that human judgment is subjective and fallible (Dahlstrom, 1993; Zimbardo, 2004). Some of the factors that can influence the outcomes of human judgment include stereotyping, personal bias, and positive and negative halo effect. Given that most decisions relating to professional psychology have significant implications for the person involved or the person who made the decision, an error in making the decision can be costly and devastating, and might not be reversible. For example, an erroneous judgment about the mental competency of a person can lead to the rights of the person being wrongfully removed. As another example, a lot of time and money could be wasted if the wrong person was hired for a job. Psychologists consider psychological tests better than personal judgment in informing decision making in many situations because of the nature and defining characteristics of these tests (Dahlstrom, 1993).

Psychological tests: definitions, advantages and limitations

In this section, we define psychological tests and discuss their advantages and limitations.

Definitions and advantages

What is a psychological test? This seems to be a difficult question to answer when one examines the plethora of published tests in the market and finds that they can differ in so many respects. While some psychological tests take only a few minutes to complete, others can take hours to administer. For some psychological tests, a respondent is required to provide only a simple yes/no answer; other tests are designed in such a way that a person has to navigate and respond in a virtual reality environment. Some psychological tests can be administered to hundreds of people at one time, and scored and interpreted by a computer, but other tests require face-to-face administration and individual scoring and interpretation that require years of training and experience.

Despite the above wide-ranging differences, all psychological tests are considered to have one thing in common; that is, they are tools that psychologists use to collect data about people (Groth-Marnat, 2009; Suhr, 2015). More specifically, a psychological test is an objective procedure for sampling and quantifying human behaviour to make an inference about a particular psychological construct using standardised stimuli and methods of administration and scoring. In addition, to demonstrate its usefulness a

psychological test requires appropriate norms and evidence (i.e. psychometric properties). To elaborate, the defining characteristics of psychological tests and their associated advantages are discussed below.

First, a psychological test is a sample of behaviour that is used to make inferences about the individual in a significant social context. The behaviour sample might be considered complete in itself or, as is more often the case, as a sign of an underlying disposition that mediates behaviour. Take, for example, a psychological test that is used to decide whether an individual will be able to understand instructional material to be used in job training. The test for this purpose might consist of sample passages from the daily newspaper. The test taker's task is to read each of the passages and report their meaning. If comprehension of most of the passages is accurate, the test taker can be judged to read well enough for the purposes of the job. As long as the difficulty level of the passages approximates that of the instructional material, the test provides a basis for inferring adequate performance in training.

In a clinical setting, a test might provide a sample of the behaviour that the client finds disturbing. For example, a client might suffer an irrational fear of objects that are not actually dangerous, such as harmless spiders. As a result of the fear, the client cannot enter a darkened room or clean out cupboards because of the likelihood of confronting a spider. To assess the magnitude of the irrational fear, the tester might ask the client to approach a harmless spider being held in a glass case. The distance from the spider that induces a report of anxiety is taken as an indication of the severity of the client's avoidance behaviour. This can be used to judge the effectiveness of any subsequent planned intervention to reduce the problem. After treatment the client should be able to approach the spider more closely than before.

In both of these cases, the sample of behaviour is complete in itself, as it assesses directly what the tester wants to know; namely, comprehending common passages of English text or avoiding an object of a phobia. The samples could be used, however, as the basis for indirect inferences, by arguing that each in its own way reflects an underlying disposition that is responsible for the individual's behaviour. Thus, the comprehension test might be used to infer the individual's level of general mental ability or intelligence, and the avoidance test could be used to infer the individual's level of neuroticism; that is, the likelihood that they will suffer an anxiety disorder. In these cases, the content of the particular sample is incidental and can be replaced by a different sample that is also thought to reflect the disposition. Thus, a sample of mathematical problem solving could be substituted for the test of verbal comprehension as a sign of general mental ability, or a set of questions about episodes of anxiety and depression could be substituted for the avoidance test as a sign of the individual's level of neuroticism. Such substitution would make no sense if the tests were being used as a sample rather than a sign.

The distinction between tests as samples of behaviour or as signs of an underlying disposition rests on theoretical differences about the causes of human behaviour. Important as these theoretical differences are, they are outside the scope of the present book. We draw attention to the distinction here for two reasons. First, it is important for the tester to be aware whether any particular test is being used principally as a sample of behaviour or as a sign of an underlying disposition. We say 'principally' because the distinction when probed is not hard and fast.

The other reason for making the distinction is that tests used in these two ways are interpreted differently. Where the test is a sample, interpretation of test performance is usually in terms of what has been called 'criterion referencing'; however, where the test is used as a sign, what is termed a 'norm referencing' strategy is usually adopted. In the case of the former, what is effective behaviour in the situation in question can be specified reasonably objectively and the individual's performance judged against this standard or criterion. Thus, a person might be expected to understand most, if not all, of what they read in a newspaper if they are to deal with instructional manuals on the job. A person free of a spider phobia can be expected to come close to a harmless spider, but perhaps not touch it. In the case of norm referencing, on the other hand, the performance of the individual is related to the performance of a group of individuals similar to the test taker in important respects (e.g. age, gender, educational level and cultural background). How well or badly a person has performed is thus assessed against what the average person can do, or what the norm is. Many psychological tests are thought of as signs of underlying dispositions and as such are norm referenced. The distinction is encountered again in Chapter 3.

The second characteristic of a psychological test, similar to other scientific measurement instruments, is that it is an **objective procedure**. It uses the same standardised materials, administration instructions, time limits and scoring procedures for all test takers. This ensures that there is no bias, unintended or otherwise, in collecting the information and that meaningful comparisons can be made between individuals who are administered the same psychological test. Unless two people are treated in the same way (e.g. same instructions, same order of questions and same time limits), it is not possible to attribute any differences in their performance to differences between them. The difference in performance could just as well be due to the difference in the ways they were tested. To ensure uniformity of test stimuli and procedures, the manual that accompanies a psychological test usually includes detailed and clear instructions for administering the test so that the same or similar score will be obtained even when the test is administered by different testers or in a different setting. The objective nature of psychological tests is one of the main advantages they have over other methods for assisting us to understand human behaviour and make

decisions about it, not least because it minimises errors of judgment relating to personal bias or subjectivity (Dahlstrom, 1993). The objective nature of psychological tests is discussed again in Chapter 2 when we explain the process and best practices in psychological testing.

objective procedure
the use of the same standardised materials, administration instructions, time limits and scoring procedures for all test takers

Third, unlike subjective human judgment, the result of a psychological test is summarised quantitatively in terms of a score or scores. Again, this characteristic is similar to that of other scientific measurement instruments that use numbers to represent the extent of variables such as weight, temperature and velocity. The quantification of psychological test results allows human behaviour to be described more precisely and to be communicated more clearly. For example, the use of an IQ score allows psychologists to provide a more fine-grained description of a person's intellectual ability. We visit the topic of psychological test scores in Chapter 3.

Fourth, a psychological test provides an objective reference point for evaluating the behaviour it measures. In the case of a **criterion-referenced test**, a standard of performance is determined in advance by some empirical method, and the test taker's performance is compared with this standard in determining whether they pass or fail. It might be, for example, the judgment of experts that determines the standard, but it is open to all to see what the standard is that is being set. It is not the personal viewpoint of the person collecting the information. In the same way, in a **norm-referenced test** the performance of a representative group of people on the test is used in preparing the test norms, and these are used in scoring and interpreting the test. The individual's performance is thus referred to that of the norming group, a reference point that is not an individual's judgment. In both cases, the psychological test allows the comparison between the individual in question and some sort of standard performance.

criterion-referenced test
a psychological test that uses a predetermined empirical standard as an objective reference point for evaluating the performance of a test taker

norm-referenced test
a psychological test that uses the performance of a representative group of people (i.e. the norm) on the test for evaluating the performance of a test taker

Fifth, possibly the most important defining characteristics of a psychological test is that it must meet a number of criteria to be a useful information-gathering device. The criteria relate to its quality as a measuring device; for example, how accurate and reproducible the scores obtained with it are, or how well it measures what it intends to measure. These criteria are referred to as **psychometric properties**. They are evaluated in the course of test construction and again subsequently, and are reported or made available to test users. This is in fact a process of quality control to ensure that the test is operating in the way the authors claim it does (these criteria are described and discussed in depth in Chapters 4 and 5). By showing that the psychometric properties of a psychological test have reached a required standard, we can have confidence in using the results obtained from this test.

psychometric properties
the criteria that a psychological test has to fulfil in order to be useful; they include how accurate and reproducible the test scores are, and how well the test measures what it intends to measure

Limitations

Although it is important to know that psychological tests have a number of advantages, it is also necessary to be aware of the limitations of tests. Not knowing these limitations can lead to an over-reliance on, or misunderstanding of, the psychological test results obtained.

The first of these limitations, as mentioned earlier, is that psychological tests are only tools. As such, they do not and cannot make decisions for test users. Decision making is the responsibility of the person who requested the use of the test and to whom the test results are made available. The person might be the psychologist who administered the test, but the two roles should not be confused. The test provides a way of gathering information and, if well chosen, will provide accurate and pertinent information, but the use of the information, including a bad decision, is in the hands of the decision maker. Not being aware of this limitation can lead the test user and the person involved to be dependent on the test results and accept them passively. Instead, psychological test results should be used as a source of data, along with other sources of data such as personal history and current circumstances, to assist the test user or the individual to arrive at or make an informed decision.

Second, psychological tests are often used in an attempt to capture the effects of hypothetical constructs. As in other scientific disciplines, psychology employs constructs that are not directly observable; rather their effects can only be inferred. As such, we need to be aware that sometimes a gap exists between what

the psychologist intends to measure using a psychological test and what a test actually measures. For example, although IQ tests were developed to measure intelligence, we need to be aware that the value of these tests in telling us about a person's intelligence depends very much on our understanding of the construct of intelligence and the type(s) of behaviours included in any particular test. Not being aware of this issue can lead to the development of unwarranted faith in psychological tests and total acceptance of the test results without being aware of their limitations.

Third, because of the continual development or refinement of psychological theories, the development of technology and the passage of time, psychological tests can become obsolete (i.e. **test obsolescence**). They might no longer be suitable for use because the theory that their construction was based on has been shown to be wrong or because the content of the items is no longer appropriate because of social or cultural change. In the early part of the twentieth century, for example, church attendance in Western countries was very much higher than it is now and a reasonable level of Bible knowledge could be assumed. A test item might draw on this fact. Although useful then, it might be far too esoteric to be of any use today. According to the Australian Psychological Society and the American Psychological Association, tests should be revised or updated regularly and normative samples should be kept current.

test obsolescence
the notion that a psychological test loses its utility because the theory that it was based on has been shown to be wrong, or because the content of its items is no longer appropriate because of social or cultural change

Finally, although it might not be the intention of a test developer, sometimes a psychological test can disadvantage a subgroup of test takers because of its cultural experience or language background. A vocabulary test that usefully discriminates levels of verbal ability among children from white, English-speaking, middle-class homes might be of no use for this purpose with children with a different subcultural experience or those who do not have English as their first language. Tests are not universally applicable and to treat them as such can do an injustice to some, but more of this in Chapter 2.

Chapter summary

In this first chapter, we have provided a brief introduction to the history of psychological testing. In addition, we have defined what a psychological test is and discussed its characteristics, advantages and limitations. In so doing, we trust you will start to appreciate why psychological tests were developed and how they

have been (and can be) used to assist individuals in our society to promote better understanding of human behaviour and to make decisions.

Questions

1. From the section on the history of psychological testing, select three developments in psychological testing and discuss why you think they have made a significant impact on our lives.
2. Select an Australian psychologist mentioned in the 'A brief history of psychological testing' section and:
   - **a.** write a short biography of this person
   - **b.** discuss his or her contribution to the field of psychological testing.
3. What are some of the ways that psychological tests have been used to assist individuals in promoting understanding and making decisions?
4. What are the five defining characteristics of a psychological test?
5. The advantages of a psychological test outweigh its limitations. Discuss.
6. Some tests (e.g. Am I a moody individual? How is your marital relationship?) in popular magazines look like but are not psychological tests. Why not?

Further reading

Dahlstrom, W G (1993). Tests: Small samples, large consequences. *American Psychologist*, 48, 393–9.
Meyer, G J, Finn, S E, Eyd, L D, Kay, G G, Moreland, K L, Dies, R R, et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, 56, 128–65.
Weiner, I B (2013). Assessment psychology. In D K Freedheim (Ed.), *Handbook of Psychology: Vol. 1 History of Psychology* (pp. 314–39). Hoboken, NJ: John Wiley & Sons.

Useful websites

Testing and assessment (American Psychological Association):
www.apa.org/science/programs/testing/index.aspx
Psychological testing (Australian Psychological Society):
www.psychology.org.au/community/topics/psych_testing/FAQs

2 Psychological Testing and Assessment: Processes, Best Practice and Ethics

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. define psychological testing and psychological assessment, and distinguish between the two
2. list the main areas where psychological testing and psychological assessment have been applied
3. list the different types of psychological tests
4. explain the main processes of psychological testing and psychological assessment
5. discuss issues relating to the best practice and ethics of psychological testing and psychological assessment
6. discuss issues in testing and assessing individuals from different ethnic backgrounds.

KEY TERMS

clinical neuropsychologist
clinical psychologist
culture fair test
educational and developmental psychologist
ethics
forensic psychologist
organisational psychologist
performance test
psychological assessment
psychological testing
self-report test

Setting the scene

- A member of the general public telephoned a psychologist because she wanted to take a psychological test to find out her IQ. After discussing her request and situation, the psychologist suggested that she needed psychological assessment (including psychological testing) to clarify her vocational interest and career goals.
- A student who graduated with an undergraduate degree majoring in psychology wanted to purchase a personality test from a publisher. Her request was refused because she did not meet the user qualification requirement for that particular test.
- The Australian Health Practitioner Regulation Agency received a complaint from a client of a psychologist who claimed that the psychologist had not provided him with a written copy of a psychological testing report.
- The psychological test librarian of a university department received a request from a member of the general public who wanted to borrow some psychological tests. The reason for the request was to prepare for a job interview. The librarian explained to the requestor that he did not fulfil the requirements for being a user. The request was turned down.
- A psychologist received a request from the personnel officer of a company. The officer wanted to obtain a copy of a psychological test report for a former client of the psychologist to assist with decision making.

Introduction

Selecting appropriate psychological tests, administering them, scoring and interpreting test results, and conducting psychological assessments are core skills of professional psychologists. Interestingly, these are common areas of complaint against psychologists lodged with registration authorities or agencies in Australia and overseas. To improve the standard of practice in psychological testing and assessment, there is a need for better education about the nature of psychological tests and the steps and processes of psychological testing and assessment. In addition, psychologists need to be aware of the ethical principles and professional guidelines relating to best practice in this area.

Is psychological testing the same as psychological assessment? What are some of the main areas in professional psychology where psychological testing and assessment have been applied? What are the steps, processes and best practices in psychological testing? What are some of the ethical issues that psychologists have to pay attention to when conducting psychological testing and assessment? What do we need to consider when testing and assessing individuals from different cultures? These are some of the questions that we aim to answer in this chapter.

Psychological testing versus psychological assessment

A distinction is made between psychological testing and psychological assessment (Groth-Marnat, 2009; Matarazzo, 1990; Suhr, 2015). When we talk of **psychological testing** we are referring to the process of administering a psychological test and obtaining and interpreting the test scores. On the other hand, when we talk of **psychological assessment** we are referring to a process that is broader in scope. Whereas psychological testing is commonly undertaken to answer relatively straightforward questions such as 'What is the IQ of a child?' or 'What is the vocational interest of a job applicant?', psychological assessment is usually required to deal with more complex problems such as 'Why is a child experiencing study problems at school?' or 'Should an applicant be appointed to a job vacancy?'

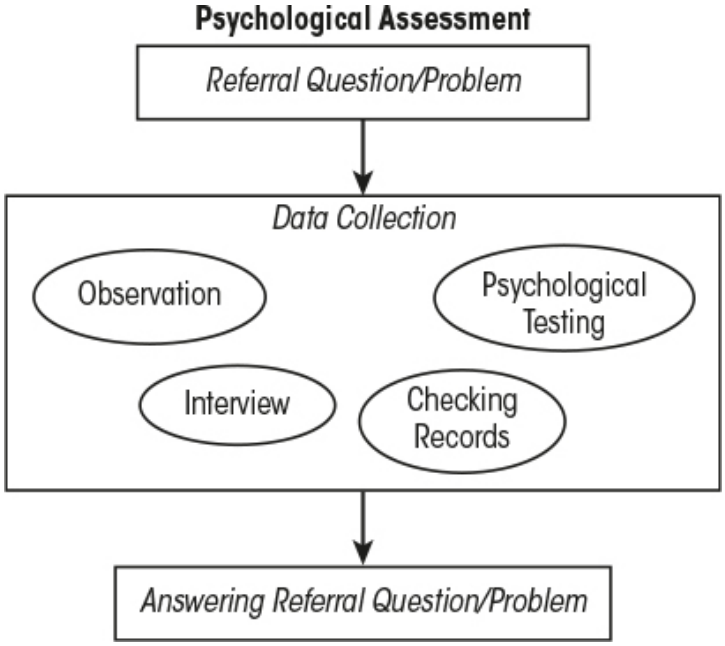
psychological testing
the process of administering a psychological test, and obtaining and interpreting the test scores

psychological assessment
a broad process of answering referral questions, which includes but is not limited to psychological testing

Maloney and Ward (1976) defined psychological assessment as 'a process of solving problems (answering questions) in which psychological tests are often used as *one* of the methods of collecting relevant data' (p. 5). According to Suhr (2015), psychological assessment 'is a conceptual, problem-solving process of gathering dependable, relevant information about an individual in order to make an informed decision' (p. 2). Thus, to answer the referral problem/question, 'Why is a child experiencing study problems at school?', a psychologist will usually administer an intelligence test such as the Stanford-Binet Intelligence Scale–Fifth Edition (Roid, 2003) or the Wechsler Scale of Intelligence for Children–Fifth Edition (WISC–V; Wechsler, 2014). However, the psychologist will also use other data-collection techniques (e.g. interviewing the child's parents and teachers, and observing the child in class) to obtain other relevant information such as medical, family, developmental and educational history to help answer the question. The importance of this distinction is that it emphasises that psychological testing forms only a part of psychological assessment and that best practice in assessment must take into account other sources of information (see

Figure 2.1). In a properly conducted assessment, conclusions drawn by the psychologist are based on data obtained from all these sources. In the hands of a skilled psychologist, scores on psychological tests are not seen as some immutable quantity possessed by the person tested, but rather as data bearing on hypotheses that need to be tested before being (provisionally) accepted. Of course, in many cases these various sources of information all point in the same direction; for example, that the person in question has some type of mental impairment, or that they are eminently suited for a particular job.

Figure 2.1 Relationship between psychological assessment and psychological testing



Areas of application

The discipline of psychology comprises both research and applied areas (Gazzaniga & Heatherton, 2003). Although psychological tests are used by research psychologists in their projects, most individuals in the general community encounter these tests through the practice of professional psychologists. In Australia and in other parts of the world, psychological testing and assessment are most commonly applied in the following branches of psychology: clinical, organisational, clinical neuropsychology, forensic, educational and developmental.

Clinical psychologists specialise in the assessment, diagnosis, treatment and prevention of psychological and mental health problems (sample referral problem: 'Is the client clinically depressed?'). **Organisational psychologists** are specialists in the areas of work, human resource management, and organisational

training and development (sample referral problem: 'Is this applicant suitable for a high level managerial position in the tourism industry?'). **Clinical neuropsychologists** are concerned with the effects of brain injury on human behaviour and provide diagnosis, assessment, counselling and intervention for these individuals (sample referral problem: 'What are the brain functions affected by this patient's stroke?'). **Forensic psychologists** are concerned with the legal and criminal justice areas and provide services for perpetrators or victims of crime and personnel of the courts and correctional systems (sample referral problem: 'What is the risk of this inmate reoffending?'). The provision of assessment, intervention and counselling services to children and adults with learning and developmental needs are the domains of **educational and developmental psychologists** (sample referral problem: 'Does this 6-year-old boy have attention-deficit hyperactivity disorder?'). In the fourth part of this book, a chapter is devoted to each of these five areas of professional psychology, with a discussion of the psychological tests and assessment procedures commonly used in each. Other branches of psychology that also utilise psychological tests and practise psychological assessment (but are not included here because of space reasons) include counselling psychology, career psychology, community psychology, health psychology, and sport and exercise psychology.

clinical psychologist
a psychologist who specialises in the diagnosis, assessment, treatment and prevention of psychological and mental health problems

organisational psychologist
a psychologist who specialises in the area of work, human resource management and organisational training and development

clinical neuropsychologist
a psychologist who specialises in understanding, assessing and treating individuals' cognitive and behavioural impairments resulting from brain injury

forensic psychologist
a psychologist who specialises in the provision of psychological services relating to the legal and criminal justice areas

educational and developmental psychologist
a psychologist who specialises in assessing and treating children and adults with

learning and developmental needs

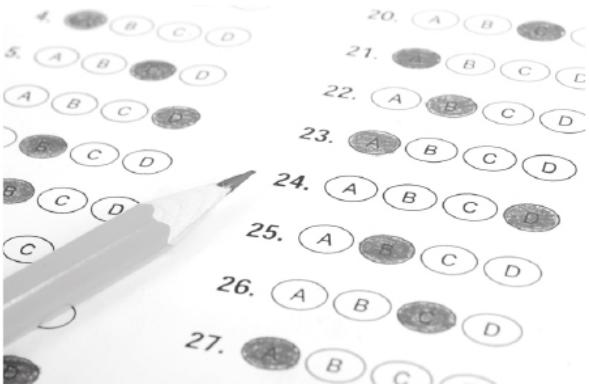
Types of psychological tests

As mentioned in Chapter 1, although all psychological tests share some common characteristics, published tests on the market differ in a number of ways. First, they differ in terms of the type of responses required from the test taker. The most common distinction is between **self-report tests** and **performance tests**. For example, the Minnesota Multiphasic Personality Inventory–2 (MMPI–2; Butcher et al., 1989) simply requires a test taker to indicate, by marking a box (yes or no), whether or not each written statement in the inventory is an appropriate description of their behaviour or experience. The Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV; Wechsler, 2008), on the other hand, requires the test taker to answer questions or solve problems, in some cases by manipulating the test materials provided. Self-report tests have practical advantages in that they usually take less time to complete and can be given to a number of people at the one time. Performance tests are usually limited to individual administration, but they provide information about what the person can actually do as distinct from what they say they can do. In practice, the two formats are used in assessing different psychological constructs. Self-report tests are most common when the interest is in typical behaviour—what the person frequently does, as in the case of personality and attitude. Performance tests, on the other hand, are used in assessing the limits of what a person can do, such as in assessing their aptitudes or abilities.

self-report test
a psychological test that requires test takers to report their behaviour or experience; these tests can be administered individually or in a group

performance test
a psychological test that requires test takers to respond by answering questions or solving problems; they are usually administered individually

Figure 2.2 Psychological testing: self-report versus performance



Second, psychological tests differ in terms of the number of individuals who can be administered the tests. The distinction is between individual versus group administration. For example, the WAIS–IV (Wechsler, 2008) is a test of intelligence that can only be administered to one person face-to-face, whereas the Raven's Progressive Matrices (Raven, 1938) is a test of non-verbal general ability that can be administered individually or to a group. Although the group-administered tests are usually more economical to administer and score, the individually administered tests allow psychologists to observe the performance of the person tested and to follow up and clarify the answers if needed.

Third, with the development of the personal computer, tests can differ in terms of whether or not a computer is used in administration, scoring and interpretation. The distinction is between human- and computer-assisted psychological testing. The National Adult Reading Test (NART; Nelson & Willison, 1991), for example, was designed to be administered, scored and interpreted by a person experienced in the use of the test. Other tests have been designed or redesigned to take advantage of computer assistance with one or more of these processes. For example, computer programs (e.g. Scoring Assistant and Report Writer) have been developed to score and interpret the performance of a person on the Wechsler Memory Scale–Fourth Edition (WMS–IV; Wechsler, 2009b). The development, practice and advantages/limitations of computer-assisted psychological testing are discussed in detail in Chapter 14.

Box 2.1

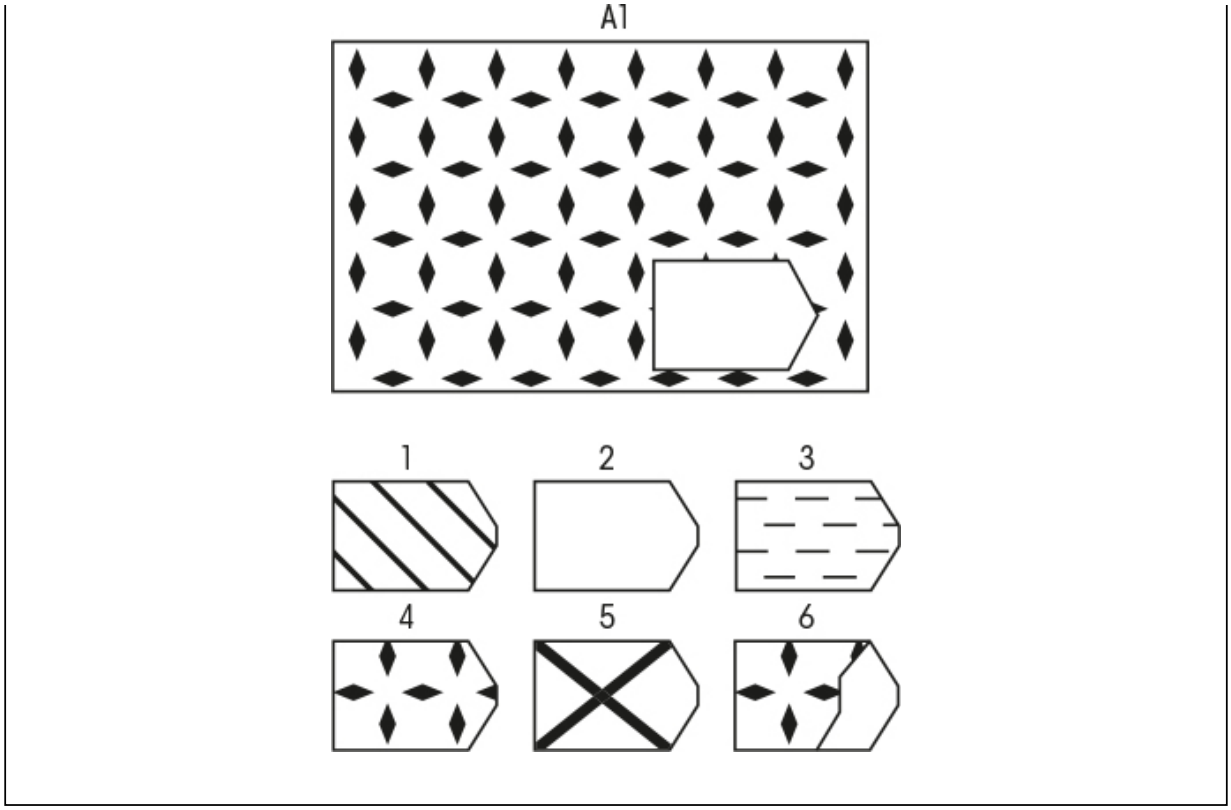
Raven's Progressive Matrices

The Raven's Progressive Matrices was invented in the 1930s by J C Raven (Penrose & Raven, 1936; Raven, 1938). Seeking a pure measure of Spearman's g, Raven devised a non-verbal reasoning test made up of two-dimensional figural

analogies. In undertaking this test, a person is required to select a picture that completes the missing element of a pattern (see Figure 2.3). Indeed, the progressive matrices are now widely recognised as the best single measure of general intelligence (Carpenter, Just & Shell, 1990) and are often used as a marker for fluid intelligence.

The progressive matrices come in three forms: a set of Coloured Progressive Matrices for children, first introduced in 1947 and revised in 1956; a Standard set for children aged six to adult, revised in 1948 and again in 1956; and an Advanced set for higher ability populations such as university students and professionals, also introduced in 1947 and revised in 1962. The Standard Progressive Matrices are composed of sixty items arranged in five sets of increasing difficulty, beginning with very easy items designed to be fairly self-evident and progressing through items drawing on various perceptual relations until reasoning by analogy in one and two dimensions is required. The progressive nature of the items means that working through them also provides training in the thought processes required for their solution. Administration can be limited to 20 minutes, or untimed, and can be given to either individuals or groups, yielding a very flexible test suitable to a wide range of applications. Its excellent psychometric properties and ease of administration have led to its extensive use in education, industrial and military settings. Its non-verbal nature has also meant that it is not subject to the same cultural influences as other tests (Jensen, 1980). As such, it has been used extensively in cross-cultural research and settings involving examinees from non-English-speaking backgrounds.

Figure 2.3 An example of a Raven's Progressive Matrices item



Finally, psychological tests can differ in terms of the frame of reference for comparing the performance of an individual on the test. This distinction is commonly called norm-referenced versus criterion- (or domain-) referenced testing. As mentioned in Chapter 1, while the former compares an individual's performance on a test with the average performance of a group of individuals (called the norming or standardisation sample), the latter compares the individual's performance with a set of *a priori* criteria of adequate or good performance. For example, the score of a person on the Symbol Digit Modality Test (SDMT; Smith, 1982), a test of attention, is interpreted by comparing it with the average score of a group of individuals (previously tested) who are similar in age and educational level. The comparison group provides an appropriate norm for describing whether the person's score is above or below average. In contrast, the performance of the same individual on the Bader Reading and Language Inventory (Bader, 1998) is interpreted based on a set of objectively specified criteria (e.g. a graded word list and graded reading passages). Most of the psychological tests developed and available commercially are norm-referenced tests. Criterion-referenced tests are more likely to be found in educational settings for assessing learning outcomes.

Processes and best practices in psychological testing

In this section, we first describe all the processes involved in psychological testing and then discuss its best practices.

Determining whether psychological testing is needed for a client

Although psychologists who conduct psychological assessments usually use psychological tests as one of their assessment techniques, this is not necessary or possible for every client. For example, a client who is referred to a psychologist might have been tested recently by other professionals or by another psychologist. Consequently, it is not necessary to repeat the testing process. As another example, some clients might refuse to undertake a psychological test because they are concerned about the potential negative impact of the test results. In addition, it should be reiterated that psychological tests are only one of the techniques of psychological assessment and the use of these tests might not be necessary for every client who needs assessment. According to Kendall et al. (1997), the skill to determine whether a client needs psychological testing is one of the characteristics of a proficient user of psychological tests. To develop this competence, a psychologist needs to be familiar with the major psychological constructs commonly assessed (e.g. psychopathology, intelligence, personality, memory and stress) and be aware of the advantages and limitations of using psychological tests (Psychology Board of Australia, 2016a; Suhr, 2015).

Selection of appropriate and technically sound psychological tests

After deciding that psychological testing is necessary for a client and settling on the particular construct or constructs to be assessed, a psychologist needs to select the most appropriate and psychometrically sound tests from the large number of instruments available in the literature and from test suppliers (Groth-Marnat, 2009; Psychology Board of Australia, 2016a). Psychometrics, as the name implies, is concerned with psychological measurement and the theories that underpin it. Part 2 of this book (Chapters 3 to 6) introduces you to the principles of psychometrics commonly employed in testing and assessment. Selecting psychometrically sound tests is a very important step because the quality and

soundness of the results and findings of a psychological assessment depend very much on this selection (Suhr, 2015). Careful consideration during this step also enables the psychologist to subsequently explain, justify and defend their choice of tests. The skills to select appropriate instruments are also considered by Kendall et al. (1997) as essential in being a competent user of psychological tests.

There are a number of resources available to a psychologist to assist with test selection. First, to find out what tests have been published, psychologists can peruse the catalogue of major publishers of psychological tests and references such as *Tests in Print IX* (Anderson et al., 2016) and *Tests* (Maddox, 2008). Table 2.1 shows a list of major publishers of psychological tests in Australia and their corresponding addresses and websites. These catalogues provide psychologists with information about which tests are available for use with which constructs; the purpose, content, length and price of a test; and other pertinent information. In Australia and overseas, test publishers usually require test purchasers to register before they are allowed to buy tests. The purpose of registration is to ensure that confidential test materials are supplied to professionals who are appropriately trained and qualified. For example, Table 2.2 shows the different user levels used by Pearson Clinical Assessment, Australia and New Zealand to restrict and regulate the supply of test materials. In the test catalogue supplied by Pearson Clinical Assessment, the test user level is clearly specified for each test listed so that potential test buyers can determine if they meet the requirement for purchasing that test.

Table 2.1: Name, address and website of major test suppliers in Australia

Name	Address	Website
Australian Council for Educational Research	19 Prospect Hill Road, Camberwell, VIC 3124	www.acer.edu.au
CPP Asia Pacific	Level 7, 369 Royal Parade, Parkville, VIC 3052	www.cppasiapacific.com
Psychological Assessments Australia	Suite 2, 96–100 Railway Crescent, Jannali, NSW 2226	www.psychassessments.com.au
Pearson Clinical Assessment (Australia and New Zealand)	Suite 1001, Level 10, 151 Castlereagh Street, Sydney, NSW 2000	www.pearsonclinical.com.au

Note: addresses and websites are accurate at time of publication.

Table 2.2: User levels used by Pearson Clinical Assessment (Australia and New Zealand) in supplying test materials

User level	Profession	Products that can be accessed*
C	Registered psychologist	A, B, C, T or HR
S	Speech pathologist	A, B, S, T or HR
B	Allied health or special education professional	A, B, T or HR
M	Medical practitioner	A or M
HR	Human resources professional	A or HR
P	Exercise physiologist and podiatrist	P or A
T	Teacher, social worker, nurse or early childhood professional	A or T
A	No qualifications necessary	A only

*Note: A registered psychologist who has a C user level can access tests that require no qualification to administer (A), along with tests developed for use by allied health or special education professionals (B), registered psychologists (C), teachers, social workers, nurses or early childhood professionals (T), and human resources professionals (HR).

Other resources such as *Tests in Print* and *Tests* are bibliographic encyclopaedias that summarise all the commercially published tests in terms of test title, intended population, publication date, acronyms, author(s), publishers, administration time, cost, foreign adaptations and references. However, it should be noted that most of the test publishers do not include critical reviews of psychological tests in their catalogues. This is because the test catalogues are designed to promote and sell tests rather than evaluate them according to scientific principles.

To obtain information about the strengths and weakness of psychological tests, a psychologist needs to turn to other sources. These include the manuals of the test under consideration, specialised test review volumes (e.g. *Test Critiques*), journals (e.g. *Assessment*, *Psychological Assessment*, *Educational and Psychological Measurement*, *Journal of Personality Assessment*, *Journal of Psychoeducational Assessment* and *Journal of Educational Measurement*) and colleagues or supervisors who are experienced in assessment in that particular area.

A psychologist can locate most of the technical information (e.g. reliability, validity, standardisation sample and norms) about a psychological test in its manual. Although professional societies have developed guidelines—for example, *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014)—regarding the kind of technical information to be included in a test manual, test developers do not always follow these guidelines. Users of psychological tests need to be wary, particularly if only a small amount of technical information can be found in the manual of a test.

Compendiums of specialist test reviews, such as the *Mental Measurements Yearbook* (see Box 2.2), provide comprehensive reviews of psychological tests. Because the reviews are written independently of the test authors and publishers, and are evaluated using a set format and carefully developed criteria, they are generally objective and critical. Technical information about psychological tests can also be found in journals that specialise in this area (e.g. *Psychological Assessment*). Compared with the specialist volumes, journal articles are less systematic in format and length, but often more up to date. Finally, colleagues and supervisors can also be a source of advice about what tests to use for a particular client.

Box 2.2

Mental Measurements Yearbook

The ***Mental Measurements Yearbook (MMY)*** is one of the oldest and most authoritative sources for test reviews. Published by the Buros Institute of Mental Measurements in Nebraska, USA, the first edition was issued in 1938 and was edited by Oscar Buros. The volumes in the MMY series are produced every three years. The latest edition, the nineteenth, was published in 2014 and edited by Janet Carlson, Kurt Geisinger and Jessica Jonson. To be included in the ***MMY***, a psychological test needs to be new or revised since the publication of the previous edition of the ***MMY***. In addition, the publisher of the test needs to be willing to provide documentation that supports the technical properties of the test. Since the publication of the fourteenth edition, psychological tests must also include sufficient documentation supporting their technical quality to meet the criteria for review. For each of the tests included, one or more reviews are provided by qualified psychologists. Each review comprises the following five sections:

- description (purpose and intended use of the test; target populations; and information on administration, scoring and scores)
- development (theoretical base, assumption and construct of the test; and details on item development, evaluation and selection)
- technical (standardisation sample and norms, reliability and validity)
- commentary (strengths and weaknesses of the test; adequacy of the theory; and assumption and construct of the test)
- summary (conclusions and recommendations).

In recent years, the Buros Institute of Mental Measurements has introduced a web-based service called Test Reviews Online (www.unl.edu/buros). This website contains all current test reviews that have been published in the *MMY* since its ninth edition. For a fee (US$15), users can download individual reviews for more than 2500 psychological tests.

Below is an example of a test review from the fifteenth *MMY* (Shum, 2003).

Learning Style Inventory, Version 3

Purpose: Designed to describe the ways an individual learns and deals with day-to-day situations.

Population: Ages 18–60. *Publication Dates:* 1976–2000.

Acronym: LSI3.

Scores: Four scores: Concrete Experience, Active Experimentation, Reflective Observation, Abstract Conceptualisation; Four learning styles: Accommodating, Diverging, Converging, Assimilating.

Administration: Group or individual.

Price Data, 2001: $79 per 10 self-scoring booklets; $50 per facilitator's guide to learning (2000, 81 pages); $38 per 15 transparencies; also available online at $15 per person.

Foreign Language Editions: French and Spanish versions available.

Time: (20–30) minutes.

Author: David A Kolb.

Publisher: Hay Group.

Cross References: See T5:1469 (13 references) and T4:1438 (12 references); for a review of an earlier edition, see 10:173 (17 references); see also 9:607 (7 references).

Review of the Learning Style Inventory, Version 3 by DAVID SHUM, Senior Lecturer of Psychology, Griffith University, Brisbane, Australia.

Description

The Learning Style Inventory, Version 3 (LSI3) is a self-report twelve-item test developed by David A Kolb to help people describe how they learn and to identify their learning style. The test describes a person's learning mode according to two polar dimensions: Concrete Experience (CE) versus Abstract Conceptualisation (AC), and Active Experimentation (AE) versus Reflective Observation (RO). Based on these descriptions, the person's learning style is classified into one of four basic types: Diverging, Assimilating, Converging or Accommodating.

The LSI3 is suitable for people between 18 and 60 years old with a seventh grade reading level or above. No special requirements for the administration, scoring and interpretation of the test are specified. According to Kolb, the main applications of the LSI3 are self-exploration, self-understanding and self-development.

The LSI3 was designed in such a way that it can be administered, scored and interpreted by the test taker. One is required to complete 12 sentences that describe learning by ranking four endings (from 4 to 1 for best description to worst description) that correspond to the four learning modes (CE, AC, AE and RO). The 12 sentences are written in easily understood language and printed on a two-part (answer and score) form. The instructions for the test are well organised and clearly written. The scores for the four learning modes can range from 12 to 48. Given the way the sentences are answered and scored, these scores are ipsative in nature.

To find one's preferred learning mode, a diagram called the Cycle of Learning is used to transform the four raw scores into percentile scores based on a normative group of 1446 adults. Two combined scores are also obtained by calculating the differences, AC – CE and AE – RO. Finally, one's preferred learning style type is determined by plotting the two difference scores on a Learning Style Type Grid.

Development

Kolb originally developed the LSI in 1971 based on Experiential Learning Theory, which in turn is based on the Jungian concept of styles or types. The LSI3 is the latest revision of the inventory and there are four main changes. First, in LSI2, the endings that represent the four learning modes were organised in the same order for all 12 sentences to facilitate scoring. To control for possible response bias, the order of the endings is randomised in the LSI3. Second, Kolb modified the wordings of the learning style type in the LSI3 (e.g. Converger to Converging) to address the concern that the old terms might give an impression that learning styles do not change. Third, the response sheet for the LSI3 was changed to a two-part colour-coded form and it is produced in such a way that answers written on the first page are automatically transferred

to the second page. Fourth, a number of experiential activities and information on career development have been added to a 19-page test booklet. An 81-page *Facilitator's Guide to Learning* was published in 2000 to accompany the test.

Technical

The technical specifications of the LSI3 are included as a six-page section in the *Facilitator's Guide to Learning*. The normative group for the LSI3 comprised 1446 adults aged between 18 and 60 years old. According to the guide, there were 638 males and 801 females, which for reasons not explained do not total 1446. Kolb states that this group was ethnically diverse, represented a wide range of career fields, and had an average education of two years of college. However, detailed description and breakdown of these demographic variables are not available. The percentile scores for all test takers are based on the average performance of this group. Separate norms for different age and gender groups are not provided. This is a concern because there seem to be age and gender differences on some of the scores (see p. 10 and p. 68 of the guide).

Evidence for the internal consistency and test-retest reliability of the LSI scores is based on the data (initial sample N = 711, replication sample N = 1052) collected by Veres, Sims and Locklear (1991) using a version with randomised sentence endings. Mean coefficient alphas for the four learning modes scores ranged from 0.53 to 0.71 in the initial sample and from 0.58 to 0.74 in the replication sample. These indices are lower than expected and they are lower than those obtained for the LSI2 (from 0.82 to 0.85). The test-retest (eight-week interval) reliabilities of the four learning modes scores ranged from 0.92 to 0.97 in the initial sample and from 0.97 to 0.99 in the replication sample. Similar statistics obtained for the LSI2 were much lower and ranged from 0.25 to 0.56. Kappa coefficients were also calculated to examine classification stability for the four learning style types and were generally high, ranging from 0.71 to 0.86 for the initial sample and 0.86 to 0.93 for the replication sample.

The *Facilitator's Guide to Learning* contains a section that discusses the validity of the LSI3, but it is only 10 lines long. In that section, Kolb directs readers to a bibliography that includes studies that tested the validity and applicability of the LSI. In other parts of the guide, Kolb refers to the validity of the LSI3. For example, on page 41 he states that 'research on the LSI has tested the relationship between individual learning styles and the careers people choose, and found a strong correspondence between the two'. On page 12 he mentions a number of studies that examined the relationships between performance on the LSI and other instruments (e.g. Myers-Briggs Type Indicator, Learning Style Questionnaire) that measure similar constructs. Nevertheless, these points are not elaborated and the studies are not referenced.

Commentary

The strength of the LSI3 lies in its brevity and its simplicity. It can be administered, scored and interpreted by most people in a relatively short time. The content and instructions of the test are clearly written, and are easy to follow and understand. The colour-coded scoring format facilitates scoring and the extensive use of graphics and diagrams in the test booklet and the guide enhances the test taker's understanding of the theory of learning and its associated constructs.

There is a concern regarding the appropriateness of the norms. According to Kolb, the comparison group used for the LSI3 is the same as the one used in the LSI2. This might not be appropriate given that the formats of the two versions are different. The order of the sentence ending for the four learning modes is the same for the 12 sentences of the LSI2, but the order of the ending of the LSI3 is randomised. Given that this change in format has led to changes in internal consistency and test-retest reliability (Veres et al., 1991) and that the equivalence of the two versions has not been demonstrated, the use of the LSI2 normative comparison group for the LSI3 might not be appropriate.

Changing the order of the sentence ending from fixed to random allows for a more accurate estimation of the internal consistency and test-retest reliability of the LSI. The internal consistency of the latest version is found to be lower than that of the previous version, and lower than expected. The test-retest reliability of the test is found to be better than that of the previous version.

Given that validity was the main concern raised in a review of the LSI2 (Gregg, 1989), it is disappointing to see that very little effort was devoted to addressing the validity of LSI3. Rather than summarising and discussing data and evidence that provide support for the various types of validity, Kolb simply refers readers to a bibliography and makes general statements about the validity of the LSI3. This lack of effort is also surprising given that interesting issues have emerged in the literature regarding the psychometric properties of the LSI3, such as whether it is appropriate to use ipsative test scores in a factor analysis to evaluate the construct validity of the LSI (Geiger, Boyle & Pinto, 1993; Loo, 1999). It is also disappointing to see that Kolb does not clarify in the validity section whether new data have been collected specifically to examine the validity of the LSI3 and whether evidence that supports the validity of earlier versions of the LSI can be used to support the LSI3. Given that the equivalence of the various versions of the LSI has not been demonstrated and that the correlations between these versions are not included in the validity section, it is difficult to evaluate the validity of this latest revision of the LSI.

Summary

The LSI3 is the latest revision of a self-report instrument for describing and identifying one's learning mode and learning style. Although the author has provided evidence to support the reliability of the instrument, he has not provided adequate and suitable evidence to support the validity of this latest version of the instrument. This is disappointing given that the LSI seems to be a popular and promising instrument in the educational and organisational literature.

References

Geiger, M A, Boyle, E J & Pinto, J K (1993). An examination of ipsative and normative versions of Kolb's Revised Learning Style Inventory. *Educational and Psychological Measurement*, *53*, 717–26.

Gregg, N (1989). Review of the Learning Style Inventory. In J C Conoley & J J Kramer (Eds.), *The Tenth Mental Measurements Yearbook* (pp. 441–2). Lincoln, NE: Buros Institute of Mental Measurements.

Loo, R (1999). Confirmatory factor analyses of Kolb's Learning Style Inventory (LSI-1985). *British Journal of Educational Psychology*, *69*, 213–19.

Veres, J G, Sims, R R & Locklear, T S (1991). Improving the reliability of Kolb's Learning Style Inventory. *Educational and Psychological Measurement*, *51*, 143–50.

Administering psychological tests

After selecting a psychological test for a client, the following points need to be considered before administering the test:

1. Ensure that the test is appropriate for use with the particular client in terms of age, educational level and ethnic background.
2. Ensure a suitable venue is selected and booked for administration of the test.
3. Check that all test materials are present, intact and in order.
4. Ensure adequate time is spent becoming familiar with the test so that standardised instructions and procedures are used (Kendall et al., 1997).

Failure to ensure that the test chosen is appropriate for the client's age, gender, educational level and ethnic background can have serious implications for the client. For example, erroneous conclusions can be drawn and wrong

decisions made based on a low aptitude test score for a client who was born overseas. Despite having aptitude in the area, the client might not have the required language skills to undertake and complete the test.

To obtain reliable and valid results for a client on a test, the venue for testing must have enough space, suitable furniture, adequate lighting and ventilation, and minimal distraction. For example, in conducting a group testing session for twenty clients, it is important to select and reserve a room that is big enough and has enough tables and chairs for everyone. Also, flat, stable and sizeable surfaces are needed for conducting tests that require writing on test booklets. In testing younger clients, children's furniture is needed to ensure that the child is seated comfortably and at the appropriate height. Finally, a room that is too hot or too cold will definitely affect the comfort and test performance of a client.

Before administering a test, a check needs to be made to ensure that all of the materials required for a test session are in the test kit and that the test materials are intact (e.g. test apparatus is not broken and test booklets or record forms are not torn or marked intentionally or unintentionally by the previous test user). Although this point applies particularly to where test materials are shared, it is good practice for any test user to spend time checking the test kit before the assessment session.

For the novice test user, or for users who have not administered a particular test for some time, it is essential that time be set aside before the actual testing session to review the details of administration (i.e. instructions, starting and ending rules, time limit and number of subtests that need to be administered). Failure to do so can lead to administration errors, embarrassment during the testing session, the waste of testing time and the collection of incorrect test responses and results.

As pointed out in Chapter 1, one of the important characteristics of a psychological test is the use of standardised materials, instructions and procedures for assessing a construct. This ensures that the results obtained for a client are comparable to the normative group and to other individuals who are administered the same test.

While not directly relating to the practical side of test materials and testing, two other issues need to be attended to before psychological tests are administered to clients; namely, building rapport and explaining the reason for testing. An experienced psychologist does not usually rush into administering psychological tests at the beginning of the testing session. Rather he or she will spend some time building rapport and explaining the purpose of testing. These are important to ensure that the client feels comfortable and cooperates with testing.

Scoring psychological tests

Despite the fact that clear instructions for scoring are provided in most psychological test manuals, errors can still occur among both novice and experienced test users. Some of the most common errors include miscalculations, incorrectly reading tables and incorrectly transferring scores on test forms. For example, Simons, Goddard and Patton (2002) found significant and serious error rates in a sample of 1452 test results collected by a national Australian private sector employment agency on a number of psychological tests, which included the Vocational Interest Survey for Australia, Rothwell Miller Interest Blank, Beck Depression Inventory–Second Edition, Myers-Briggs Typology Indicator–Form M, Competing Values Managerial Skills Instrument, ACER Higher Tests ML and MQ, and Multifactor Leadership Questionnaire–5 X Revised. As another example, Charter, Walden and Padilla (2000) found that many different types of clerical errors (i.e. in addition, in using conversion tables, and in plotting scores) were found for 325 test performances of the Rey-Osterreith Complex Figure Test administered by a psychologist and two test technicians who were well trained and experienced. These errors can have significant implications for the summary and interpretation of results, and in conclusions and recommendations for the client.

Interpreting results of psychological tests

The ability to interpret the results of psychological tests for a client is an essential requirement for competent test use, although this is one of the more difficult skills to teach and acquire (Groth-Marnat, 2009; Suhr, 2015). To interpret test results properly and meaningfully, it must be recognised that, compared with measurement in the physical sciences, measurement in psychology is less precise and more prone to error. It follows that the final score obtained cannot be taken as absolute. Rather, there is a margin of error for the score obtained and allowance for this fact must be made during interpretation (see the discussion of the standard error of measurement in Chapter 4). Moreover, frequently there are interpretative guidelines provided in the test manual or established interpretative procedures for the test published in the research literature, and these need to be followed in test interpretation. Finally, test results of a client cannot be interpreted in isolation. Rather, they should be interpreted within the context of all the other relevant background information and assessment data collected (e.g. answers to interview questions, educational background, school grades, and developmental and medical history). It is also the case that extraneous factors such as anxiety, depression, medication and lack of sleep can influence test performance, and these need to be ruled out as alternative explanations before drawing conclusions based on the test results.

Communicating the findings of psychological testing

To be useful, the results of psychological testing should be communicated to the client or the referral agent in a clear and timely manner. This is usually done in the form of a written report, often supported by an oral feedback/explanation. There is an accepted format for a psychological report and agreement about what information should be included (more discussion of psychological report writing is included in Chapter 9). To be understandable, a report needs to be written in language that is free from jargon and conforms to accepted standards of spelling, grammar and usage. Most importantly, a psychological report should directly and adequately answer the referral questions and include suggestions or recommendations that are based on the results obtained; these suggestions or recommendations should be logical and implementable.

Keeping case records

One of the aspects of psychological testing and assessment that is not usually emphasised—or sometimes not even discussed—is the importance of maintaining a clearly labelled and well-organised file of cases that have been seen (Vandecreek & Knapp, 1997). A good system facilitates the filing of information for clients and the speedy retrieval of records when they are required for retesting, legal consultation or other purposes. In this regard, there are usually legal requirements in keeping records that might differ across countries and states and that ethical practice dictates need to be observed. As an example, the Australian Psychological Society *Code of Ethics* (2007) specifies that unless legal or organisational requirements specify otherwise, psychologists keep client records for a minimum of seven years from the last contact. In the case of information and data collected when the client was less than 18 years old, records are required to be kept at least until the client reaches 25 years of age. While advances in computer hardware and software (e.g. scanners, message storage devices and cloud storage) have increased the ease of storage and size of storage space, they can introduce other issues such as importance of regular backup, protecting privacy and preventing unauthorised access.

Ethics

No discussion of psychological testing and assessment would be complete without consideration of the ethical issues involved. Indeed, one of the most

extensive sets of ethical guidelines issued by the Australian Psychological Society is concerned with psychological testing and assessment. Clearly this topic is a very salient one to psychology as a profession.

Consideration of ethical behaviour can be traced through the millennia, from the writings of ancient Greek philosophers such as Pythagoras, Plato and Aristotle through medieval religious scholars to modern philosophers such as Hobbes, Locke, Mill, Kant and Rousseau. Contemporary professional ethics is more concerned with standards of daily practice within the domain of a profession than with the development of a complete ethical system, although many of the principles underlying appropriate professional behaviour can be found in philosophical and religious writings. Indeed, it can be said that one of the defining features of a profession is adherence to a code of ethics, and most professional bodies, including legal, medical and psychological societies, are concerned with developing such codes.

Ethics can be defined simply as the formulation of principles to guide behaviour—in this case *professional* behaviour—with respect to clients, colleagues and the general public. Codes of ethics are an attempt at self-regulation by a group of professionals. Self-regulation and a sense of propriety and ethics are among the defining features of any profession. It has been said that more careers have been damaged by a lack of ethical knowledge than by a lack of technical knowledge or subject matter (Francis, 1999), so knowledge of a code of ethics and a sense of ethical behaviour is vital for any professional. Indeed, professional practice relies heavily on professional reputation, and reputation can be easily ruined by unethical conduct. The good news is that virtually all ethical problems are preventable and they arise more through carelessness than through malice.

ethics
a set of principles for guiding behaviour; in the case of psychological testing and assessment, for guiding professional behaviour

Sometimes students new to studying ethical issues ask: 'If I don't belong to a professional society, does that mean I am exempt from their code?' A better question might be to ask yourself: 'Do I really think I can sidestep broadly held standards of behaviour within my profession and get away with it?' Irrespective of your professional membership, is anyone going to take you seriously after that? Like it or not, you will be held accountable to the standards of your profession and any psychologist registered by the Psychology Board of Australia is automatically bound by the code of ethics formulated by the Australian Psychological Society. As such, it behoves all students of psychology to remain vigilant to potential ethical problems—society at large certainly will (see Box 2.3)

—and study of a code of ethics is a step in that direction. One obvious advantage of employing a psychologist for conducting testing and assessment is that they are bound by a code of professional ethics.

Some commentators have claimed that ethics is something that cannot be taught, learnt from a book or captured in a code; that it can only be acquired through experience. Indeed, Kohlberg's famous psychological theory of moral reasoning puts ethical understanding at the most advanced stage of development and it is certainly possible that some individuals might never advance to that level (Kohlberg, 1981). Conversely, it must be conceded that no one is born with a sense of ethical behaviour. There is no such thing as an ethics gene, so each of us must develop an ethical mindset through experience and conscious consideration of ethical issues. Learning and reading about ethical issues can take you a long way towards developing an ethical mindset.

Box 2.3

Lessons from Chelmsford

Between 1963 and 1980, a psychiatrist working at a private clinic in outer metropolitan Sydney developed his own unique method of treating mental illness (Slattery, 1989a, p. 47). Having the legal authority to prescribe medication, he massively sedated patients and confined them to bed. This so-called 'deep sleep therapy' (DST) was based on the idea that patients could literally sleep off their mental illness and wake up well. Patients were kept unconscious, sometimes for weeks on end, during which time they were also subjected to daily bouts of electroconvulsive shock. It certainly made for an easily managed psychiatric ward, described by some as 'quiet as a tomb' (Slattery, 1989b, p. 32). There were no patients wandering around in a confused state or shouting meaninglessly into the air. Tragically, some of them started to die. Psychiatric treatments are not supposed to be fatal and eventually people began to take notice.

Under pressure from several media exposés into the hospital, the New South Wales state government eventually set up a Royal Commission into the affair in 1988. The Royal Commission into Deep Sleep Therapy, chaired by Justice John Slattery, tabled its final report in 1989, totalling fourteen volumes of evidence and discussion. Royal Commissioner Slattery concluded that at least twenty-four deaths between the years 1964 and 1977 could be directly attributed to DST (Slattery, 1989b, p. 25), with at least two cases of brain damage caused by the treatment (Slattery, 1989b, p. 1). Unfortunately, the psychiatrist checked

himself in for DST in 1978 and committed suicide seven years later in the lead-up to the Royal Commission.

What has this awful tragedy got to do with psychological assessment? Well, one of the main sources of evidence used by the psychiatrist to support his continued use of DST was psychological assessment reports provided by a Sydney-based psychologist. These reports purported to show an improvement in patient symptoms and functioning as a result of the treatment. Because of this, the terms of the Royal Commission were expanded to include psychological testing in 1989, and two volumes of the final report directly related to these matters. There is little doubt that the psychiatrist truly believed DST was helping his patients and he sought what he believed to be quality scientific evidence to back up his procedures.

Although there was no strong correlation between the assessment reports and patients being given DST (Slattery, 1989c, p. 70), with the final decision to use DST the province of the psychiatrist, the psychologist's assessments were used to support DST in a number of ways. First, they were used to help explain to patients their particular condition and why they needed DST. Second, they were used to show that there appeared to be no adverse effects associated with DST, and indeed that DST was beneficial.

An expert panel of psychologists comprising local academics and experienced practitioners was formed at the behest of the Royal Commission to independently evaluate the assessment reports. The panel criticised the assessments on several grounds (Slattery, 1989c, p. 68). First, some of the tests used lacked reliability. Second, their validity had not been established in some cases, and certainly not for the use to which they were put; that is, diagnosing improvements in psychiatric conditions. Third, some of them had inappropriate norms; and fourth, the psychologist in question was 'idiosyncratic' in his application and use of the tests; that is, he scored them differently on different occasions, sometimes combining scores in undocumented ways. Readers will become aware that all of these shortcomings are in areas directly covered by this textbook. We have chapters on reliability, validity and the use of norms. The criticism of idiosyncratic use clearly indicates a lack of standardised administration and scoring procedures. Such processes are undocumented and unable to be replicated and therefore cannot claim to have any scientific basis.

The psychologist in question claimed to have an 'eclectic approach to testing' and justified his idiosyncratic use of the tests on the grounds that it was a legitimate application of his clinical judgment and experience (Slattery, 1989c, p. 70). A number of other psychologists making submissions to the Royal Commission concurred with these views. Psychological assessment, they claimed, was broader than the mechanical administration of tests and inevitably required the amalgamation of diverse information even to the extent of coming up with one's own set of unique composite scoring rules.

It has been said, paraphrasing Newton's third law of motion, that for every expert there is an equal and opposite expert, and nowhere is this more true than in psychology. The two volumes dedicated to psychological testing by the Royal Commission contain argument and counter argument by respected psychologists on the use of psychological tests, and sometimes it is difficult to determine who is right. This is the hallmark of an ethical dilemma. There is also the suggestion that prior to treatment the psychologist might have unwittingly used his clinical judgment to exaggerate patients' symptoms, which were subsequently found to be reduced after DST during post-test (Slattery, 1989c, p. 70). We can only reiterate that psychological testing emphasises standardisation of procedures for good reason and that the objective information provided by mechanical procedures is probably the only safeguard against wishful thinking clouding one's clinical or subjective judgment.

In the end, the Royal Commissioner concluded that it was not possible to determine which conclusions were based on the results of the tests and which were based on the psychologist's subjective clinical opinion, but nevertheless implicated the tests in contributing to the continued use of DST (Slattery, 1989d).

What conclusions can be drawn from this sorry tale? Looking at the code of ethics in Box 2.4, we can see the relevance of paragraphs B.13.2, B.13.3, B.13.4, B.13.5 and especially B.13.6. It is probably no coincidence that registration of psychologists was introduced into New South Wales in 1990 just after the completion of the Royal Commission, almost 20 years after registration occurred in other parts of Australia. Without a register of psychologists it was impossible for authorities to impose any disciplinary action other than legal proceedings; and these, as we have seen, only pertain to minimal standards of behaviour. With a code of ethics and a register, you can be struck off. Moreover, all states and territories in Australia now have a national health practitioner regulation agency.

How do ethics differ from related issues like morality or the law? Considerations of morality, law or virtue are usually more general than an ethical code. Morality pertains to a pervasive set of values to live by, whereas ethics focuses on principles to guide behaviour in certain situations. Conversely, the law seeks to define minimum standards of acceptable behaviour, and many people are satisfied with behaving just within those minimum standards. Ethics, on the other hand, seeks to define the highest standards of behaviour. The law can be slow to change, whereas codes of ethics are readily amended. For example, the Australian Psychological Society code has been updated several times since its introduction in 1968. It is important to realise that you can be considered to have acted unethically even though you might have done nothing illegal.

Code of ethics of the Australian Psychological Society

The latest update of the Code of Ethics of the Australian Psychological Society (2007) is based on three broad principles. These are respect for the rights and dignity of people and peoples, propriety and integrity. Based on each of these principles, a number of ethical standards are derived (see Table 2.3). The ethical standards relating to psychological assessments are summarised in Box 2.4.

Table 2.3: General principles and ethical standards of the Australian Psychological Society

General principles	Ethical standards
A. Respect for the rights and dignity of people and peoples Psychologists regard people as intrinsically valuable and respect their rights, including the right to autonomy and justice. Psychologists engage in conduct that promotes equity and the protection of people's human rights, legal rights and moral rights. They respect the dignity of all people and peoples. (p. 11)	A.1. Justice A.2. Respect A.3. Informed consent A.4. Privacy A.5. Confidentiality A.6. Release of information to clients A.7. Collection of client information from associated parties

General principles	Ethical standards
B. Propriety Psychologists ensure that they are competent to deliver the psychological services they provide. They provide psychological services to benefit, and not to harm. Psychologists seek to protect the interests of the people and peoples with whom they work. The welfare of clients and the public, and the standing of the profession, take precedence over a psychologist's self-interest. (p. 18)	B.1. Competence B.2. Record keeping B.3. Professional responsibilities B.4. Provision of psychological services at the request of a third party B.5. Provision of psychological services to multiple clients B.6. Delegation of professional tasks B.7. Use of interpreters B.8. Collaborating with others for the benefit of clients B.9. Accepting clients of other professionals B.10. Suspension of psychological services B.11. Termination of psychological services B.12. Conflicting demands B.13. Psychological assessments B.14. Research

General principles	Ethical standards
C. Integrity Psychologists recognise that their knowledge of the discipline of psychology, their professional standing and the information they gather place them in a position of power and trust. They exercise their power appropriately and honour this position of trust. Psychologists keep faith with the nature and intentions of their professional relationships. Psychologists act with probity and honesty in their conduct. (p. 26)	C.1. Reputable behaviour C.2. Communication C.3. Conflict of interest C.4. Non-exploitation C.5. Authorship C.6. Financial arrangements C.7. Ethics investigations and concerns

Box 2.4

Section B.13 of the Australian Psychological Society's Code of Ethics: Psychological assessments

B.13.1 Psychologists use established scientific procedures and observe relevant psychometric standards when they develop and standardise psychological tests and other assessment techniques.

B.13.2 Psychologists specify the purposes and uses of their assessment techniques and clearly indicate the limits of the assessment techniques' applicability.

B.13.3 Psychologists ensure that they choose, administer and interpret assessment procedures appropriately and accurately.

B.13.4 Psychologists use valid procedures and research findings when scoring and interpreting psychological assessment data.

B.13.5 Psychologists report assessment results appropriately and accurately in language that the recipient can understand.

B.13.6 Psychologists do not compromise the effective use of psychological assessment methods or techniques, nor render them open to misuse, by publishing or otherwise disclosing their contents to persons unauthorised or unqualified to receive such information.

Acting ethically means more than memorising a list of dos and don'ts, even though this is what most codes of ethics appear to be. Ethical principles extend beyond circumstances specifically mentioned in any code, and learning to behave ethically is more about understanding the principles that underlie the code than being able to recite it point for point.

Accommodating the differently abled

One of the major advantages of psychological tests is that performance on a test can be compared across individuals because all individuals complete it in the same standard way. If testing were to vary from individual to individual there would be no basis for comparison across individuals because any differences could be due to the individuals or to the ways the test was administered. Psychological tests are, however, used in psychological assessment and this must consider the whole person. Where people are differently abled this needs to be taken into account. An obvious example is that of a blind person and a test that requires the person being tested to read. Here there needs to be some accommodation of the test (e.g. presentation using Braille) if the result of testing is to contribute to assessment of the person.

When and how to accommodate psychological tests to meet individual needs are important practical questions. A good treatment of some of the answers is outlined in Reynolds and Livingston (2014). Essentially, tests can be modified where the modification is not central to the construct being assessed. In the above example, if the test was one of acuity of vision or reading speed, then adapting it by using a Braille format would defeat its purpose. That is, the test user needs to be clear about the purpose of the assessment in judging whether a modification is appropriate. As to how the modification is to be done, this depends on the abilities of the person being assessed and the ingenuity of the tester. Among the options, one test could be substituted for another or not all items of a test administered.

Accommodations need to be noted for their possible influence on the result and therefore its comparability with the results for other individuals. Some suggest that this might be discriminatory, and there is legislation in Australia relevant to the issue that needs to be recognised in the case of psychological testing as with other activities. The Australian Psychological Society (2015) has published guidelines for testing people with disabilities. Practical assessment involves the use of judgment that is informed by the best research available.

Cultural differences, testing and assessment

Australia is not a monoculture in which everyone shares the same language of origin, religion, customs and worldview. If they are to be effective, professions seeking to provide a service to Australians must necessarily recognise cultural difference as a critical feature of the lives of their clients and of their practitioners. Over the years, professional organisations have developed guidelines for best practice in this regard (e.g. Australian Psychological Society *Code of Ethics*, A1. Justice). It follows that the use of psychological tests and the conduct of assessment must recognise cultural differences. The psychological testing and assessment enterprise became aware of the importance of cultural differences quite early in its history. The Binet test originated in France but was adapted by Terman for use in the USA. In so doing, Terman and his associates had to translate the verbal items into English and consider more generally the applicability of all test items for use with children who had not received the typical French kindergarten and primary education. They commented thus: '[The] Binet scale requires radical revision to make it at all suitable to conditions in this country' (cited by Rogers, 1995, p. 199). That is, simple translation was not enough and the relabelling of the test the Stanford-Binet was warranted given the changes that needed to be made for it to be used effectively in the USA. Nowadays, it is more usual to talk of adapting tests for use in another culture rather than translating them, and the ways this is done are discussed a little later.

First, it is important to realise that the impact of culture on a test goes beyond the language used to issues of the information matrix in which the test is embedded and the ways of thinking that are necessary to solve the questions posed. The questions might be posed in English, but they assume a great deal more of the test taker than access to the language. There are taken-for-granted understandings of the culture in which the test is developed that go to make the testing exercise possible (see Goodnow, 1976). If these understandings are not shared between tester and test taker, then the task is not presenting to each of them in the same way, and the interpretation of performance on the test is not straightforward.

For example, Reddy, Knowles and Reddy (1995) showed differences between Australian and Indian samples on self-report tests of well-being, although the language of instruction for both groups had been English. There were similarities in response when the test was presented as a series of adjectives to be endorsed, but differences when the test was in the form of sentences that required an understanding of context. Carstairs et al. (2006) compared performance on a number of tests of cognitive functioning for groups from a non-English-speaking background whose first language was other than English, from a non-English-speaking background whose first language was English, and from an English-speaking background (ESB). They observed that two factors affected performance. One was language, which affected verbal tests, and the other was sociocultural background, which had a greater effect on non-verbal tests. The

role of language in testing was examined in some detail in a special issue of the *International Journal of Testing* (Zenisky, 2015).

The effects of language and culture on tests are so pervasive that repeated attempts to develop tests that are 'culture free' have proved unsuccessful. The Queensland Test (McElwain & Kearney, 1970, 1973) was specifically developed for use with Indigenous Australian people without the need for communication in English. However, the test was found to be sensitive to the degree of contact Indigenous people had with mainstream non-Indigenous culture, with the greater the contact the higher the score. In reviewing cognitive and neuropsychological tests in use in Asian countries, Chan, Shum and Cheung (2003) concluded that 'our early experience of language may exert subtle effects on the performance of non-verbal tests' (p. 258). Cattell (1979), having tried for many years to develop a 'culture free' test of intelligence, settled for a 'culture fair' test in which the effects of language and culture were reduced rather than eliminated (Cattell, 1940, 1979).

What, then, is a culture fair or culture reduced test? Essentially a **culture fair test** is one for which there is no systematic distortion of scores resulting from differences in the cultural background of the test takers. Test scores are subject to error, but when this is random the scores can move up and down and its effects can cancel out, at least over a group of test takers or a number of occasions of testing. When the error is systematic it is more insidious because all scores are moved in one direction, making them higher or lower than they should be. Producing a culture fair test requires that test items have the same meaning in each of the cultures in which the test is to be used, and that it acts as a predictor of socially relevant criteria in each culture in the same way. That is, there must be an equivalence across cultures in what is termed the test's construct validity and in its predictive or criterion validity (these concepts are discussed in Chapter 5). This does not mean that there must be equivalence in the average scores for different cultural groups. It is not too difficult to find a set of items that will give rise to difference between groups on average score on the test. The Koori IQ test (http://docslide.us/documents/koori-iq-test.html), for example, is a set of twenty words, the meanings of which are likely to be known by a majority of Indigenous Australians living in New South Wales, but are unlikely to be known by people in that state who have been educated in mainstream non-Indigenous culture. The average scores for representative samples from the two groups would likely differ, and in favour of the Indigenous sample, but it is not at all clear what meaning to attach to the difference or what, if any, consequences it has. These are questions that take us beyond the face validity of the items and require an examination of the construct and predictive validity of the test.

culture fair test
a test devised to measure intelligence while relying as little as possible on culture-

specific knowledge (e.g. language); tests are devised to be suitable across different peoples, with the goal to measure fluid rather than crystallised intelligence

Differences in average scores can be a problem when they occur, a situation that is usually termed adverse impact (Landy & Conte, 2007), but it is best addressed in its own right rather than taken as a necessary indication of cultural bias. For example, within-group norming can be applied, as described in Chapter 7. Rather than use the total raw score on the test to make decisions (e.g. who to employ), total scores are expressed as deviations from the average for the group to which a test taker belongs. Different raw scores can give rise to the same deviation score and, if decisions are made on the basis of the deviation scores rather than the raw scores, adverse impact is minimised.

To define bias in a test on the basis of difference in average scores between groups is to rule out other possible causes of the difference. If test scores of two cultural groups differ because members of one cultural group have been disadvantaged in various ways for some time, then the test might simply be identifying a genuine problem rather than introducing a systematic error. Requiring that the averages be the same in order for a test not to be judged culturally biased might in fact be disguising a problem better revealed so that it can be addressed. But just as a difference in averages on a test between cultural groups is not necessarily evidence of bias, the failure to show a difference between group averages is not evidence that a test is unbiased. Absence of evidence is not evidence of absence (see, for example, Altman & Bland, 1995). Bias must be determined independently of average differences.

Equivalence in the meaning of a test across cultures depends on showing that the items comprising the test and the total score that is calculated from the items mean the same in both the culture in which it was developed and the culture to which it is to be applied. This involves both a consideration of item content and the procedure for administration, and an examination of the psychometric properties of the test. Meaning involves linguistic considerations, but, as noted earlier, it is broader than this, touching on questions of the familiarity that the task itself engenders. As for the psychometrics of meaning, a major consideration is that the items must hang together in the same way for samples of participants drawn from the two cultures. That is, they must show similar patterns of intercorrelation in the two samples, and when subjected to the statistical technique of factor analysis (discussed in detail in Chapter 5) the results should be highly similar—if not identical. In addition, the relative contribution of the items to the total score on the test should be the same across the two samples. These are criteria that Reddy, Knowles and Reddy (1995) applied in their examination of the differences between Indian and Australian samples on tests of well-being described earlier. Some would add a further requirement: both

cultural groups must be included in the development of norms (see Chapter 3) for the test.

In terms of equivalence of predictive validity across cultures, a number of requirements have been proposed, with the most stringent being those by Cleary et al. (1975). They argued that the relation between test score and any socially relevant criterion that is to be predicted by the test (e.g. employability for a job or the need for a remedial education program) should be the same in both cultural groups. Given these ways of assessing possible bias in tests in current use, how well do tests perform? The matter has been most extensively examined in the case of cognitive tests because this is where the greatest danger from cultural differences is perceived to lie (this evidence is considered in more detail in Chapter 7). Although the evidence is extensive, it is drawn almost wholly from studies conducted outside Australia, principally in the USA involving whites, African Americans and Hispanics. We must therefore extrapolate from this body of evidence to answer the question about cultural differences in other parts of the world.

Although not unanimous in outcome, the weight of this evidence is that psychological tests are not biased by cultural differences (see, for example, Gottfredson, 2009; O'Boyle & McDaniel, 2009). Such a conclusion is not grounds for complacency. Instances of bias can be revealed in further studies or new methods of analysis might show up limitations in what has been done to date. Nor is a conclusion of no evidence of bias a reason to be any less stringent in developing new tests or adapting older ones for use in different cultures. A good deal is known about best practice in this regard, and the International Test Commission (ITC) has provided a set of guidelines on how tests are to be adapted (www.intestcom.org/page/16).

Ideally, the process begins with an expert panel with cultural competence (e.g. knowledge of the language and customs) in the cultures in which the test is to be used (e.g. English and Chinese) and with understanding of the psychological construct or constructs to be assessed in the test. The panel evaluates old and any additional items proposed in terms of their cultural specificity, the extent to which they represent the diversity of cultural expression of the underlying trait, and any offence they might unintentionally give respondents. Items chosen on the basis of this evaluation are then translated into the language of the other culture, using if possible a different panel of bilingual experts. This is often termed forward translation, but it involves more than just a literal translation of the words used in the items or in the instructions. Rather, it attempts to capture in a culturally appropriate way the meaning of the original. Once done, a third bilingual panel translates the test back into the language of the original test and the original panel then assesses the equivalence of the back translation to the original test. Only if the meaning has been preserved can the content of the test be considered equivalent in the two cultures. This is an

expensive exercise and it is not surprising that few adaptations to date have met such stringent standards. A notable exception is the work of Taouk, Lovibond and Laube (2001) in adapting the Depression Anxiety Stress Scales (DASS–21) developed in Australia for English-speaking samples for use with Mandarin speakers. In the future, however, test constructors and test users will be expected to meet the ITC standards. A good model for this is the report of Ægisdóttir and Einarsdóttir (2012) of the adaptation for use in Iceland of the Beliefs about Psychological Services scale (I–BAPS), originally developed in the USA.

The examination of cultural differences has rightly commanded a good deal of research, because the social consequences of ignoring them are considerable. Those consequences have led some to give up on testing as an appropriate technology in dealing with members of different cultures. Graham Davidson is an Australian psychologist who has spent more than 30 years working with Aboriginal and Torres Strait Islander people in most Australian states and territories. He has argued that the suspicion of psychological testing by Indigenous people is so great, and the difficulties with tests so limiting, that apart from some specialised forms of neuropsychological testing, the use of tests with Indigenous Australians should be abandoned and in its place a more individualised form of assessment be practised (Davidson, 1995). The history of psychological testing with Indigenous groups is not a happy one, confounded as it is with ideas of racial inferiority and cultural deficits (Klich, 1988; Rickwood, Dudgeon & Gridley, 2010; Ross, 1984), and even now there are few validated tests for use with Indigenous peoples (Dingwall & Cairney, 2010). Although the present authors are not as pessimistic about the limitations of tests as Davidson —and this view is shared by Dyck (1996), who wrote a rejoinder to Davidson's paper (but also see Davidson, 1996)—Davidson's observations about the deep suspicion of psychological testing held by Indigenous people must be accepted and their view respected. A good starting point for this are the chapters in Dudgeon, Milroy and Walker (2014), particularly the chapter on assessment by Adams, Drew and Walker (2014). Our position is that:

- Cultural differences can lead to bias in the use of psychological tests.
- Several criteria need to be applied to adequately assess whether cultural differences are biasing test results.
- To date, most studies applying these criteria have not found evidence of bias, but this can only be conditional on the outcome of further research. The cultural background of the person to be tested (the client) must be appreciated and respected if the psychologist is to perform the task competently and ethically.
- Assessment is more than testing because it involves decisions about whether a test should be used in the first place and, if it is, how the test score is to be

interpreted against the background of a full knowledge of the person, including their cultural experiences.

Chapter summary

In this chapter we have explained the difference between psychological testing and assessment. The latter is a broader process that aims to answer referral questions. Psychological testing is one of the tools that is commonly used in psychological assessment. Other tools include observation, interview and record checking. We have also provided some examples to illustrate the different types of psychological tests. We introduced and described some of the best practices and ethical principles relating to psychological testing and assessment, and finally discussed the testing and assessment of individuals with disability and individuals from different cultures. The guidelines and principles outlined in this chapter are important for ensuring the quality of assessment services provided by psychologists.

Questions

1. Is psychological testing the same as psychological assessment? Discuss.
2. What are some of the major differences among psychological tests?
3. What is ethics and why are ethical principles needed?
4. Why do we need ethical principles to guide the practice of psychological testing and assessment?
5. When one wants to purchase a new psychological test, where can one go to find information to guide the purchase?
6. Go to the ACER website (www.acer.edu.au) and find out about its system for supplying tests to users. How is this system different from that used by Pearson Clinical Assessment (www.pearsonclinical.com.au)?
7. A psychologist wants to use the WISC–IV to test the IQ of a 5-year-old boy. Find out who supplies this test and how much it costs. Do you think it is a suitable instrument for this purpose?
8. What can a psychologist do if someone with a disability cannot undertake the test?
9. How can cultural differences influence scores on psychological tests?
10. What needs to be considered in determining whether or not a test is culture fair?

Further reading

Adams, H E & Luscher, K A (2003). Ethical considerations in psychological assessment. In W T O'Donohue & K E Ferguson (Eds.), *Handbook of professional ethics for psychologists: Issues, questions and controversies* (pp. 275–83). Thousand Oaks, CA: Sage.
Camilli, G (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–56). American Council on Education. Westport, CT: Praeger.
Geisinger, K. (2013). Testing and assessment in cross-cultural psychology. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology. Vol. 10: Assessment psychology*. (pp. 114–39). Hoboken, NJ: Wiley.
Matarazzo, J D (1990). Psychological assessment versus psychological testing: Validation from Binet to the school, clinic and courtroom. *American Psychologist*, *45*, 999–1017.
Zenisky (2015). Not a minor concern: Introduction to the themed issue on the assessment of linguistic minorities. *International Journal of Testing*, *15*(2), 91–3.

Useful websites

Buros Center for Testing: http://buros.org
Careers in psychology (Australian Psychological Society): www.psychology.org.au/studentHQ/careers-in-psychology
Ethics (Australian Psychological Society): www.psychology.org.au/about/ethics
International Test Commission: www.intestcom.org

PART 2 METHODOLOGICAL AND TECHNICAL PRINCIPLES OF PSYCHOLOGICAL TESTING

3 Test Scores and Norms

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. explain the difference between norm referencing and criterion referencing in interpreting test scores
2. explain the difference between linear and non-linear transformation of test scores
3. define what is meant by a standard score and explain how it is interpreted
4. define what is meant by a percentile and describe the ways in which percentiles can be used in psychological testing
5. explain what norms are and the characteristics that determine their value.

KEY TERMS

criterion referencing
deviation IQ
linear transformation
local norms
norm referencing
norms
percentile
standard score
sten score
stratified sampling
T score

Setting the scene

- High school students in different Australian states complete different forms of final examination and yet compete on the same terms for places in Australian universities. How is this done?
- The average IQ is 100. Why is this so?
- A high school student obtains scores on tests in five different subjects (English, mathematics, history, science and French). Can we combine those scores into an overall aggregate mark indicating school performance?
- Can we use the performance of samples of US citizens on intelligence tests to evaluate the performance of, say, New Zealanders?

Introduction

Scores on psychological tests do not have direct meaning but must be interpreted. The most common form of interpretation of a score on a psychological test is to compare it with the scores that similar individuals obtain on the test. Essentially, the question asked is: how likely is it for others who are similar in important respects to the person tested to obtain this score? To the extent that the score is unlikely, it attracts our interest. To interpret a test score, then, we need data from samples of individuals on how they score on the test (technically, a set of **norms**), and a way of expressing the individual test score so that the likelihood of obtaining it becomes apparent. This chapter is concerned with common ways of expressing test scores and with the construction of adequate norms for interpreting them. We consider the two major ways for transforming scores on tests to allow their interpretation, plus their strengths and limitations, and the relationships between them. We also consider the major considerations that need to be borne in mind when developing norms for psychological tests, and examples of these as applied to some of the major tests in use.

norms
tables of the distribution of scores on a test for specified groups in a population that allow interpretation of any individual's score on the test by comparison to the scores for a relevant group

Interpreting test scores

A psychological test is made up of a number of questions or tasks that the person taking the test must answer or complete. The term 'item' is used to refer generically to the questions or tasks that make up a test. The response to the item

is scored by applying a consistent rule. In the simplest case of an item permitting only a right or wrong answer, the item would be scored 1 for correct and 0 for incorrect. Where an item permits more than one right answer but some answers are better than others, it is possible to formulate rules allowing partial credit for the item. Each item thus comes to have an **item score** for each person taking the test. The **raw score total** on a psychological test is the score obtained by summing the item scores on the test. Consider, for example, a test of ability that comprises 50 general knowledge questions (e.g. What is the capital of Papua New Guinea?). Each question or item can be scored in terms of whether the respondent provided the correct or the incorrect answer (Port Moresby is the correct answer for our example). In this example, the item score for the question for a particular respondent would be 1 or 0 and the raw score total would be a number between 0 and 50.

item score
the score for each item on the test

raw score total
(or raw score) the total score on the test found by summing item scores

Raw score totals on psychological tests typically are of little use by themselves and require some way of acquiring meaning. To know that Person X obtains a score of, say, 35 on the general knowledge test tells us very little, because there are many more questions that could have been included in the test. Even if it is assumed that the set of items included is a good sample of the population of general knowledge items, we need to interpret the score of 35 in some way. We might say that a raw score total of 35 constitutes 70 per cent of the total that could be obtained, and because, conventionally, 50 per cent is the 'pass mark' on a test, this represents a reasonably good result. The 50 per cent mark is a useful convention in some circumstances: it indicates that the person knows as much as they don't know and is thus at a threshold point of achievement. In other circumstances—for example, assessing the competence of a brain surgeon—one might want a greater grasp of what is to be known.

When it is possible to specify what is to be known with some precision, the raw score can have meaning in itself. Driving a motor vehicle involves a set of skills, such as engaging the engine, steering into a lane of traffic, stopping and turning. For a person to be judged a competent driver, they need to be able to show mastery of this skill set. To know, for example, how to start the car but not how to stop it, or to go straight but not how to turn, would be considered insufficient. The nature of the task determines the items on the test and gives a score on the test its meaning. The term **criterion referencing** (see Chapter 1) is

sometimes used to describe this situation; the task itself is the yardstick (criterion) to which performance is referred (Allen & Yen, 1979).

criterion referencing
a way of giving meaning to a test score by specifying the standard that needs to be reached in relation to a limited set of behaviours

Not many variables in psychology allow this form of interpretation because the potential item pool for a test often cannot be determined with precision. What, for example, is the possible set of behaviours that lead to a person being described as hostile, depressed or intelligent? It is possible to list a number of these, but the list is far more open-ended than it is in the case of skills such as driving a motor vehicle. Thus, for most psychological variables, the raw score total on a test cannot be directly interpreted.

To give a raw score total meaning in these circumstances, test developers have resorted to 'norm' referencing rather than criterion referencing. That is, they have sought to relate the raw score to the average score (or norm) of a representative group of people similar to the person being tested. A simple example of **norm referencing** occurs when a parent attempts to give meaning to their child's result in a spelling test at school by asking how other children in the class performed on the test. A score of, say, 55 per cent takes on very different meaning if most children obtained scores of less than 30 per cent or if most obtained scores of better than 70 per cent.

norm referencing
a way of giving meaning to a test score by relating it to the performance of an appropriate reference group for the person

The idea of norm referencing is simple, although the way it is put into practice requires some understanding of statistics. The idea is to express the raw score total in terms of its position in a distribution of raw score totals for a sample of individuals with whom it is sensible to compare the individual being tested. The meaning of the raw score total is thus established by its place in the distribution of scores: if it is towards the top end of the distribution the person's performance is better than most; if it is towards the bottom end, performance is poorer than most. The use of this approach to interpretation has, it must be acknowledged, proved controversial in some circles, because the comparison process is thought to be demeaning or as having adverse motivational consequences. The approach seems to be saying to some critics that the individual only has importance when considered in the context of other individuals. Alternatively, to be told that one's score is low relative to the scores of one's peers might lead to feelings of failure and

possibly less effort in the future. Without wishing to minimise the importance of these issues, suffice it to say that these concerns arise with respect to the way tests are used and are not intrinsic to the tests themselves.

We begin our discussion of scoring with norm referencing before briefly considering some other methods.

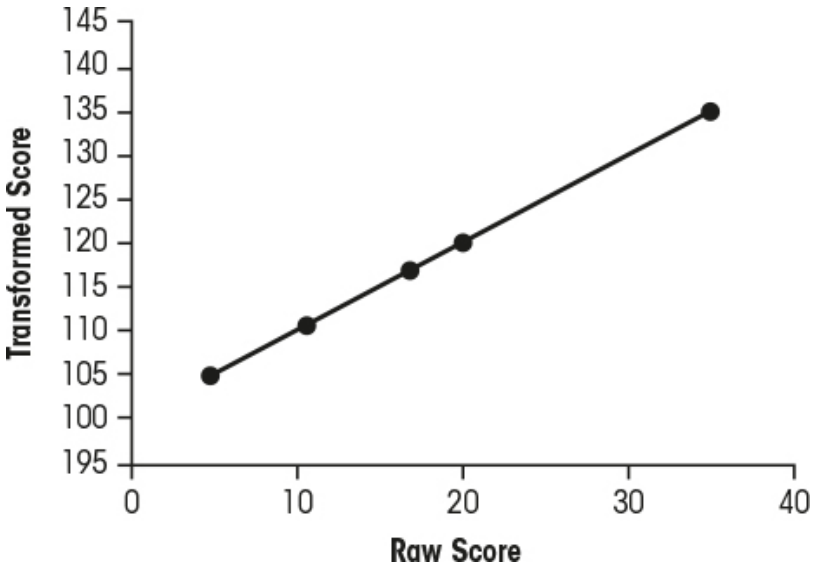
Transforming scores for norm referencing

To refer a raw score total to an appropriate reference group, the raw score has to be changed or transformed to a score that has normative information. Two basic forms of transformation are typically employed: linear and nonlinear. The term 'linear' means that there is a straight-line relationship between two variables, in this case between the raw and the transformed scores. That is, if one were to plot the transformed score against the raw score, the plot would be a straight line.

A simple **linear transformation** is the addition of a constant to all raw scores. Consider the following scores for five students on an achievement test that has a maximum score of 40: 35, 20, 17, 11, 5. A constant value, say 100, is added to each of the raw scores, and a new set of scores results, but it is one closely related to the first: 135, 120, 117, 111, 105. Note that plotting the transformed scores against the raw scores results in a straight line, as in Figure 3.1. The transformation is linear. An essential feature of this type of transformation is that the differences between the raw scores are maintained in the transformed scores, although their magnitude has changed. Thus 17 minus 11 in the raw score set is equivalent to 11 minus 5, and this equivalence is maintained among the transformed scores (17 – 11 = 11 – 5 = 117 – 111 = 111 – 105).

linear transformation
a transformation that preserves the order and equivalence of distance of the original set of scores

Figure 3.1 Plotting the raw scores in our example with the transformed scores



The same effect would be produced if instead of adding a constant we subtracted (or divided or multiplied) each number by a constant. In case of division by a constant, say 5, the set of numbers that results is 7, 4, 3.4, 2.2, 1. A plot of transformed scores against the original again results in a straight line and the equivalence of differences in the original set (e.g. 17 – 11 = 11 – 5) remains in the transformed set (3.4 – 2.2 = 2.2 – 1). In this case, the scale of the differences has changed (the absolute differences in the two sets are not equal as they were with the addition of a constant) but this is easily dealt with (if we want to) by simply multiplying the differences by the constant that was used for division in the first place: 17 – 11 = 11 – 5 = 5 × (3.4 – 2.2) = 5 × (2.2 – 1). It is worth noting that the linear transformation is not limited to one operation as long as the straight-line relationship is preserved. For example, we could both divide by 5 and add 100 to all numbers. You might want to satisfy yourself that the transformation is linear. In general, a linear transformation is one in which the transformed scores are related to the raw scores in terms of a straight-line function, and this means that the equivalence of distances between points on the raw score distribution is maintained in the transformed distribution.

A linear transformation is not the only form of transformation possible. There are many forms of **nonlinear transformation** that are used for different purposes. In the case of psychological tests, we sometimes find that numbers in the raw score distribution are bunched in the middle of the range of scores, affording little discrimination in that region. A test developer might want to draw out the differences in the middle of the range while leaving the values in the tails of the distribution unchanged. This means a nonlinear transformation of the raw scores because in these circumstances the plot of transformed and raw scores will not produce a straight line. Nonlinear transformations are as legitimate as linear ones, but we need to bear in mind that the equivalence of differences only holds with

linear transformations. We will say more about the nonlinear transformation when we encounter one.

nonlinear transformation
a transformation that preserves the order but not the equivalence of distance of the original scores

The straight line is fundamental to the distinction being made here between linear and nonlinear transformations. In geometry, the straight line has an equation that is useful in transforming test scores. The equation is:

$$Y = mX + c$$

It can be read for our present purposes as meaning that the value of a linearly transformed score (Y) is equal to a weighting factor (m) multiplied by the raw score (X) plus a constant (c). In terms of the linear transformations considered to this point, the equation can be written as follows.

For addition of a constant 100, the constant is 100 and the weighting factor is 1:

$$Y=1 \times X+100$$

For division by 5, the weighting factor is 1/5 and the constant is 0:

$$Y=\frac{1}{5} \times X+0$$

For division by 5 and addition of 100:

$$Y=\frac{1}{5} \times X+100$$

The same basic equation characterises these transformations because they are all linear. (Note that if Y were a function of some power of X, say $Y = X^a$, the relationship would not be linear.)

In transforming psychological test scores to give them normative meaning, linear and nonlinear transformations are used. The most common form of linear transformation is the **z score** and the most common form of nonlinear transformation is the **percentile**. Percentiles have been widely used in education and for some users are more intuitively understandable than z scores. For that reason percentile equivalents are sometimes given, even when the basic transformed score being used is the z score. It is necessary to understand both and the relationship between them that holds when the distribution of scores being transformed is normal or nearly so. In what follows, we use the term 'percentile', although the term 'centile' is also used in the testing literature. The two have exactly the same meaning.

z score
a linear transformation of test scores that expresses the distance of each score from the mean of the distribution of scores in units of the standard deviation of the distribution

percentile
an expression of the position of a score in a distribution of scores by dividing the distribution into 100 equal parts; also known as 'centile'

Standard scores and transformed scores based on them

Students first encounter standard (or 'z') scores in discussions of the **normal curve**, a statistical distribution that has a characteristic bell shape and many interesting properties. A normal distribution is symmetrical about the mean, with half the scores below the mean and half above. When scores are specified in terms of their distance from the mean, the mean is 0 and the standard deviation is 1. The **standard score** is a way of specifying where, in a normal distribution, a score lies with reference to its mean. The procedure is simple—subtract the mean from the score and divide the result by the standard deviation:

normal curve
a bell-shaped distribution of scores that conforms to a particular mathematical

function that is a good approximation for random variables that cluster around a single mean

standard score
the distance of a score in a normal distribution from the mean expressed as a ratio of the standard deviation of the distribution

$$z = \frac{(X - M)}{SD}$$

If the number is positive, the score must be larger than the mean; that is, it lies in the distribution above the mean. If negative, the score is less than the mean and it lies below it in the distribution. The magnitude of the z score can be read as a proportion: how far the score is from the mean as a proportion of a standard deviation. Consider again the scores on the achievement test discussed earlier. Each can be converted to a z score, as shown in Table 3.1.

Table 3.1: Calculating z scores from raw scores

	Raw score	X – M	$Z=(X-M)SD$
	35	17.4	1.54
	20	2.4	0.21
	17	–0.6	–0.05
	11	–6.6	–0.58
	5	–12.6	–1.11
Mean	17.6		
SD	11.3		

The first score, 35, is equivalent to a z score of 1.54. This means that it lies just over one and a half standard deviations (1.54) above (positive) the mean, whereas the score of 11 is just over half a standard deviation (0.58) below (negative) the mean. The z score thus locates the individual score in relation to the mean of the distribution of scores, which is what we want for norm referencing purposes (i.e. where the individual's score lies with respect to those of others).

The z score transformation is linear, because we can write it in terms of the equation for a straight line:

$$Y = mX + c$$
$$Y = \left(\frac{1}{SD}\right) \times X + \left(\frac{-M}{SD}\right)$$

The weighting factor is $\frac{1}{SD}$ and the constant is minus the mean divided by the SD.

The z transformation is useful because if we can assume a distribution of scores is normal (or nearly so), the properties of the normal curve can be invoked in interpreting a z score (see the Technical Appendix). We can always calculate z scores from a raw score distribution, but their interpretation depends on being able to make a reasonable guess about the distribution from which they come.

The assumption of normality is a necessary one, otherwise incorrect inferences can be drawn. For example, if a distribution is badly skewed (i.e. scores are bunched towards the top or bottom end of the distribution), the actual proportion of cases derived from the normal curve tables will not apply. This becomes even more of a problem if two scores are being compared that are drawn from distributions skewed in different directions. However, because many psychological variables come from distributions that are sufficiently close approximations to normal, this necessary assumption is less limiting than it might first appear.

The z score is the basic linear transformation used in psychological testing, but often transformed scores are not expressed simply as z scores. The reason for this is that z scores are 'untidy' numbers, with negative as well as positive signs in front of them and decimal fractions following them. By a further linear transformation of the z score (which by definition leaves the z score distribution unchanged in just the same way as the z transformation left the original raw score distribution unchanged), a tidier set of numbers can be produced.

Instead of having the mean at 0 with a standard deviation of 1, which is what we have with z scores, we can set the mean to be, say, 100 and the standard deviation to be 15 and adjust all scores accordingly. The equation of the straight line is again of use:

$$Y = mX + c$$

But the X now is read not as the raw score, as it has been up to now, but as the z score we have calculated, and the weighting factor is the new standard deviation and the constant is the new mean:

$$\mathrm{Y}=15 \times \mathrm{z}+100$$

We have transformed a transformed score, but again linearly. You might try transforming the z scores in Table 3.1 using this equation. If you ignore the decimal points in your new transformed score, a regular set of numbers results. Plotting the new set of numbers against the z scores results in a straight line and the equivalence of differences in the new and the z score distribution is maintained.

Rather than calculate the z scores and then transform them to the new distribution with the mean at 100 and a standard deviation of 15, we could do this in one step, again using the equation of a straight line:

$$\mathrm{Y}=\mathrm{mX}+\mathrm{c}$$

$$\text{NewScore}=\frac{\text{SD}_{\text{NewScore}}}{\text{SD}_{\text{OldScore}}} \times (\text{OldScore - Mean of OldScores}) + \text{Mean of NewScores}$$

For example:

$$100 = \frac{15}{1} \times (0 - 0) + 100$$

This is the procedure used originally by Wechsler (1955) in developing his Adult Intelligence Scale (now the WAIS–IV). He expressed an individual's score as a z score using the mean and standard deviation from a sufficiently large age-appropriate sample and then transformed these z scores to a distribution with a mean of 100 and a standard deviation of 15. He selected the mean to be 100 because an earlier formulation of scores on Binet's test of intelligence by Terman

led people to think of the average IQ as 100. Wechsler used the term **deviation IQ** to capture the essential link between his metric for intelligence and the z score.

deviation IQ
a method that allows an individual's score to be compared with same-age peers; the score is reported as distance from the mean in standard deviation units

Within his original adult intelligence test, Wechsler used the z score to describe performance on each subtest (of which there were initially 11). In this case, a reference group of 500 cases aged 20 to 34 years was used to furnish a mean and standard deviation, and the z score so computed was transformed to a distribution with a mean of 10 and a standard deviation of 3. Wechsler used the term **standardised score to** describe this form of z score.

standardised score
a score based on a z score, but set to a distribution with a particular mean and standard deviation considered convenient for a particular purpose

Other test developers have used the z score as the basis for a transformed score. Hathaway and McKinley (1951) in developing the MMPI derived **T scores** as the way of expressing aggregate response on each of the subscales of the test. (Being a personality test, there are no right or wrong answers in the way there are on an ability or intelligence test. Answers either indicate the personality characteristic or interest or they do not, and the raw score total for each subscale is thus the sum of responses indicative of the personality characteristic of interest.) The mean of the T distribution was set at 50 and the standard deviation at 10. A score of 60 on a subscale is thus 1 standard deviation above the mean. Dahlstrom and Welsh (1960) suggested that, as a rule of thumb, a score of 65 (1.5 standard deviations from the mean) or greater should be considered as unusually high.

T score
a score standardised to a distribution with a mean of 50 and a standard deviation of 10

One other variation of the z score was used by Cattell (1957) in developing the 16 PF. This is a ten-point scale with a mean of 5.5 and a standard deviation of 2, and is referred to as a **sten score** (an abbreviation of 'standard ten' score; see, for example, Cattell, 1957; Russell & Karol, 1994).

sten score
a point on a scale that has 5 units above and 5 units below the mean, which is set at 5.5 with a standard deviation of 2

Percentiles and transformed scores based on them

The z score, with its several transformations, is a widely used method of giving meaning to total raw scores obtained from a psychological test. Almost as popular is the nonlinear transformation known as the percentile. This should not be confused with a percentage correct score, which is just the expression of the raw score as a proportion of the total possible score.

The percentile scale expresses each raw score in a distribution in terms of the percentage of cases that lie below it. Thus a raw score at the 50th percentile is larger than 50 per cent of the raw scores in the distribution of scores, a score at the 63rd percentile is larger than 63 per cent of cases, and so on. Note that the percentile does not indicate the percentage correct on the test but the percentage of cases below the given value of the raw score. For example, a raw score that represents 50 per cent correct on the test would fall at the 63rd percentile, if 63 per cent of those tested obtained scores lower than 50 per cent correct. Percentile and percentage correct are separate concepts and must not be confused.

The term 'percentile point' is sometimes used to describe the point in the raw score distribution and the term 'percentile equivalent' to refer to the percentile score that expresses the raw score. That is, the raw score has a percentile equivalent, which is the point on the percentile scale. The distinction is correct but too subtle for most users, who recognise a raw score total and the percentile corresponding to it. The term 'percentile rank' is more widely used and refers to the percentage of scores that fall below the percentile point.

The value of the percentile scale is that it allows scores to be ranked in such a way that their position in the distribution is immediately apparent. A percentile rank of 80 expresses a score that is larger than a percentile rank of 70, but as well as knowing this we know that 80 per cent of the cases lie below the percentile rank of 80, and 70 per cent below the rank of 70. That is, we know approximately where in the distribution the scores lie, as well as their standing relative to each other. Because of its intuitive appeal, the transformation is popular in educational and psychological measurement.

It must be recognised, however, that the transformation is non-linear: it is not based on the equation of a straight line and it does not therefore preserve the equivalence of distances between scores in the raw score distribution. Scores in the middle of a normal distribution of scores are stretched apart on the percentile scale, whereas those at the tails are pushed closer together to form what is sometimes called a rectangular scale. Think of it this way: in a distribution of

scores that approximates a normal bell-shaped distribution, scores in the middle of the range of scores occur more frequently (by definition) and therefore we do not need to move far along the score range to aggregate any fixed percentage of scores: say, 10 per cent. By comparison, in the tails where scores are less frequent we need to move further to aggregate the same fixed percentage of scores. Therefore, scores that are an equal number of percentiles apart are not necessarily an equal distance apart in the raw score distribution. In comparing differences in percentiles, we need to bear in mind where the percentiles are on the percentile scale.

There are several ways of calculating percentiles, aside from using computer software. Two of these are described in the Technical Appendix. A third way involves computing the z score for each raw score and then, if one can assume a normal distribution or one that is nearly so, reading from the tables of the normal curve the proportion and hence the percentage of cases below each particular z score. Use a table like A1 (see the Technical Appendix) and read from column 2 (for z score above the mean) or column 3 (for z scores below the mean) the proportion (and hence percentage: proportion times 100) of cases below that z value. Although the z score is being used here to compute percentiles, it does not follow that the percentile distribution is a linear transformation. A check of some actual values will show that equal distances between z scores do not correspond to equal differences in percentages of cases. For example, a difference of 0.2 between the pairs of z scores in Table 3.2 does not convert into the same difference when the pairs of scores are expressed as percentiles. (Note you can verify the first transformation of z scores to percentiles by consulting A1.)

Table 3.2: Equal differences in z scores do not mean equal differences in percentiles

	z score	Percentile
	0.25	60
	0.45	67
Difference	0.20	7
	2.00	98
	2.20	99
Difference	0.20	1

The fact that z scores can be used to calculate percentiles indicates the close relationship between the two, a point that is discussed further below. Because whatever method is used to calculate percentiles is tedious (unless a computer is

used), a test developer typically publishes a table of percentile equivalents for all possible raw scores on the test as part of the test manual. The user simply consults the table to find the percentile equivalent. The user must of course understand what a percentile is, and its limitations, to make intelligent use of the table.

The link between z scores and percentiles has led some test developers to use this to 'normalise' non-normal distributions of test scores. The normal distribution, as noted earlier, is a desirable distribution for test development because of the known properties of the normal distribution and because many psychological variables are distributed in a nearly normal way. Where a distribution of test scores departs from a normal distribution, some test developers are inclined to force the distribution into a normal form.

Normalised standard scores constitute an easy way of doing this. The first step is to determine percentiles for scores in the raw score distribution and then to calculate the z scores that correspond to the percentiles using, say, the tables of the normal curve. The process is the reverse of that used earlier in finding percentiles using z scores. In the earlier case, the user enters the tables of the normal curve with a set of z scores (calculated from the raw score distribution) and reads off the proportion of cases associated with each of these to express them as a percentage of cases below each; that is, a percentile rank. In the case now being considered, rather than enter the tables of the normal curve with z scores and read off percentages (proportions), the user enters the tables with percentiles (calculated from the raw score distribution) and reads off z score equivalents. The z scores are, by definition, normally distributed; however, because the starting point is with percentiles (a nonlinear transformation), the normalised standard scores that result from this procedure are nonlinear transformations of the original raw scores, and the limitations of this must be recognised.

normalised standard score
a score in a distribution that has been altered to conform to a normal distribution by calculating the z scores for each percentile equivalent of the original raw score distribution

A variant of the percentile that is used by some test developers is the **stanine** scale. This was developed to facilitate recording of scores because it required only nine numbers, all single digits, to describe all possible raw scores. (This was of value when recording and manipulating scores was done manually.) The stanine (or 'standard nine') scale grouped percentiles into bands and assigned the numbers 1 to 9 to these bands, as shown in Table 3.3. The stanine distribution has a mean of 5 and a standard deviation of approximately 2. Stanines (a nonlinear transformation) should not be confused with stens (a linear transformation), discussed earlier.

stanine
a score on a nine-point scale with the points set in terms of percentiles

Table 3.3: Percentile ranges corresponding to stanine scores

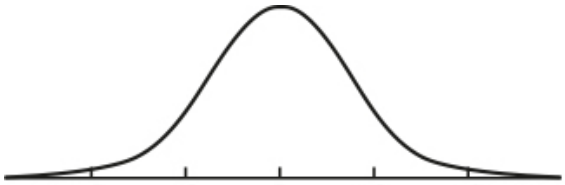
Stanine	Percentile range
1	Up to 4th percentile
2	4th to 11th percentile
3	11th to 23rd percentile
4	23rd to 40th percentile
5	40th to 60th percentile
6	60th to 77th percentile
7	77th to 89th percentile
8	89th to 96th percentile
9	96th percentile and beyond

Based on Allen & Yen (1979)

Relationships among the transformed scores

As noted above, there is a link between the linear and nonlinear transformations of raw scores in popular use, and that link is the normal distribution. Where distributions are normal or nearly so, the properties of the normal curve can be used to express any particular raw score in that distribution in terms of its deviation from the mean in units of the standard deviation, and then the resulting z score can be referred to the tables of the normal curve. From these tables, the proportion of cases corresponding to a z score can be determined and this can be expressed as a percentage of cases below that point; that is, a percentile. Scores can be expressed using a value for a mean and a standard deviation of a user's choosing without distorting the essential meaning of the z score. Thus, deviation IQs, standardised scores and T scores are all variations on the basic idea of the z score. Figure 3.2 demonstrates this relationship among transformations.

Figure 3.2 Relationships among linear and nonlinear raw score transformations



–2	–1	0	+1	+2	z score
70	85	100	115	130	Deviation IQ
2	16	50	84	98	Percentile
30	40	50	60	70	T score
1.5	3.5	5.5	7.5	9.5	Sten

Other methods of scoring
Using difficulty level to score ability tests

Another way to approach scoring of test items, particularly tests of cognitive ability, is to use the difficulty level of the item. In developing the first intelligence test, Binet and Simon drew on this approach when they arranged their test items in order of difficulty. In completing the test, a child would be administered the items in order of increasing difficulty, and when items were reached where the child began to fail, an estimate of ability was obtained. Difficulty level of the item thus becomes a means of assessing ability.

Item response theory makes use of item difficulty in determining a person's score. In the case of the **Rasch model** (see also the Technical Appendix), items are scaled for their difficulty level using a mathematical model known as the logistic function. A person's position (θ) on the trait assumed to influence response to the item and the difficulty of the item (b) are related to the probability of response to the item (in the case of a dichotomously scored ability test, passing the item). Specifically, the difference between a person's position on the trait and the difficulty level of the item are related to the odds of the person passing the item (odds = $e^{(\theta - b)}$). Odds, as in horse racing, is the probability of passing the item (the horse winning the race) divided by 1 minus the probability; where the probabilities of passing and failing are equal the odds are even (50/50). The logarithm of the odds (to the base e) is then the difference between θ and b:

Rasch model
a model that relates the probability of response of a particular sort (e.g. right/wrong) to the difference between a person's standing on a latent variable and the difficulty of the item

$\ln(\text{odds}) = (\theta - b)$

An individual's θ can be found by determining those items where the odds of passing are 50/50, because here the log is zero and θ = b. Thus, if we know about the difficulty level of the items, we can find out about the ability level of the person.

The ln(odds) scale is a **logit scale**. Values in logits vary from plus to minus infinity but in practice a range of ± 3 to 4 logits is used.

logit scale
an equal interval scale that locates the person's standing on the underlying trait of interest in terms of the percentage of items they get correct on the test and the average difficulty level of the items

Tests based on the Rasch model (such as the Woodcock-Johnson III and the Stanford-Binet 5 discussed in Chapter 13) use the logit as the basic score unit. Just as transformations have been used with the z score to eliminate negative numbers and untidy decimals, so transformations have been used with the logit scale. In the case of the Woodcock-Johnson Test of Cognitive Abilities, which is based on the Rasch method, a transformation to a W scale has been applied:

$$W = 9.1024 \text{ logit} + 500$$

The constant 9.1024 tidies up the numbers and the 500 eliminates the zero and negative numbers. Ws greater than 500 are positive logits and those less than 500 are negative. The usual range is then approximately 430 to 550, or about 13 logits. When the difference between ability and difficulty is +20W, the probability of getting the item correct is 0.9, when it is +10W the probability is 0.75, when it is –10W the probability is 0.25, and when it is –20W the probability is 0.1.

The W scale and the logit scale on which it is based might seem esoteric, but they do have advantages. As an interval rather than an **ordinal scale**, differences between scores mean the same throughout the range of the scale (e.g. an increase from 350W to 360W is the same as that from 420W to 430W) and are not dependent on a reference group. Jo's z score on a norm referenced test might not show any change from age 5 to age 10 if the reference group all change by the same amount, but her W score on a Rasch-based test will, showing for example that her probability of passing an item might now be 90 per cent compared with 50 per cent previously.

ordinal scale
a scale that has the property of a nominal scale, but also identifies an ordering of objects in terms of the attribute

Age and grade equivalents

To examine a child's performance in terms of expected developmental change, age or grade equivalent scores are sometimes computed. The idea here is to refer the child's level of performance to the typical performance of children of the same age or grade level. The median age or grade score for a sample of children is set as the age or grade equivalent. If, say, for children aged 10 years the median score on the test of interest is 20, then a raw score of 20 is given an age equivalent of 10 years. In the case of grade equivalents, the median score for children of a given grade is the grade equivalent, say 7 in the case of children in grade 7. To continue with the example for grade, for children in Grade 6 the median might be 17 and a raw score of 17 is thus a grade equivalent of 6. Raw scores between 17 and 20 are given grade equivalents by interpolation. A raw score of 18 would have a grade equivalent of 6.3 (i.e. a third of the way between 6 and 7 because 18 is a third of the way between 17 and 20). The decimal place can be given a score in months (e.g. 3 months for a 10-month school year).

Although a legitimate way of expressing a score, authorities (Cronbach, 1990; Allen & Yen, 1979) warn against their misinterpretation for a number of reasons. For example, children at different ages have different levels of understanding and preparedness for different types of learning, even though they might have the same age or grade equivalent scores.

Expectancy tables

Where there is a single well-specified criterion of interest and it is possible to obtain a very large sample, a test score can be given meaning in terms of the likelihood of reaching a given point on the criterion. For example, the probability of successfully completing recruit training in the military could be studied in terms of scores on a selection test, with the frequency and percentage of a successful outcome tallied for recruits at each score level on the test. For the percentages to be reasonably stable, the numbers of recruits at each score level need to be large (e.g. 100+), but if they are, then a score on the test can be directly interpreted in terms of probability of success. Not many situations, however, meet the requirements for use of **expectancy table**s.

expectancy table
a table that presents the probability of an outcome on a criterion of interest in terms of score on score range on a test

Norms

In all transformations of raw scores that use the idea of the z score, the mean and standard deviation that are used are critical to the meaning that is given to the score. In norm referencing, the raw score is referred to a relevant group for comparison purposes. If the comparison is not with an appropriate group, the transformation, although technically correct, fails to convey meaning or, worse yet, opens the score to misinterpretation. To say that a person's score is 1 standard deviation above the mean only has value if the mean is for a group the person is like in some way. For example, to say that an adult's score on reading is 1 standard deviation above the pre-schoolers' mean for reading does not usually convey any information because we would expect this, unless we had some reason to suspect severe educational disadvantage. The pre-schooler group is not an appropriate point of comparison in most cases.

It follows that selecting an appropriate reference distribution and ensuring that the mean and standard deviation are well estimated are essential aspects of the norm referencing approach. What constitutes an appropriate reference group may vary even for the same test taker from time to time, depending on the interpretation of the individual's test score that is to be made. In testing for aptitude for business, for example, one might want to compare the individual's score to that of all students of the same age, or to just those students of the same age who are interested in business, or to business people who have a reputation for being good at business. The reference group that is appropriate varies depending on the use to be made of the score. More than one group is usually studied in preparing norms so that a number of interpretations become possible and a number of different individuals can be evaluated with the test.

For characteristics that vary with age, age norms are valuable. Intelligence tests are normed for various age ranges in the population. Where intelligence is known to increase rapidly, as in young children, the appropriate age range for norms might only be a year, whereas for adults the age range could be 10 years. A related concept is that of grade norms, which might be useful for some forms of educational tests. Although grade usually depends on age, it can be useful to develop norms for grade levels to allow comparisons of, say, reading ability.

The issue of an appropriate norm group helps make the point that the test score is not an immutable fact of the person, but a sample of performance that needs to be interpreted in an appropriate context. The same performance can lead to different interpretations depending on the context. A score that earns the

judgment that a student is in the top 20 per cent of their classmates in terms of business aptitude might lead to a judgment that they are below average if the reference group is that of a group of experienced business people.

Deciding on an appropriate norm group is an exercise of judgment, which takes into account the uses to be made of the test. Not all uses can be anticipated and the test user might be left with a situation for which an appropriate norm group is not provided. In such a case the test score needs to be interpreted cautiously or no interpretation should be offered and an alternative test with appropriate norms sought (see Box 3.1).

Box 3.1

Using US norms with Australian populations

Test development is expensive and good psychological or educational tests require a substantial market for a test to justify the investment in its development. Not surprisingly, many of the tests used in Australia and other countries in the region are developed overseas, principally the USA, and used here with only minor modifications. For example, a question on a general knowledge test that asks the name of the US president might be altered to ask about the Australian prime minister. There is usually some pilot work done with the altered item to check that it is performing as expected, but there is seldom any large-scale examination of a test in its new cultural environment.

An exception was the development work in the Macquarie University Neuropsychological Normative Study (MUNNS; Carstairs & Shores, 2000) in which 399 healthy young adults from the Sydney metropolitan area were tested on a battery of neuropsychological tests used for rehabilitation and medico-legal assessments. The test battery comprised eleven tests, including the Wechsler Memory Scale–Revised (Wechsler, 1987), Rey Auditory Verbal Learning Test (Lezak, 1995), Wechsler Adult Intelligence Scale–Revised (Wechsler, 1981) and the Depression, Anxiety and Stress Scales (Lovibond & Lovibond, 1995a). A stratified random sampling plan ensured the representativeness of the sample in terms of age, gender, language background, socio-economic status and level of education. Participants were screened for prior head injury resulting in loss of consciousness, use of certain therapeutic or recreational drugs, inability to understand English, and physical or intellectual disability that interfered with performance on the tests. Over 10,000 people were contacted in order to find sufficient numbers to participate in the study, giving some idea of the effort required to produce a set of good-quality **local norms**.

local norms
norms developed for specific population groups or geographical regions

Where this is not done there would seem to be a serious problem, given that we have made much in this chapter about using the correct norms for interpreting test results and, indeed, commentaries appear in the professional literature (e.g. McKenzie, 1980) from time to time criticising the use in Australia of psychological tests with US norms. The reason should be obvious from this chapter; but to give a concrete example, consider the situation in which the Australian mean on, say, a test of intelligence is in fact higher than that of the mean for the US population. In these circumstances, a score for an individual on the test could be below the Australian mean but still above the US mean. The probability of this occurring increases as the distance between the Australian and US means increases. Consider now that an individual with a score sufficiently below the mean is eligible for some special form of intervention—for example, remedial education—and that failure to receive it is to their disadvantage. If the US norms are used in this situation, the person's score will be interpreted as being above the cut-off, whereas if Australian norms were to be used it could well be that their score might be sufficiently below the mean to warrant their access to the special program. In this situation, testing with the US norms has done the individual a disservice.

How likely is this scenario to occur? There are some data that point to Australian means on tests like the Wechsler differing from those reported in the US standardisation samples (see Holdnack et al., 2004). As Holdnack et al. point out, however, the Australian samples on which these observations are based are typically small, and unrepresentative in terms of the sampling design used for their collection. Where large samples with better claims to representativeness are employed (Howe, 1975), the means for Australian samples are on most factors of cognitive ability close to those reported for large US samples. Given the similarities in language and media exposure of the Australian and US populations, to find otherwise would be surprising. This is not, however, an argument for complacency. Where differences exist between cultures—for example, in educational practices—there is reason to expect differences in means between US and Australian samples on some characteristics.

In the light of this discussion, we suggest the following rules of thumb:

1. Check the source of the norms for any test that one is using or for which one is evaluating the results.
2. Ask whether the norms are relevant to the situation in which the test is being used or to which results might be generalised.

3. If there is concern about their relevance in terms of country of origin, ask what is known about the susceptibility of the measure to cultural differences.
4. Consider how the test result is being used. Is it being used with reference to a cutting score for describing the individual or determining a course of action with respect to the individual, and is it the primary or only basis for this?
5. Ask whether it is possible to check the result in some way using another test for which norms are available or by reference to non-test information, but beware small and unrepresentative samples.
6. Explain in any report the basis for the description or recommended action in terms of the norms employed, and any qualifications that should in prudence be considered.

Selecting an appropriate reference group (or groups) is the primary decision, but once made there is a need to ensure that the mean and standard deviation that are determined for the group are accurate. Accuracy depends on two principal considerations: the manner by which the sample is drawn from the population in question, and the size of the sample.

Sampling is a technically complex matter. A distinction is usually made between probability and non-probability methods of sampling from a population. Probability methods increase the likelihood of the sample matching the population in all respects that are important to the researcher and permit the calculation of the degree of precision in estimating a parameter of interest in the population (e.g. the average IQ). **Random sampling** is a case in point. Here members are drawn from the population but in such a way that every member of the population has an equal opportunity of being selected and the drawing of one member does not influence in any way the likelihood of any other member being selected. Non-probability methods, on the other hand, might produce a biased estimate of the parameter and the precision of the estimate is unknown. Non-probability methods include accidental or convenience sampling. In these cases, a sample is gathered in a way that is easy or convenient to do. For example, the test developer might stop individuals in a shopping mall and ask them to participate. There is no way of knowing how representative such a sample is or even what population it might be a sample of (people who frequent shopping malls, possibly). Although probability methods have clear advantages, they are expensive to implement and might in fact not be practical. For example, if the population cannot be specified or compliance of those sampled cannot be guaranteed then a probability sample cannot be obtained.

random sampling
a procedure in which every member of a population of interest has an equal probability of being selected and the selection of one member does not affect in any way the selection of any other member

In norming psychological tests, best practice is to draw a sample from a population in such a way that it matches as closely as possible important characteristics of the population. For example, in norming his intelligence test, Wechsler (1955) considered the major demographic characteristics that research had shown relate to intelligence (e.g. age, gender, education level, geographic region of residence and ethnic background) and sought a sample that resembled the population (all US citizens) in these respects. That is, the sample was to have the same age distribution as the population, the same distribution of educational attainment and so on. A sampling plan was drawn up that specified the number of participants with the selected characteristics that needed to be included for the sample to match the population as determined from the most recent US Census at that time (e.g. x per cent of white, high-school-educated males from the south-west of the USA aged between 25 and 34 years; and y per cent of black American females, from the north-east, with college level education, aged 35 to 45 years). Wechsler's research assistants were then dispatched to interview the number required by the sampling plan. This was an ambitious attempt to ensure that the sample on which the norms for his test were to be based resembled the population in terms of factors that influence the construct being measured. Not all norms are as well based as these, largely because of the considerable cost involved.

Wechsler's method was an approximation to the probability sampling method termed stratified random sampling. Sampling was not, however, random in that a number of decisions intervened—such as which cities or towns to include—and research assistants could only test those citizens prepared to volunteer. This sampling method is better described as **stratified sampling** or quota sampling, a non-probability sampling method.

stratified sampling
a method of sampling in which the sample is drawn from the population in such a way that it matches it with respect to a number of characteristics that are considered important for the purposes of the study

A further consideration in developing norms is the size of the sample that is employed. Size is important because the requirement is to estimate the mean and standard deviation with precision, and sample size has a potent influence on the standard errors of these statistics. In the case of estimating the mean, the standard error is proportional to the standard deviation of the distribution divided by the square root of the sample size.

As sample size increases, the denominator becomes larger and the standard error smaller. Note, however, that the effect is not a linear one: doubling sample size does not halve the standard error (or double the precision). It is not sample size but the square root of sample size that is the denominator. Thus, to halve the sampling error one must increase the sample size by a factor of 4. What this means in practice is that, beyond a certain point, increasing the sample size will have little discernible effect on the standard error. On the basis of these considerations, Bartram and Lindley (1994) proposed the rules of thumb outlined in Table 3.4 for evaluating samples for norming purposes.

Table 3.4: Bartram and Lindley's recommendations for sample sizes for purposes of test norming

Sample size	Evaluation
Under 200	Inadequate
200–500	Adequate
500–1000	Reasonable
1000–2000	Good
2000+	Excellent

With these considerations in mind, how well do some of the major psychological tests fare in terms of the norms they provide? Table 3.5 provides a brief summary.

Table 3.5: Some examples of sampling methods and sample sizes for widely used psychological tests

Test	Sampling method	Sample size
WAIS–IV	Stratified: age, gender, education level, race/ethnicity, geographical region	2200
MMPI–2	Convenience (seven US states)	2600
16 PF	Stratified: age, gender, education level, race	2500

One other consideration needs to be borne in mind in the use of norms for tests of general mental ability, and that is the age of the norms being used. The raw score mean on many of the commonly used tests of intelligence has been rising for

at least the previous half century for reasons that are not well understood at present. This increase has been named the 'Flynn effect' after the researcher who first observed it (see Chapter 7). What it means is that when a new or recently re-normed test is used to retest a person whose score has been previously determined, it might appear that the person's intelligence level is lower than it was. For example, testing with the WISC–III indicates on average a five-point drop in IQ compared with initial testing with the WISC–R, because the norming process for the WISC–III (the more recent test) has adjusted for the upward trend in intelligence over time (Kanaya, Scullin & Ceci, 2003). In view of the Flynn effect, the test user needs to be particularly vigilant in assessing what is and what is not a substantial change in IQ from one testing occasion to the next. If different tests with different norms are involved, the test user needs to allow for the Flynn effect in any inferences that are drawn from an apparent change in IQ.

Chapter summary

Test scores must be interpreted either by direct reference to the behaviour they reflect (criterion referencing) or by reference to the performance of other individuals with whom a comparison is appropriate (norm referencing). In the case of the latter, test scores are transformed linearly (e.g. the standard score) or nonlinearly (e.g. percentiles) to aid in interpretation. There are important differences between these two sorts of transformations that need to be borne in mind, but when the distribution of scores is normal or nearly so, one form of transformation can be expressed in terms of the other.

Questions

1. Define: raw score, scaled score, standard score, standardised score and percentile.
2. What are norms? Why are they needed for psychological testing?
3. What are the characteristics of a good normative sample for a psychological test?
4. Compare and contrast z score and percentile.
5. What is the relation between the sten score and stanine score scales?
6. What deviation IQ does a T score of 70 correspond to?
7. Find out the sample size used in norming the Wechsler Memory Scale, Wechsler Memory Scale–Revised and Wechsler Memory Scale–Third Edition, and evaluate them according to Bartram and Lindley's recommendation.
8. What reference point is used in evaluating a mastery test?
9. What is the value of having local norms for a test?
10. Why do authorities on testing warn about the use of grade equivalent scores?

Exercises

1. For the following set of raw scores:

   52, 54, 56, 58, 60, 61, 61, 63, 67, 68

   express each score as a z score, and then transform each to a score in a distribution with a mean of 100 and an SD of 15.

2. Assume that a large Year 10 class took achievement tests known to be highly reliable and valid in the areas of geography, spelling and mathematics. The scores on all three of these tests were normally distributed, but the tests differed in the following respects:

	No. of items	Mean	Standard deviation
Geography	75	60	10
Spelling	150	100	20
Mathematics	40	25	5

Assume that you are particularly interested in comparing the performance of three of the students who took these tests (Hassan, Brett and Zhang Wei). First, you are interested in how each student has performed across the three tests, which is his best performance and which is his worst. Second, you are interested in comparing students in terms of their performance on each test. Third, you want to identify the student who performed best across all areas. The students' scores are as follows:

	Hassan	Brett	Zhang Wei
Geography	46	72	60
Spelling	110	100	97
Mathematics	30	33	37

a. Prepare a table showing the percentage correct scores for each student on each test. Note, however, that because the percentage correct scores on each test come from different distributions, they cannot be justifiably averaged across tests or otherwise compared.

b. Prepare a table showing linearly derived z scores for Hassan, Brett and Zhang Wei. Note that although z scores can be averaged, the presence of decimals and negative values will make it more difficult to do so than it would otherwise be.

c. Using Table A1 (see Technical Appendix), determine the percentile equivalents for each z score.

d. Convert each of the z scores into T scores (mean = 50, SD = 10) and prepare a table showing them and showing the average T score for each student. The initial goal of obtaining scores that are

intra- and inter-individually comparable will have been achieved most suitably with this final step.

3. Assume that 100 students took a test and that the test scores were normally distributed with a mean of 20 and a standard deviation of 2.

   a. What are the z scores for the following raw scores?
   16, 18, 19, 20, 21, 22, 24

   b. Using Table A1, with the z scores you have just obtained, determine the percentage of the scores that fall between the following raw score ranges:
   18 and 22, 19 and 21, 16 and 24

4. a. The percentile of a score with a z score of 1.0 is _____

   b. The z score of a score at the 98th percentile is _____

   c. The T score for a score with a z score of 2.0 is _____

5. Given that a test has a mean of 30 and a standard deviation of 10, complete the following table.

Raw score	z score	Percentile
40		
	0.5	
		75

6. How many cases in a normal distribution lie between a z score of 1.0 and a z score of 1.15?

7. Before undertaking a reading enrichment program, Nehir and Tanya obtained scores on a Rasch-based test of reading ability of 490 and 510, respectively. On conclusion of the program Tanya has a score of 520 on the same test. What score should Nehir obtain to show the same improvement as Tanya?

   a. 520

   b. 499

   c. 500

   d. no comparison is possible

8. Paula scores 500 on the Woodcock-Johnson test of ability. Is it more or less likely that she will pass items with the following logit values of difficulty?

   a. +1.5

   b. +0.5

c. –0.2

9. What would the item difficulties in question 8 be if expressed in W units? What is Paula's probability of getting each of these items correct?

10. Pilot testing with a sample of 100 indicates that the mean score on the test has a standard error of 0.5. To halve this, what size sample is required, if it is assumed that the standard deviation remains the same? How does this new sample compare with the Bartram and Lindley guidelines?

Further reading

Osterlind, S J (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. Upper Saddle River, NJ: Pearson.
Reynolds, C R & Livingston, R B (2014). *Mastering modern psychological testing; Theory and methods.* Harlow, UK: Pearson.
Rust, J & Golombok, S (2008). *Modern psychometrics: The science of psychological assessment* (3rd ed.). London, UK: Routledge.
Wasserman, J D & Bracken, B A (2003). Psychometric characteristics of assessment procedures. In J R Graham & J A Naglieri (Eds.), *Handbook of psychology: Vol 10: Assessment psychology* (pp. 43–66). Hoboken, NJ: John Wiley & Sons.

Useful websites

Interpreting ACER test results: www.acer.edu.au/files/PATM-Interpreting-Scores.pdf
Norming and norm-referenced test scores: http://ericae.net/ft/tamu/Norm.htm

4 Reliability

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. explain why the concept of reliability is important in testing and assessment
2. specify the various ways reliability is estimated
3. define Cronbach's alpha and describe how it is calculated
4. define the standard error of measurement and explain how it is applied.

KEY TERMS

Cronbach's alpha
domain-sampling model
equivalent forms reliability
generalisability theory
inter-rater reliability
reliability
reliability coefficient
split-half reliability
standard error of measurement
test-retest reliability

Setting the scene

- A young man who has suffered a motorcycle accident is experiencing some memory loss and those managing his recovery want to track the severity of this problem using standard memory tests as he rehabilitates. They want to know what tests are most suitable for repeat testing.
- In a compensation case, evidence is presented that the litigant has shown a deterioration in scores on measures of planning ability from before to after the accident that is the source of the claim. The legal counsel for the insurance company asks for evidence that not only have the scores changed but that the change is greater than that to be expected had the accident not intervened.
- A psychologist notes that there is a discrepancy between a client's scores on measures of verbal and spatial ability taken from the same test of general mental ability and wonders whether the difference should be taken seriously.
- Two tests are being considered for use in a situation where decisions will have a substantial impact on what happens to those involved. Although the tests, on the data available, appear similar in many of their features, one is much shorter than the other and is being favoured for use for this reason.

Introduction

Psychological tests are used in a number of different situations and we need to know how appropriate they are to use across these different conditions. Answering this general question involves consideration of a number of issues and this chapter begins an examination of these. Chapter 5 takes the exploration further, but for the present we focus on the reliability of psychological tests for use for particular purposes. Note that we talk of the reliability of a test for a particular purpose and not of the reliability of a test in general, because the latter is not correct, even though we might find ourselves saying that for economy of expression. Reliability, as with validity (which we take up in the following chapter), is not a property of a test itself but a property of a test as used in a particular situation.

In this chapter we discuss the meaning of the concept of reliability, how it is expressed in particular situations, how it can be estimated, and how it can be applied in considering the appropriateness of a test when used in a particular way.

The meaning of reliability

The word **reliability** has as its ordinary meaning: dependability. To say that a person or a car is reliable is to say that they can be depended on. The person will be true to their word; the car will run and not let you down. The term reliability

in psychometrics has much the same meaning. To ask about the reliability of a psychological test score is to ask about how much it can be depended on. When, for example, an intelligence test yields a score for a person that indicates their mental ability to be well above average, how much confidence can we have in this finding? Because of the importance of reliability, theorists and researchers have paid a good deal of attention to working out how the reliability of a test can be determined. The important ideas that have emerged from this century-old exercise are the subject of this chapter.

reliability
the consistency with which a test measures what it purports to measure in any given set of circumstances

Tests, like cars, can be unreliable for two sorts of reasons. Imagine one of the tyres on your car deflates over the course of the day. The first time this happens you cannot use the car when you want to. But once you have become aware of the problem you can deal with it by building time into your schedule to pump the tyre up each day. Similarly, your battery gradually loses its charge but, again, you know you can arrange to have it recharged during the evening so that the car is drivable when you want it in the morning. These defects make the car unreliable but they are predictable, regular or systematic—once you become aware of them. Compare this with, say, a problem in the electrical system of the car that is difficult to trace and intermittent in its effects. The car is rendered inoperable at various times over which you have no control. This is an unsystematic source of unreliability.

So, too, psychological tests have both systematic and unsystematic sources of unreliability. Unlike cars, however, the systematic sources of unreliability in a test can be hard to detect unless a lot is known about it. The test might appear to be functioning well, but is not really testing what you want it to be testing. You might think, for example, that your test is one of 'anxiousness', but what it is really testing is a mixture of anxiousness and the desire of the test taker to present himself or herself in a favourable light to the tester—a concept known as **social desirability bias**. The test is systematically wrong in the assessment of anxiousness because of the confounding with social desirability. This problem is taken up again in Chapter 5 when we examine threats to the validity of tests. The major concern of the present chapter is with the unsystematic sources of unreliability in tests.

social desirability bias
a form of method variance common in the construction of psychological tests of

personality that arises when people respond to questions that place them in a favourable or unfavourable light

The domain-sampling model

One of the earliest ways of thinking about the problem, and one that is still of considerable value, is founded on the idea that a psychological test is a sample of responses or behaviours from a much wider population of responses (Nunnally, 1967). For various reasons, not the least of which is practicality, an assessor interested in an individual's status on some trait or condition can only ask a limited number of questions or present a limited set of tasks. The test or assessment device thus draws from a larger possible set of items to give a score for the person on the trait or condition. It is recognised that, because the sample of items is limited, the score obtained is an estimate of the person's actual or true position on the trait rather than a direct expression of that position. If all possible questions had been asked we would have the true position, but what we have, in fact, is a sample of questions and hence an estimate that is likely to be in error. The important issue is how good an estimate do we have? That is, how close is the score obtained from the sample to the score that would have been obtained if all possible questions had been asked?

Put this way, the question of test reliability becomes a problem of sampling; however, it is not one of sampling people from a specified population, but of sampling items from a domain of all possible items. The **domain-sampling model**, as it is called (Nunnally, 1967), is one important way of thinking about the question of reliability. In applying it, we can think of the score a person receives on a test as one of the scores that would be obtained if samples of the items were put to the person repeatedly. That is, imagine drawing a finite set of items, say 20, from the domain of all possible items that might be asked, and presenting them to the test taker. Once those have been completed, you draw another sample of 20 items and administer those, and then another 20, and so on. The patience of the test taker is not infinite and this is an illustration of what is implied by the model rather than the beginning of a real study. The scores obtained from each of those 20 item tests would not be the same; there would be some variation due to sampling. The mean of the scores from all possible samples would tell us, however, the true position of the person on the trait in question; the person's 'true score' as it is called in classical test theory (see the Technical Appendix). The standard deviation of the distribution of scores from all possible samples about the true score would tell us about the likelihood of obtaining any particular sample score. It is referred to as the standard error of measurement and indicates the precision of our estimate of the true score.

domain-sampling model
a way of thinking about the composition of a psychological test that sees the test as a representative sample of the larger domain of possible items that could be included in the test

The situation is hypothetical but serves to illustrate the essential idea. In practice, we have only samples and the true score eludes us, but we can use what we know from the sample to make estimates of the likely true score for an individual, and the interval in which it lies, with a stated degree of confidence. If the interval is very large, clearly we have a great deal of imprecision in the measurement process and we cannot depend on any score we obtain with this sample of items. The value of thinking about the problem in this way is that it leads to two quantitative indexes of reliability that allow us to be more precise than verbal labels allow. To say that a test is 'not very reliable' or has 'satisfactory reliability' is to make a statement that is open to misinterpretation. Quantitative indexes, when their meaning is understood, provide for a more precise form of communication.

One of the indexes we have encountered already is the **standard error of measurement** (SEM). The other is the **reliability coefficient** (r). They are intimately related but serve slightly different purposes in practice. Their actual mathematical derivation is beyond the scope of this chapter, but the interested reader is referred to Nunnally (1967) for a statistical treatment of domain-sampling theory and its implications. The relationship between the two indexes is:

$$\mathrm{SEM} = \sqrt{(1 - r)}$$

for scores expressed as standard normal deviates (z scores). For ordinary scores we simply multiply the right-hand side of the equation by the standard deviation of the obtained score distribution. We leave it in deviation score form for the present to illustrate the essential relationship.

standard error of measurement
an expression of the precision of an individual test score as an estimate of the trait it purports to measure

reliability coefficient
an index—often a Pearson product moment correlation coefficient—of the ratio of true score to error score variance in a test as used in a given set of circumstances

The fact that drawing samples repeatedly from a domain gives rise to variation in obtained scores can be understood in terms of the mixture of true score and error score variability that makes up the observed score. When this variability is defined in a particular way (as variance; that is, the sum of the squared deviation of each score from the mean of the scores), we can say that the observed score variance is equal to true score plus error score variance. (We are assuming here that there is no relation between true and error components of the observed score.) How much true score variance makes up the observed score variance is of course of considerable interest to us. If it constituted the whole, our measure would give us the true score we ideally want to know and we could claim it was perfectly reliable. We can define the reliability coefficient, then, as the proportion of observed score variance that is due to true score variance. In practice, the proportion will be less than 1.0 and in some cases a good deal less. If the proportion is only 0.5 (i.e. r = 0.5), 50 per cent of the variance in the scores obtained with the test is due to variance in true scores and the other 50 per cent to errors of measurement. For some, 0.5 would be a minimal level of reliability, beyond which point the test is reflecting more of what we are not interested in than what we are.

If we return to the formula above, we see that, in the unlikely situation that r = 1.0 (perfect reliability), the SEM is zero; that is, there is no error in estimating the true score. If, on the other hand, the proportion of true score variance is zero (r = 0) then the SEM = 1, which is the standard deviation for a standard normal distribution. That is, our obtained score gives us no more information about the true score than any other score we might have obtained at random.

In practice the two indexes have different applications. The reliability coefficient is, in general terms, used in forming judgments about the overall value of a particular test (e.g. is this a better test for some given purpose than another test?), whereas the standard error of measurement is used in making judgments about individual scores obtained with the test (e.g. how much error might be associated with this score as an estimate of the trait in question?). The reliability coefficient is determined from data obtained with the test and the standard error is then calculated using the above formula.

Calculating reliability coefficients

The reliability coefficient is determined in three main ways. The oldest is in terms of the correlation between equivalent forms of the test. Knowing that the

problem of reliability has to be faced at some stage, the test developer from the outset devises two tests rather than one; that is, they draw two samples from the domain of possible test items. The two forms will have the practical benefit of minimising practice effects if, subsequently, a person is to be tested on two separate occasions, because one form can be used on the first occasion and the equivalent form on the second. For present purposes, however, the existence of equivalent forms allows the test developer to examine how well two samples of items from the same domain agree in the scores they yield. This is a far cry from all possible samples, but it is a good beginning. If the two samples do not yield comparable scores, the test, or at least one form of it (although we do not know which one), cannot be depended on.

The product moment correlation between scores on equivalent forms of the test for a reasonably large sample of test takers gives an estimate of the reliability coefficient. If equivalent forms of a test are not available, the reliability can be calculated by splitting the test into two equivalent forms; for example, all the even numbered items are used for one form and all the odd numbered items for the other. By correlating the scores obtained on the two halves for a sample of test takers of reasonable size, one again obtains an estimate of reliability. This is called the **split-half reliability**. The coefficient is usually corrected for the fact that, when the test is used as a whole, it is twice as long as either of the two halves and larger samples are better estimates of a population mean than smaller samples. (When the sample is the same size as the population, its mean is the population value.) The formula for estimating the reliability of a test that is longer than the original test by some factor is given by a formula named after the two people who derived it independently of each other: Spearman and Brown. The Spearman-Brown formula is sometimes termed the Spearman-Brown prophecy formula because it purports to tell us about an otherwise unknown state of affairs. The formula is discussed later in this chapter.

split-half reliability
the estimate of reliability obtained by correlating scores on the two halves of a test formed in some systematic way (e.g. odd versus even items)

How to split the test into two to determine its split-half reliability is something of a problem. The odd-even method is a practical solution that at least ensures that any factors that might influence scores late in the test (e.g. fatigue) have an equal influence on both halves. Even with this proviso, however, when speeded tests are being examined (those that must be completed within a time limit), this method of estimating reliability is not recommended. But the odd-even method is arbitrary and different reliability estimates can result from the one test split in different ways. In terms of the domain-sampling model, this

outcome is not at all surprising because each sample provides only an estimate and estimates are likely to vary.

Cronbach (1951) suggested one way round the problem. He proposed that the test be split into subtests, each one item in length; that is, think of the test as made up of k tests, where k is the number of items in the test. All subtests are then correlated with all other subtests and the average correlation calculated. This average correlation becomes the estimate of reliability. This method is often described as determining the internal consistency of a test. The formula for calculating it is simple (see Box 4.1) and is referred to as **Cronbach's alpha**. It is the same as a formula, arrived at in a somewhat different way, by Kuder and Richardson and known after them as the KR20 formula (Kuder & Richardson, 1937). Because Cronbach's alpha is so easy to calculate, with a program for it provided in major software suites, it is frequently used to determine reliability. It does, however, have its limitations.

Cronbach's alpha
an estimate of reliability that is based on the average intercorrelation of the items in a test

Box 4.1

Cronbach's alpha

Assume a five-item true/false test has been administered to ten participants, who have responded as shown in the Table 4.1, where 1 indicates a 'true' response and 0 indicates a 'false' response.

Table 4.1: Calculating Cronbach's alpha for a five-item test

	Item							
Person	1	2	3	4	5	Total score	(x – M)	(x – M)2
1	1	1	1	1	1	5	2.5	6.25
2	1	1	1	1	1	5	2.5	6.25
3	1	1	1	1	1	5	2.5	6.25
4	1	1	1	1	1	5	2.5	6.25
5	1	1	1	1	1	5	2.5	6.25

	Item							
6	0	0	0	0	0	0	–2.5	6.25
7	0	0	0	0	0	0	–2.5	6.25
8	0	0	0	0	0	0	–2.5	6.25
9	0	0	0	0	0	0	–2.5	6.25
10	0	0	0	0	0	0	–2.5	6.25
Sum								62.5
Mean	0.5	0.5	0.5	0.5	0.5	2.5		
Variance	0.25	0.25	0.25	0.25	0.25			6.25

The data are artificial but serve to illustrate a point. Note from the table that all items are consistent in the responses they elicit across participants. For half the sample, all items elicit a 'true' response and for the other half a 'false' response. Knowing how an individual has responded to one of the items means we know how they have responded to all other items. Cronbach's alpha is calculated using the standard formula. It requires that we know the number of items in the test (five in this case), the items' variances, and the variance of total score on the test. The variance of a dichotomously scored item is simply the product of the proportion of individuals (p) who answer in one way (say, true) and the proportion who answer in the opposition direction (1 – p). The variances are shown for each item. The variance of total score on the test is calculated in the usual way for calculating the variance for a sample (not the population estimate that uses N – 1). It is shown as the average of the squares of the deviations of each total score from the mean of the total scores.

When the formula is applied to the calculated values, alpha is shown to be 1. That is, the test shows perfect internal consistency, which is not surprising as the exercise was designed with this in mind.

Alpha is thus given by the following formula:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum \sigma i^2}{\sigma t^2}\right)$$

where:

α is coefficient alpha
k is the number of items in the test
$\sigma_i 2$ is the variance of an item
$\sigma_t 2$ is the variance of total score on the test.
In this example,

$$\begin{aligned} \alpha &= \left(\frac{5}{5-1}\right)\left(1 - \frac{1.25}{6.25}\right) \\ &= 1.25 \times [1 - 0.2] \\ &= 1 \end{aligned}$$

One limitation is that a test can be developed to have a high internal consistency by having items with highly similar content. Although faithfully sampling a domain, the domain itself might be so constricted as to be trivial. On the other hand, high internal consistency does not in itself guarantee that the items are all reflecting the one thing. It means that the items are interrelated but not that they are homogenous; or, as it is technically referred to, unidimensional (Hattie, 1985). If there are multiple factors (traits) underlying performance on the test, alpha can overestimate the reliability of the factor thought to underlie the test (the one referred to in the label on the test), because alpha estimates the reliability of the labelled factor and all other factors being measured. This has led some authors (e.g. Yang & Green, 2011) to argue that it is better to approach the question of reliability through the methods of confirmatory factor analysis (CFA; see Chapter 5) and assess whether the items conform to a single-factor or congeneric model of measurement. What is being measured by a test with high internal consistency needs further investigation, which is the problem of validity to be discussed in the next chapter. High internal consistency is an important attribute in a psychological test but it is not of itself a 'seal of approval'.

It needs to be pointed out again that the statements being made about reliability and standard error are for the test when used in a particular way. A test, once constructed, is not reliable in all situations in which it could be used. The variance of observed scores on the test is likely to differ depending on the particular sample of individuals we choose to study, and we cannot assume that the reliability will remain constant across different samples. Studying creativity in a group of people purposely selected because they are high in intelligence means that variability in creativity is likely reduced compared with an unrestricted sample from the population. Artificially restricting the range of scores on

creativity means that reliability will need to be calculated again. It is better to think of the reliability coefficient and the SEM as applying to a test when applied to a particular type of sample and not as a property of the test itself. This might seem a highly conditional way of speaking, but it is more accurate and makes us pay particular attention to the circumstances under which the claim of reliability is being made.

Extending the domain-sampling model

Thus far we have talked about the domain-sampling model and the ways this leads us to assess reliability. Although very important, it is not the last word on reliability and some other ideas about it need to be understood.

The first is that reliability can be thought of in relation to the time of testing. Having determined a score today, how likely is it that the same score would be obtained by the test taker if the test were administered tomorrow, next week, in a month, or in 12 months' time? **Test-retest reliability** is a long-standing approach used by researchers seeking to evaluate reliability because its meaning is intuitively obvious. If the characteristic we are attempting to measure is in fact likely to be stable over time (e.g. mental ability is likely to be stable, but mood is not because by definition it varies from day to day or even within the same day), then scores obtained on two different occasions should correlate highly, and reliability can be assessed by the product moment correlation between test scores on two occasions. To the extent that the two sets of scores do not correlate, the test lacks reliability.

test-retest reliability
the estimate of reliability obtained by correlating scores on the test obtained on two or more occasions of testing

Note that, in this case, the sample of items employed is the same on the two occasions and hence there has not been sampling from a domain of items as required by the domain-sampling model. There has been, however, a sampling from a domain of occasions, in that the choice of the second occasion for testing (tomorrow, next week or in 12 months) is arbitrary. There would be the same interest in the outcome if the second occasion were, for example, eight days as if it were one week. Occasions are sampled from a wider possible set of occasions. But this is not what domain-sampling theory is about. Cronbach proposed that the original theory be extended to include not just items but also occasions, in what he termed **generalisability theory** (Cronbach et al., 1972). In obtaining a score on a test, the user, according to Cronbach et al., seeks to generalise beyond

the particular score to some wider universe of behaviour. Generalisability theory asks the user to specify what generalisation they are seeking to make, and then ask whether there are data that support such a generalisation. The detail of the theory is challenging and the ways that it is implemented in practice require a good understanding of the statistical technique of analysis of variance, but the essential idea that extends domain-sampling theory is a valuable addition to our perspective of what reliability of measurement is about.

generalisability theory
a set of ideas and procedures that follow from the proposal that the consistency or precision of the output of a psychological assessment device depends on specifying the desired range of conditions over which this is to hold

Its value can be shown when we extend our thinking about reliability to include cases in which human judgment is the basic assessment tool; for example, a diagnosis of a psychiatric condition or the rating of a person on some characteristic (such as leadership ability). Here the question of reliability arises in terms of whether or not a different judge of similar expertise would make the same diagnosis or rating. In this case, correlating scores across judges provides a means of estimating reliability. Reliable judgments are those that involve high inter-rater agreement. For continuous measures, the intraclass correlation is the appropriate index of inter-rater reliability (see the Technical Appendix). For category data (the patient has the condition or does not), per cent agreement among raters can be used or the kappa coefficient can be computed. Kappa (see, for example, Howell, 2002) is a better index of reliability than per cent agreement when most of those being rated fall into one of the two categories (e.g. most patients are not rated psychotic).

In terms of Cronbach's generalisability theory, the problem of estimating **inter-rater reliability** is one of generalising over judges rather than over occasions or items. The logic remains: what grounds do we have for generalising from this particular sample—the judgment of one individual—to the wider universe in which we are interested; for example, the judgment of psychiatrists in general when presented with this patient or the judgment of leadership experts in general when observing this individual's leadership behaviour. The particular statistical techniques used to assess reliability in any particular instance should not hide the fact that the question being asked is basically the same.

inter-rater reliability
the extent to which different raters agree in their assessments of the same sample of ratees

Put in this wider context of generalisability, a question that sometimes is asked about reliability is shown not to be a good way of thinking about the issue. The question is: Given all these various ways of indexing reliability, which is the correct way? The answer depends on the generalisation, in Cronbach's terms, that you wish to make. Often the interest is in generalising to a domain of items, not all of which it is practically possible to administer. In this case, the methods first discussed (equivalent forms, split-half and internal consistency) are the appropriate ones. But for some purposes the question of generalising over occasions of testing might be quite important and reliability needs to be assessed in terms of some version of the test-retest procedure. Consider, for example, a patient who has suffered a head injury that has produced some cognitive deficits. Those responsible for the care and management of the patient need to know whether these cognitive deficits are getting worse, remaining the same, or perhaps improving as the result of the passage of time or some remedial intervention. In this situation, test-retest reliability of the measure being used is a prime concern. If a test is known to have scores that drift over time, then it is of little use for this type of assessment.

Some special issues

How reliable does a test need to be? Again this is not a good question, because it depends on the circumstances in which the test is being used. If the result of the test has serious consequences for an individual, then a very high level of reliability is required. If, however, the test is in the process of being developed, then one might be content to persevere with a much lower level of reliability, expecting that in time one may be able to improve the low figure obtained. Nunnally (1967) gave the following rule of thumb for assessing reliability: 0.5 or better for test development; 0.7 or better for using a test in research; and better than 0.9 for use in individual assessment. Like all rules of thumb this one needs to be treated cautiously, as Pedhazur and Schmelkin (1991) cogently argue.

How good are the reliabilities of tests in use? The best answer to this question can be found by checking the manual that comes with each commercially produced psychological test or diagnostic procedure, because reliabilities vary considerably. Some conclusions are, however, possible. Tests of cognitive abilities have the highest reliabilities, followed by self-report tests of personality. Jensen (1980) reviewed the reliabilities of widely used individual and group tests of general mental ability and reported that the Stanford-Binet showed a median alternate forms reliability over twenty-one samples of 0.91. The latest version of the Stanford-Binet (Fifth Edition) has reported reliabilities of 0.95 to 0.98 at the scale level and 0.84 to 0.89 at the subtest level. Earlier versions of the Wechsler tests, which cover the age span 4 to 74 years, show reliabilities for Full Scale IQ of

from 0.95 to 0.97. The latest version of the WAIS (IV) has reported reliabilities of 0.98 at the scale level and 0.78 to 0.94 at the subtest level. For thirty individual tests of general mental ability the average reliability reported by Jensen was 0.9. At the other end of the scale are projective measures of personality. Entwisle (1972) summarised findings on the reliability of measures of achievement motivation based on the Thematic Apperception Test (see Chapter 8), which has been used extensively in research, although seldom for decision making in the individual case. Her review indicated that test-retest reliability over periods of one to two months was no better than 0.26, split-half reliability about 0.27, and equivalent forms at best 0.48 and in some cases as low as 0.29. Some advocates of projective techniques (e.g. Atkinson, Bongort & Price, 1977; Winter & Stewart, 1977), it should be noted, would dispute the application of psychometrics to an evaluation of these techniques. The reliabilities of self-report tests are closer to those of cognitive tests, in the order of 0.75 to 0.85 for commercially produced tests (Fiske, 1966).

What are the implications of differing levels of reliability across different tests? One is in terms of assessment of the individual case. Consideration of the standard error of measurement helps to make the point. Suppose we have an individual's IQ result of 105 and wonder whether this means that the person is of at least average intelligence. If the test has a reliability of 0.9 the SEM is 0.31, and if it is 0.7 the SEM is 0.54 (for raw scores expressed as standard scores; to express the SEM in raw score form we simply multiply the standard scores by the standard deviation of the raw score distribution). That is, the interval within which the individual's true score lies is almost twice that at the lower reliability. If the raw score standard deviation is, say, 15 for both tests, this means that for one test the true score is likely to lie within the range 105 ± 5 on 68 occasions in 100 on which we check it, whereas for the other the range is 105 ± 8. We can have more confidence with the first test than with the second that the person's IQ is at least 100, although both test scores involve error. Note that a more accurate assessment would involve calculating the predicted true score and setting the confidence intervals about it, rather than about the obtained score, because the error is about the true score. For most practical purposes there will be little substantive difference in the judgments made, unless the test has a very low level of reliability.

A second reason for being concerned about varying reliabilities among tests is that the reliability of a test affects the magnitude of the intercorrelation of the test with any other variable. The logic of this is straightforward. Thought of in terms of **equivalent forms reliability**, an unreliable test is one that does not correlate with itself. How then can it be expected to correlate with anything else? The effect can be made explicit in terms of the following formula:

equivalent forms reliability
the estimate of reliability of a test obtained by comparing two forms of a test constructed to measure the same construct

$$r_{xy} = \sqrt{r_{xx} \times r_{yy}}$$

where r_{xy} is the intercorrelation between tests x and y, and r_{xx} and r_{yy} are the reliabilities of the two tests.

Although the theoretical maximum correlation coefficient is 1.0, as the reliability falls the maximum possible correlation falls, too. With low reliabilities we may conclude that two variables are unrelated when in fact the magnitude of the correlation has been reduced ('attenuated' is the term used for this in technical writing) by poor measurement of one or other of the variables.

Can reliabilities be improved if found wanting in any particular case? The answer here depends on the nature of the reliability being considered and the practical constraints on what is possible in any particular situation. Where one is sampling from a domain of items, reliability can often be improved by extending the sample; that is, lengthening the test. The Spearman-Brown formula referred to earlier can be used to give an indication of the number of items that need to be added to a test to bring its reliability from a given level to some desired level.

$$k = \frac{r_{yy}(1 - r_{xx})}{r_{xx}(1 - r_{yy})}$$

where k is the factor by which the test has to be lengthened to take the reliability from its current level (r_{xx}) to the desired level (r_{yy}) (Allen & Yen, 1979). The formula makes important assumptions about the nature of the items being added (e.g. that the interrelationships among them duplicate those of the original set of items), which in practice are not always easily achieved. Brief reflection on the formula will show that the relationship between increasing the number of items and changes in reliability is not linear; doubling the number of items, for example, does not double the reliability. On the other hand, improvement of inter-rater reliability calls for better training of raters about the

characteristic being judged and the meaning of the points on the rating scale being used.

Chapter summary

Reliability is an important property of a test or any assessment device because it allows the user to generalise from the score obtained to some wider domain of interest. It is estimated in a number of ways depending on the generalisation one is interested in making, but usually results in an estimate in the form of a correlation coefficient or a standard error of measurement. The former indicates the proportion of variance in the measure that is dependable and the latter allows the user to set a confidence interval on an obtained score to specify the range within which the test taker's true score is likely to lie at the given level of confidence.

Questions

1. Define reliability. Why is it an important concept for psychological testing?
2. Compare and contrast systematic and unsystematic sources of unreliability and give some possible reasons for each.
3. Name the different types of reliability and briefly explain how they can be calculated.
4. Compare and contrast SEM and Cronbach's alpha.
5. Compare the test-retest reliability and SEM of the WAIS–IV and Stanford-Binet (Fifth Edition) at the scale level.
6. A study indicates that the variance due to stable individual differences in a test is 0.36 and that the variance due to other random sources is 0.14. What is the reliability of the test?
7. What is the best estimate of the average intercorrelation of the items of a test?
8. How is test-retest reliability calculated?
9. Why might raters disagree in their ratings of the impulsivity of a group of school children?
10. Can generalisability theory be applied to the items of a psychological test?

Exercises

1. A psychological test has 16 items. The mean and SD for each are as follows:

0.13, 0.33; 0.11, 0.32; 0.11, 0.37; 0.06, 0.24; 0.21, 0.41; 0.08, 0.28; 0.08, 0.27; 0.19, 0.39; 0.11, 0.31; 0.23, 0.42; 0.01, 0.12; 0.10, 0.30; 0.15, 0.36; 0.01, 0.13; 0.11, 0.31; 0.01, 0.09.

The mean score on the test was 2.0 with a standard deviation of 1.91. What is the coefficient alpha for the test?

2. The standard error of measurement (SEM) of a psychological test score, as its name suggests, is an index of measurement error. It tells us something about the reliability/accuracy of scores obtained by that test. The scores of a test are more reliable/accurate if the SEM of that test is small (small error, more accurate).

   This statistic can be calculated if we know the reliability coefficient and standard deviation of a test.

   The following table summarises the reliability coefficients and standard deviations of four psychological tests.

Test	Reliability coefficient	Standard deviation	SEM
A	0.85	15	
B	0.85	5	
C	0.55	15	
D	0.55	5	

   a. Before calculating the SEMs for the above four tests, try to guess which one of these tests has the most reliable/accurate test scores.

   b. Use the following formula to calculate the SEMs and see if your guess is correct:

$$SEM = SD\sqrt{1 - r}$$

   c. Sometimes psychologists want to know if the score on one subtest is significantly higher or lower than the score on another subtest. To answer this, one needs to calculate the *standard error of the difference* between two scores. The equation for this statistics is as follows:

$$SE_{diff} = \sqrt{(SEM_1)^2 + (SEM_2)^2}$$

   Looking at this equation, do you think the *SEdiff* is larger or smaller than the SEM of the two subtests?

3. Entry to a university requires a score of at least 115 on the Australian version of the Scholastic Aptitude Test (ASAT). The ASAT has been found to have a standard deviation of 15 and reliability of 0.90 using an applicant sample.

   a. How would you interpret the reliability of this test?

   b. Would you admit a person with a score of 112 on the test?

4. Two clinicians rate five patients on improvement after psychotherapy on a 100-point scale.

Patient	Clinician A	Clinician B
1	75	60
2	80	70
3	60	45
4	65	50
5	59	40

The Pearson product moment correlation between the two sets of ratings is 0.99. Do you think the ratings are as highly reliable as the correlation coefficient implies?

Further reading

Osterlind, S J (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. Upper Saddle River, NJ: Pearson.
Reynolds, C R & Livingston, R B (2014). *Mastering modern psychological testing: Theory and methods*. Harlow, UK: Pearson.
Rust, J & Golombok, S (2008). *Modern psychometrics: The science of psychological assessment* (3rd ed.). London, UK: Routledge.

Useful website

Measurement (The Personality Project): www.personality-project.org/readings-measurement.html

5 Validity

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. explain the role of validity in psychological testing and assessment
2. specify the key ways in which validity is established
3. describe the approach to validity in terms of the statistics of regression
4. explain the value of thinking of validity in terms of decision theory
5. explain how psychological theory is involved in examining validity
6. describe the ways factor analysis can be used in examining validity.

KEY TERMS

concurrent validity
construct validity
content validity
convergent and discriminant validity
factor analysis
incremental validity
method variance
multitrait–multimethod matrix
predictive validity
standard error of estimate

Setting the scene

- A counsellor who specialises in vocational assessment is interested in knowing how well a test of aptitude for computer programming predicts results in a technical college course in programming in Visual Basic.
- A group working with young boys asks if there is any relationship between the score on a test of 'delinquency proneness' and the likelihood of coming to the attention of the criminal justice system.
- A personnel manager who has introduced a selection test for those working in clerical positions in his organisation is interested in knowing whether decision making about whom to employ in these positions has improved as a result.
- A psychologist is surprised to find that a test that is reported in the literature as highly valid does not seem to be useful in the hospital in which she is working.
- A journalist is planning to write a magazine article about a new measure of 'ecological intelligence', but has cold water poured on the idea by a psychologist friend who questions whether there is any evidence to show that 'ecological intelligence' is any different from intelligence as it has been traditionally measured for over 100 years.
- An experienced manager is firmly of the view that by reading a psychological test carefully you can always tell whether it is any good.

Introduction

In this chapter we explore some practical issues about the use of psychological tests. How do we evaluate how well they predict socially relevant outcomes to do with performance or well-being that society might be interested in, such as success at school or university, the likelihood of suffering from a psychological disorder or in engaging in delinquent or criminal activity? What sorts of errors can be made with psychological tests if they are used to make decisions, and are the errors all equally important? How can we be sure a test is measuring what its authors claim it is measuring or appears to be measuring? These are questions that form part of the literature on psychological testing that is usually considered under the heading of validity. The literature on validity is extensive because the issues can be approached in a number of different ways. The purpose of the present chapter is to acquaint you with the major issues that need to be thought about when examining validity.

The meaning of validity

The **validity** of a test has been traditionally defined as the extent to which the test measures what it purports to measure (e.g. Nunnally, 1967). Test developers make claims about their tests; the most obvious are the labels they place on them. The question of validity asks about the justification for the claims made. Take the example of a test developer who publishes a new test of 'social intelligence'. The test community—those who use or develop tests—and the community more generally have the right to ask the developer of the new test about the extent to which it is in fact a measure of social intelligence and not a measure of, say, 'verbal intelligence' or general education level. The onus is on the developer to justify the claim. Statements of the sort, 'I know a lot about social intelligence, more than most, and I say it is' are not adequate. What is required is an empirical demonstration, and preferably more than one. That is, are there observations about scores on the test that are consistent with the claim that it is a measure of social intelligence? It is in this spirit that the *Standards for Educational and Psychological Testing* define validity as: 'The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test' (American Educational Research Association et al., 1999, p. 184). The 2014 revision of the *Standards* maintains this definition.

validity
the extent to which evidence supports the meaning and use of a psychological test (or other assessment device)

Validity is a central requirement for the use of a psychological test. Without it we have a set of items (questions and tasks) without meaning. The sets of data the developer and, subsequently, other interested test users gather about the test and its relationship to other measures help to give meaning to the test scores. Note that the earlier and the more recent definitions speak of 'the extent to which' and 'the degree to which', which implies that the question 'is the test valid or not' is never one that can be answered 'Yes' or 'No'. Rather the answer is always conditional. For example, we might conclude that on the basis of *what we currently know* this test is a good measure for the purpose for which it is to be employed, but this is not to say that the test cannot be found wanting as new evidence accumulates and it does not mean that it is necessarily valid when used for a new purpose. How we form views about test validity requires some statistics and a lot of critical thinking. The statistics help to crystallise some of the thinking, but, as with most important issues, ultimately an exercise of judgment is involved.

The need to evaluate test validity was recognised early in psychology and has been an ongoing topic of discussion in the test community as ideas are refined and added from the broader domains of psychology and social science. Binet and

Simon (in Ittenbach, Esters & Wainer, 1997), for example, in their pioneering work on the assessment of intelligence, saw the need to justify their test for the purpose for which it was to be used. Although Binet had been called on by the Office of Public Instruction in Paris to develop the test because of his expertise in the field, he did not base the value of his test on his reputation. Instead, he applied two criteria to establish its worth. If the test measured intelligence, then children identified by their teachers as 'bright' children should perform better on it than those identified as 'dull'. Second, older children should perform better than younger children. Only items that met both these criteria were included, irrespective of the merit Binet or his co-workers saw in them.

Binet's criteria were practical; for several decades after his work became widely known, test developers focused on the practical aspect of test validity. The way it came to be framed was in terms of the extent to which scores on the test predicted some criterion measure external to itself (teachers' judgments and children's age were the criteria Binet employed). The use of tests for selection in the First World War and subsequently in industry supported this widespread interpretation of validity. If a test is being used to select which of a pool of applicants will perform best in a particular job, it makes sense to ask about the validity of the test in terms of the prediction of job performance from the test score. This remains an important aspect of discussions of validity, but it was subsequently seen as too limiting in terms of the types of tests for which it is relevant and in terms of its limited integration with developments in the mainstream science of psychology.

The term **construct validity** was introduced to capture a wider core of meaning for validity than predictive validity provided. A construct is a hypothetical concept: a way of talking about features of the world that can make them more comprehensible. Constructs might be found to be unhelpful and are then discarded. In psychology, constructs are 'invented' by theorists in an effort to make sense of aspects of people's behaviour. Intelligence is one construct; anxiety is another. Because we see certain commonalities in the way people solve problems or adapt to their surroundings, we speak of intelligence. It is not a thing in the way a chair or a computer is a thing. It is an idea that potentially makes sense of differences in the way people solve problems. In the same way, anxiety does not exist other than in the responses that individuals make in certain situations; for example, when they are under threat. Regularities in these responses lead theorists to use anxiety as a convenient way of talking about them and linking them to other phenomena. Constructs as ideas are tied to the world of observation by certain 'operations': things that we do to identify them. Answers to a word quiz can be used as one operational way of tying down the construct of intelligence, but constructs have surplus meaning and are not reducible to sets of operations. To show that one particular test of intelligence lacks validity is not to show that this is true of the construct.

construct validity
the meaning of a test score made possible by knowledge of the pattern of relationships it has with other variables and the theoretical interpretation of those relationships

Construct validity sees the test as an operation for giving a construct meaning and asks how well it does that. The value of this approach is that it moves test development into the mainstream of psychology rather than having it as a technology on the periphery. The general approach to theory development in psychology thus becomes available to evaluate the quality of psychological tests, and they in turn can inform psychological theory. To find that scores on a presumptive test of intelligence do not behave as a theory of intelligence predicts might mean that there is a fault with the test (it is lacking validity) or, and this would not be the first alternative accepted, that there is a problem with the theory. It is this interaction between test and theory that attracted psychologists to the thinking about construct validity, and some argued that predictive validity could be seen as a special case of construct validity. One review of the literature on validity (Cizek, 2012), however, concluded that there was value in continuing with the distinction between justifying the use of a test in a practical context and justifying the inferences that are made about the score on a test.

In this chapter, we consider predictive validity separately from construct validity. But before discussing either of these we need to comment on the idea of content validity.

Content validity

The content of the items that constitute a test gives rise to inferences about the nature of the test and there is some evidence that individuals can guess reasonably well what some tests are attempting to assess from reading through the items (Fiske, 1971). In some areas of testing, **content validity** is a sufficient basis for justifying use of a test. An end of semester examination in a psychology course, for example, is validated by demonstrating that the questions asked are drawn from the material set for the course and only from this source, and that the course material is adequately sampled. Beyond achievement testing, content validity is often a poor guide to test validity. Although test developers often use items that 'look' as if they are appropriate so that the layperson can guess their purpose, the use of such questions does not provide the evidence necessary to demonstrate test validity. Tests that 'look' valid, that have what is sometimes called 'face validity', might not be valid when subjected to the more rigorous requirements of predictive and construct validity, and some tests can be useful

even when they include items that have little if any face validity (e.g. some of the items in the Minnesota Multiphasic Personality Inventory; see Chapter 8).

content validity
the meaning that can be attached to a score on a psychological test (or other assessment device) on the basis of inspection of the material that constitutes the test

Predictive validity

As noted above, a claim that a test has **predictive validity** is evaluated in terms of the extent to which scores on the test allow us to estimate scores on a criterion external to the test itself. If the estimates the test provides are good, then we are likely to accept the test as a valid measure of the criterion in question. Thus, a scholastic aptitude test should allow us to estimate to some degree how students will perform in an academic examination. For example, those who obtain high scores on the test should perform well in the examination and those who obtain low scores should perform poorly. As another example, scores on a test of anxiety should predict the ratings of anxiety that psychiatrists make of patients in therapy. Psychiatrists' ratings, as with examination results, are criteria external to the test, which should be estimated from scores on the two types of tests, if they are in fact valid measures.

predictive validity
the extent to which a score on a psychological test (or other assessment device) allows a statement about standing on a variable indexing important social behaviour independent of the test

These examples imply that there is some difference in time between administration of the test one is seeking to validate and assessment on the criterion measures. The scholastic aptitude test is administered, for example, at the beginning of the semester and the examination at the end, or the anxiety test is administered before therapy begins and the ratings are made, say, at the end of the first session. This is often the case but it is not necessarily so. The test and criterion can be administered at the same point in time and the logic still holds. This is often the case when the criterion external to the test is another test. In developing a short form intelligence test, the test developer could administer the short form along with a test that can be considered a well-validated test, such as the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV). The term 'predictive validity' is sometimes restricted to those instances where the test is

administered before the criterion is evaluated and the test is then predicting a future event, a common meaning of prediction. The term **concurrent validity** is then used to characterise those situations in which the test and criterion are administered jointly.

concurrent validity
a form of predictive validity in which the index of social behaviour is obtained close in time to the score on the psychological test (or other assessment device)

It is important to recognise this difference, if only because prediction introduces potentially more error than is the case in concurrent assessment of validity. Events that have nothing to do with the validity of the test might intervene in the interval between test and criterion, and these can reduce artificially the validity of the test. A family crisis can mean that a student does not do as well in an examination as he or she might and hence their actual performance is less than that predicted, but this is not what the test purports to measure (reactivity to a family crisis). Although it is important to use the terminology of predictive and concurrent validity correctly, it is important to note that concurrent validity is a special case of the more general idea of predictive validity; that is, prediction when the time interval is minimal.

The regression approach to predictive validity

The way predictive validity has traditionally been indexed is in terms of the regression coefficient or its close relative, the Pearson product-moment correlation coefficient (see Box 5.1).

Box 5.1

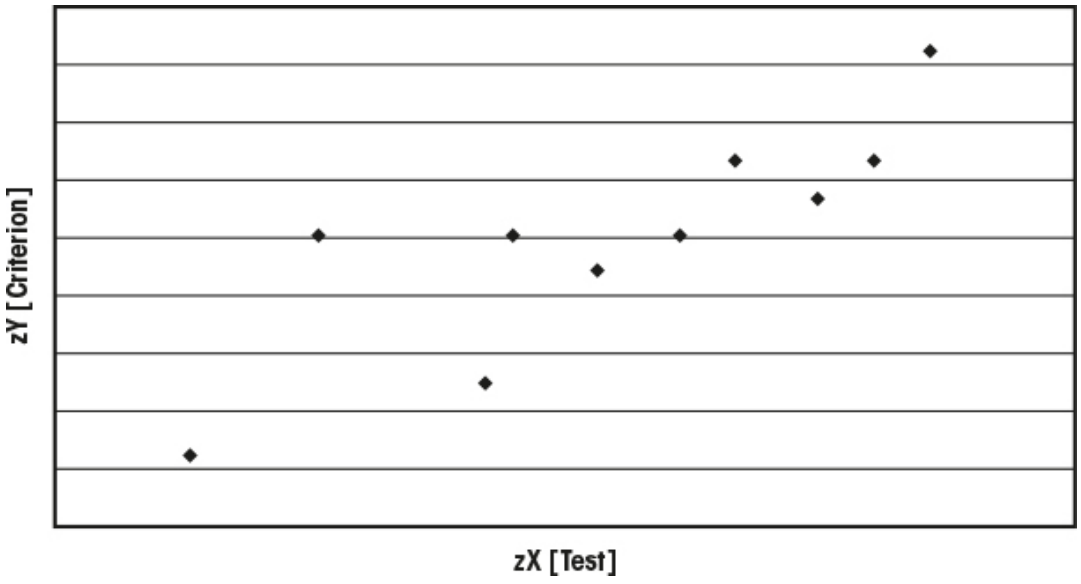
An example that illustrates the calculation of predictive validity

Consider a situation in which we have scores on the test we are seeking to validate and scores on an appropriate criterion for ten persons. We would of course seek a considerably larger sample than this, but for purposes of illustration we will use an N of 10. The test could be one of General Mental Ability that yields scores with a mean of 100 and a standard deviation of 15 and the criterion is a rating of performance on a seven-point scale, from 1 indicating

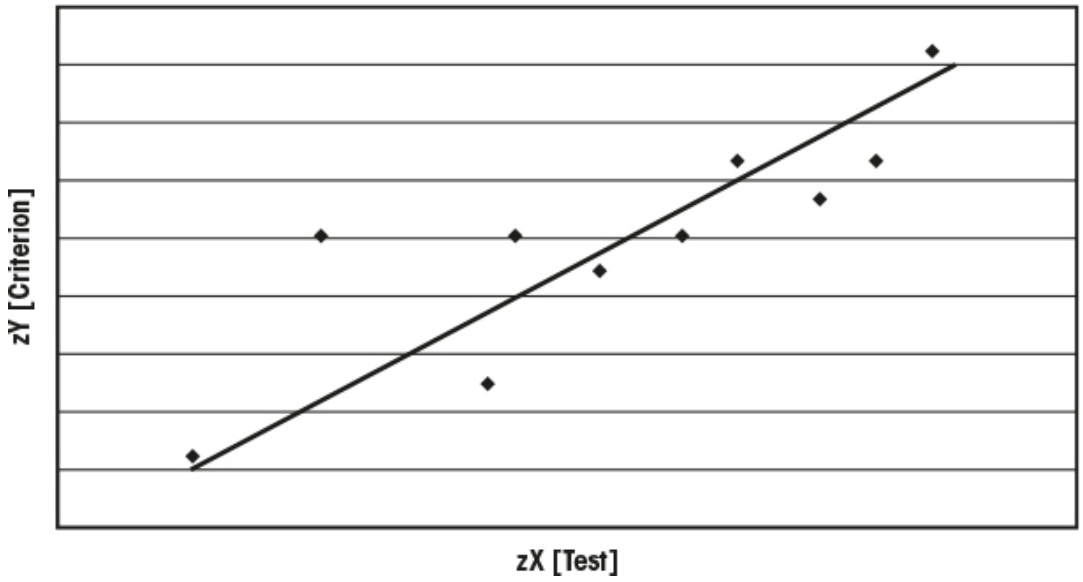
a low score on the criterion to 7 indicating a high score. Suppose the scores are as follows.

Person	Test (X)	Criterion (Y)
1	100	1
2	102	4
3	108	2
4	109	4
5	112	3.5
6	115	4
7	117	5
8	120	4.5
9	122	5
10	124	6.5

First, we plot the criterion score as a function of the test score, with both expressed as z scores.

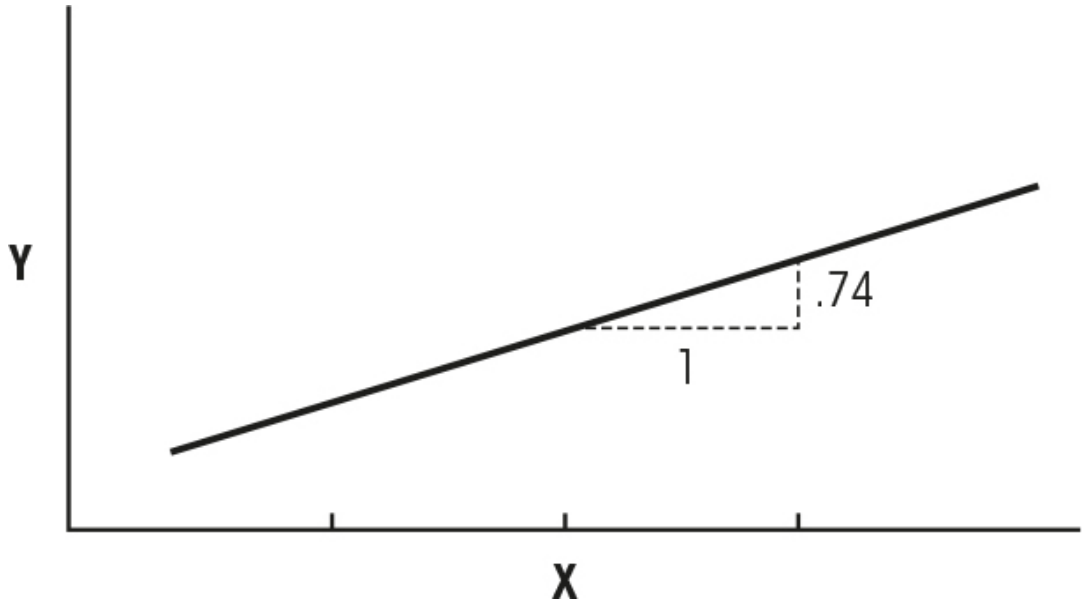


Next, we fit a straight line to the points. This line has been fitted by eye to make the distances of the points above the line about the same on average as the distances below the line.



Instead of fitting the line by eye, we can use a mathematical solution, known as the 'least squares' solution. This fits the line in such a way that the sum of the distances of the points from the line when squared is a minimum.

With a line fixed to the points, we can find its slope: how much do Y values change per unit change in X values (when X changes by 1, by how much does Y change)? We could do this manually by measuring distances on the figure:



That is, Y changes by 0.74 for a change of 1 in X. Alternatively, it can be solved mathematically using the following formula:

$$\begin{aligned} \mathrm{r} &= \frac{\sum(z_x z_y)}{N} \\ &= 0.74 \end{aligned}$$

We can now predict a value of Y for any given value of X, because we have found that zY = 0.74zX.

Whether this is a reasonable thing to do depends on considerations outside the mathematics (e.g. is our sample sufficiently large and representative to warrant generalisation to cases not included in the sample?). If it is, then we can say that the test has a predictive validity of 0.74 for this purpose.

In this example, we have used z scores but we could have used raw scores (see the Technical Appendix). In this case the slope is given by the regression coefficient (b) and there is an additional term in the equation representing the intercept of the prediction line on the Y axis of the plot. When we use z scores this is 0 (the line passes through the origin.)

The correlation between test and criterion is often evaluated in terms of its square (r^2), the coefficient of determination, which is an estimate of the amount of variance in the criterion accounted for by variance in the test. Thus a correlation of 0.3 means that scores on the test account for 9 per cent of the variance in the criterion and a correlation of 0.6 for 36 per cent of the variance. There is an argument for considering the magnitude of the correlation as a direct index of the accuracy of the estimate (Ozer, 1985). Considered in this way, a test with a validity of 0.3 improves the prediction of the criterion by 30 per cent (0.3 × 100) over prediction by chance, and a validity of 0.6 improves prediction by 60 per cent. Prediction, however, is seldom by chance and one usually needs to consider what method of prediction would be used if a test was not used for this purpose. Thus, one could use years of education rather than score on a psychological test to predict job performance, and the accuracy of the estimate would then be better evaluated in terms of the improvement in prediction the test affords over that provided by the demographic information. The improvement is sometimes referred to as the **incremental validity** of the test (Hunsley & Meyer, 2003). Although the absolute magnitude of the validity coefficients for psychological tests is not high (e.g. 0.2 to 0.3 for personality tests and 0.6 to 0.7 for cognitive tests), they often add to the estimate available without them (see Chapter 10).

incremental validity
the extent to which knowledge of a score on a test (or other assessment device) adds to that obtained by another, pre-existing test score or psychological characteristic

A further consideration in evaluating the estimation a test provides is the magnitude of the error involved in any particular instance. We predict from a score on the test (X) that the individual's score on the criterion (Y) will be a particular value. We might expect in a fallible world that the prediction will not, however, be exact; that it will involve some error. The actual scores will be somewhat larger or smaller than we predict. If we inspect the plot in Diagram 2 in Box 5.1, it is clear that the line does not fit the plotted points exactly.

An index of this error is to take the average of the difference between the estimated and the actual values. Where this is large, we can expect in any particular case a good deal of variation of predicted score from actual score. This index is termed the **standard error of estimate** and can be determined from knowledge of the correlation between test and criterion and the standard deviation of the criterion:

standard error of estimate
an index of the amount of error in predicting one variable from another

$$\mathrm{SEe} = \mathrm{SD_Y}\left(1 - \mathrm{r}^2\right)$$

The standard error of estimate (SEe) can be thought of as the standard deviation of the distribution of the differences between actual and predicted scores, with a large SEe indicating considerable difference and hence the greater likelihood of error in any particular case. Although the SEe is a useful index for some purposes, it is more common in test evaluation to use the correlation coefficient, or validity coefficient as it is sometimes referred to in this context.

The regression approach can be used to evaluate predictive validity for one test or it can be used to evaluate a battery of tests; that is, a number of tests used in combination to predict a criterion. In the latter case, a multiple regression analysis is performed, which takes into account the correlation among the tests in the battery. If the tests are uncorrelated (an unlikely situation in practice) the validity of the battery is equal to the sum of the individual validities. When the intercorrelation of tests in a battery is taken into account the validity is less than the sum of the individual validities, and where there is a good deal of overlap

among tests considerably less. The optimal arrangement is where each test in the battery predicts the criterion but accounts for some variance not accounted for by the other tests. There are instances, however, where a test with no relation to the criterion is added to a battery and the validity increases. These cases of 'suppression', however, are rare.

The decision-theoretic approach to predictive validity

As noted above, interest in predictive validity was encouraged by the widespread use of psychological tests in industrial and educational settings after the First World War. This use centred on the value of tests in decision making, itself an issue that came into prominence during and after the Second World War because of its significance in a number of military contexts, from signal detection by the radar operator to choice of a particular plan of attack. One of the earliest members of the test community to recognise the importance of the expanding relevance of decision theory to psychological testing was LJ Cronbach. With Gleser (Cronbach & Gleser, 1965), he wrote an important but difficult text on the application of decision theory to the evaluation of tests. (A more accessible source is Wiggins, 1973.) Some of this thinking informs what follows.

The simplest decision, relatively speaking, that can be made with a test is when it is used to decide which of two categories a person belongs to: successful worker versus unsuccessful worker; a prisoner who is likely to reoffend if released on parole versus a prisoner who is not likely to reoffend; or a patient who is suffering psychotic symptoms versus a patient who is not. There are only two categories possible. More complex decision problems involve more than two categories, but we will stay with the simple case. To make the two-choice decision, a cutting score on the test is determined (by prior research) and those with scores that fall above the cutting score are assigned to one of the categories and those with scores below the cutting score are assigned to the other. The problem can be summarised as follows (see also the diagrams in Box 5.2).

The X-axis represents the range of test scores and the Y-axis the range of outcomes on the criterion variable. Rather than concentrate on the continuous range of scores as in the previous presentation of validity, we think now in terms of grouped scores. A **cutting point** is established on the X-axis, with scores above indicating one type of predicted outcome and those below the other type of predicted outcome (see Diagram 1 in Box 5.2). The actual state of affairs is either consistent with prediction or contrary to it. Framed in this way, it is clear that the use of the test might lead to correct and to incorrect decisions. There are two sorts of correct decision. **Valid positive decisions** are those where the person is predicted to show the characteristic of interest (a successful worker or a

patient with the condition in question) and this is in fact the case. **Valid negative decisions** are those in which the prediction is that the person does not show the characteristic of interest and this is the case. There are, as well, two sorts of errors. **False positive decisions** are those in which the prediction is that the person has the characteristic but in fact does not, and **false negative decisions**, in which the prediction is that the person does not have the characteristic of interest but does.

cutting point
(or cutting score) the test score or point on a scale, in the case of another assessment device, that is used to split those being tested or assessed into two groups predicted to show or not show some behaviour of interest

valid negative decision
a decision that correctly allocates a test taker or person being assessed to the category of those predicted not to show some behaviour of interest on the basis of their score on a test or other assessment device

valid positive decision
a decision that correctly allocates a test taker or person being assessed to the category of those predicted to show some behaviour of interest on the basis of their score on a test or other assessment device

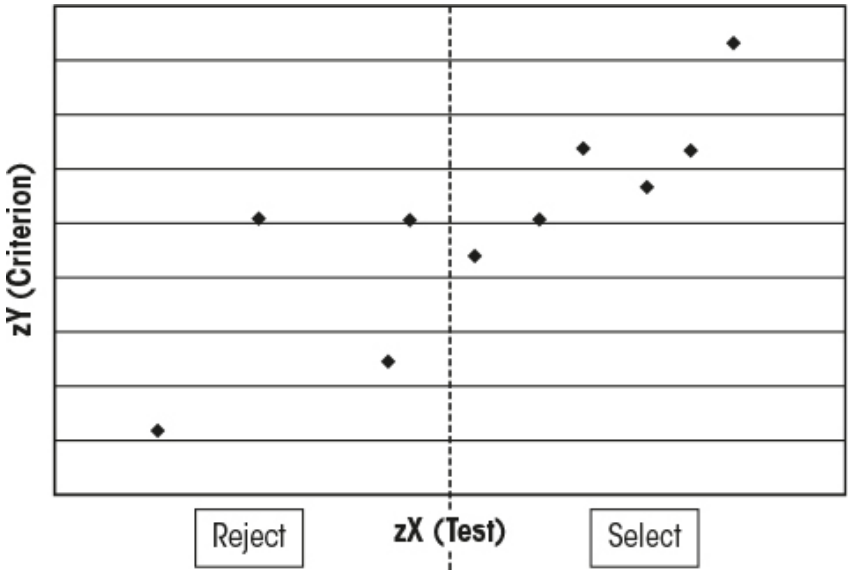
false negative decision
a decision that incorrectly allocates a test taker or person being assessed to the category of those predicted not to show some behaviour of interest on the basis of their score on a test or other assessment device

false positive decision
a decision that incorrectly allocates a test taker or person being assessed to the category of those predicted to show some behaviour of interest on the basis of their score on a test or other assessment device

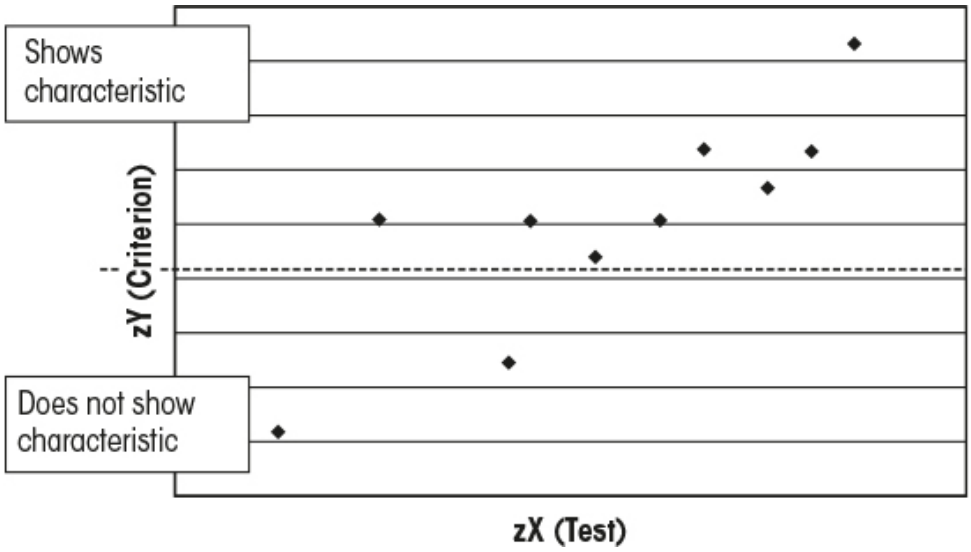
Box 5.2

An example that illustrates the decision-theoretic approach to predictive validity

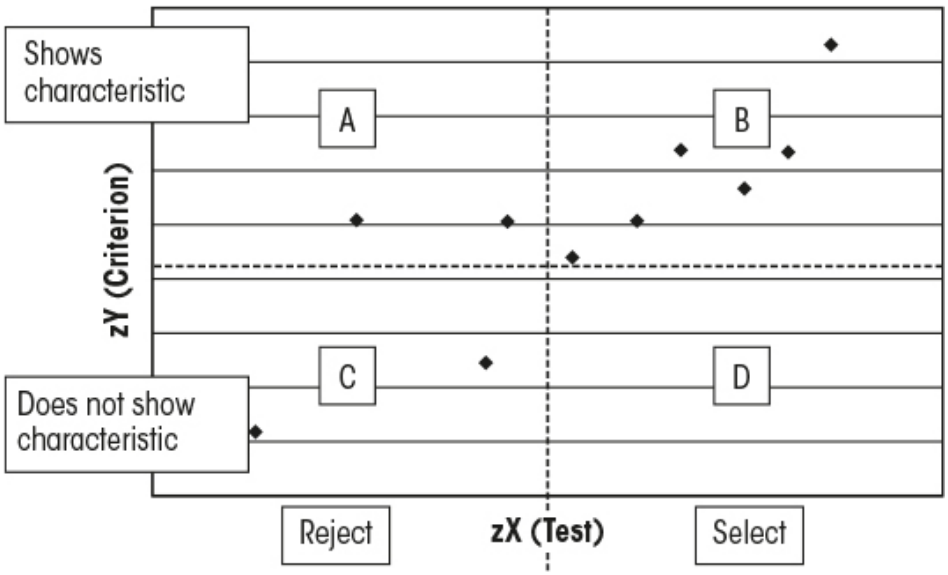
We begin with the example in Box 5.1 of scores on test and criterion for a sample of N = 10, and assign a cutting score on the predictor variable, X.



And then impose a cutting score on the criterion, Y.



When both categorisations are combined, we have four cells: A, B, C, D.



These have the following names:

A False negatives (predicted not to have the characteristic but actually do)

B Valid positives (predicted to have the characteristic and do)

C Valid negatives (predicted not to have the characteristic and do not)

D False positives (predicted to have the characteristic but do not).

The 2 × 2 contingency table formed by splitting the plot into four quadrants based on predicted and actual outcomes has marginal totals. For example, the frequencies in A and B can be added to find a marginal total, or A and C likewise.

Some of these marginal totals reflect factors operating in the context of decision making over which the test user has little if any control. This means that the entries in the table are in fact constrained and are not free to take on all possible values.

A	B
C	D

(A + B)/N = Base Rate (BR) of criterion behaviour in the population of interest

(C + D)/N = 1 – BR

(B + D)/N = Selection Ratio (SR) under the prevailing conditions

(A + C)/N = 1 – SR

(Dividing by N is each case expresses the frequency of cases as a proportion of the total.)

If the marginals are fixed then only one value in the table is free to vary.

A	B	BR
C	D	1 – BR
1 – SR	SR	N

If BR is 0.8, and SR is 0.6, then only one of A, B, C, D can vary.

If B = 0.4, for example, then

D must be 0.2 (0.6 – 0.4)

A must be 0.4 (0.8 – 0.4)

C must be 0 (0.4 – 0.4)

What is the benefit of thinking about the validity of a test in this way? First, it takes us closer to one way tests are used in practice. To say that a test has a predictive validity coefficient of 0.3 does not tell us a great deal about the test in use. To say, on the other hand, that with the test 60 per cent of the predictions on average will be correct is more immediately meaningful. Second, it makes us think about the errors that will be made with the test: the false positives and the false negatives. Although both are errors, they are not always of the same significance. If one is predicting success in pilot training, for example, the false positives are, from the point of view of the organisation doing the selecting, far more important than the false negatives. To say that a person will be a successful pilot and to find that this is not the case will involve the organisation in an expensive training program without a result and could lead to the loss of an expensive aircraft. To say that a person will not be a successful pilot and then find

that he or she does so might have a consequence to the individual but, unless there is a great shortage of applicants for pilot training, no adverse result for the organisation.

In another context, the relative costs or the two types of errors can be reversed. Imagine a situation in which a test is being used to screen for a central nervous system malignancy. If the test predicts the person falls into the category of persons with the malignancy, then there is an extensive neurological examination; if the test predicts the person is clear then there is no follow-up. In this situation, a false positive has only minor consequences: the person must undergo a neurological examination, which admittedly takes time and might involve some inconvenience, but which has a relatively small cost. On the other hand, a false negative is of considerable significance. The person does not undergo the examination that would identify the potentially life-threatening malignancy. The decision theory approach draws our attention to the fact that errors are made in using tests—none is perfect—and makes us think about the consequences of these errors in the way the test is used. For example, any false negative rate in the screening for malignancy might lead us to dispense with the screening approach even though it has a high valid positive rate.

There is a further good reason to consider the decision-theoretic approach to test validity. To demonstrate this we need to first consider the relationship between the classical approach to validity and that based on decision theory. In the classical approach, in which we consider predictor and criterion continuous, we examine the slope of the line relating the two and use the slope as a basic index of validity. In the decision-theoretic approach we set up a 2 × 2 contingency table, in which we cross-tabulate scores above and below a cut-off on the predictor with one of two outcomes on the criterion. A 2 × 2 contingency table permits an index of association to be computed that describes the relationship between the two variables that are cross-tabulated. The index of association frequently used in these cases is the phi coefficient, which is a form of the product moment correlation used in the classical approach. That is, we could compute a form of validity coefficient from the 2 × 2 approach if we wished.

What the decision-theoretic approach adds, however, is the recognition that the magnitude of the association, the validity coefficient, is constrained by the marginal totals in the 2 × 2 table (see Box 5.2). In practice, these marginal totals are often not under the control of the users of the test. One of the marginal totals (the sum of the valid positives and the false negatives) is referred to as the **base rate** (or prevalence) of the characteristic in the population where the test will be used. Without any test being administered, there is a certain number of individuals in the population who have the characteristic of interest (e.g. can do the job, or are recidivists). A second of the marginal totals (the valid positives plus the false positives) is termed the **selection ratio**, the number who can be assigned to the category of persons showing the characteristic, and is defined by

practical considerations unrelated to testing. In the case of personnel selection, for example, the selection ratio is the number of workers the organisation can employ divided by the number who apply. If there are, for example, only ten jobs to fill, then the selection process cannot yield twenty successful outcomes. Similarly, if there are only a fixed number of beds in a psychiatric facility, more patients cannot be admitted to the facility than there are beds to take them.

base rate
the proportion of individuals in the population who show the behaviour of interest in a given psychological testing or assessment situation

selection ratio
the proportion of those tested or assessed who can be allocated to the category of showing the behaviour of interest in a given psychological testing or assessment situation

The base rate and the selection ratio are often fixed by the population on which and the conditions under which the test is to be used. These values, the marginal totals of the 2 × 2 contingency table constrain the values in the cells because the 2 × 2 contingency table has only one degree of freedom: once one of the cell frequencies in the table is set, the remaining cell frequencies cannot vary (see Box 5.2). What this means is that the association between predictor and criterion, the validity coefficient, is set by the conditions under which the test is used and is not some property of the test that holds irrespective of the situation. A test of anxiety proneness that is validated by comparing equal numbers of patients diagnosed with a neurotic disorder and those not so diagnosed has a validity coefficient for a situation where the base rate of anxiety proneness is artificially set at 0.5 (equal numbers in the two groups). If the test is now used in a situation in which the base rate is much lower (or higher) than this—for example, it is employed in an unselected sample from the normal population—the validity coefficient will necessarily be lower.

It is important therefore to know the base rate of the characteristic in the population in which the test is to be used, a consideration that does not necessarily arise with the classical approach to predictive validity. A similar consideration applies to the selection ratio. This at times can be manipulated, and if it can then a higher valid positive rate can be obtained even with a test of low predictive validity. This phenomenon was described many years ago and Taylor and Russell (1939) compiled a table that specified the change in effectiveness possible with tests of varying validities when selection ratios of different magnitudes apply. As well as alerting us to potential problems in the interpretation of predictive validity, the decision-theoretic approach provides

some further indices of test validity. These are discussed in the Technical Appendix.

The regression and the decision-theoretic approaches to validity provide useful insights into the use of psychological tests in practical situations. They do not, however, address the important issue of the social consequences of testing. This was raised in Chapter 1 when we discussed the discriminatory ways tests can be used, in Chapter 2 under the question of cultural differences and their impact on test results, and will be noted again in Chapter 7 on intelligence. The use of tests in educational, organisational, clinical and forensic contexts has consequences for individuals and for society, recognised as early as the 1920s, with Lippman's critique of the use of the Army alpha and beta tests (Rogers, 1995). Social consequences inevitably involve a political dimension, and in a democracy these are matters for everyone and are not the exclusive province of psychologists. Legitimate questions about test use can be raised in good faith, and test developers and test users cannot dismiss or foreclose on them. Understanding of predictive validity can help clarify these debates when they arise, but it is not the last word.

Construct validity

To this point we have considered the practical context in which the validity of tests is studied. In a landmark paper, Cronbach and Meehl (1955) shifted the emphasis in discussions of validity from the practical to the theoretical and argued that the validity of a test depends on the extent to which it truly reflects the construct that it purports to measure. If a test developer claims a test measures intelligence, how well does it do that job? Because constructs are theoretical entities, an answer depends partly on the power of the theory in which the construct resides. A weak theory will have poorly defined constructs that are poorly operationalised, and in these circumstances it will be difficult to conclude with any precision on the validity of a presumed measure of the construct.

The approach is theoretical but moves consideration of validity of a test very much within the mainstream of psychology. Psychological tests become tests of theory in the sense that the ability to develop a valid test of a construct adds some confidence to the theory from which it is drawn. The failure of a test to behave as predicted might bring into question the theory or suggest amendments to it, if the test has been developed along sound lines. Badly constructed tests have no value for testing theories, but well-constructed ones do. With this approach the testing enterprise is no longer a technology, but very much part of theory development.

Cronbach and Meehl proposed ways in which construct validity can be evaluated. They introduced the idea of the **multitrait–multimethod** (MTMM) matrix as a tool for evaluating validity. The basic idea is that the variance in scores on a test arises from three sources: (a) variation due to the underlying disposition the test developer is seeking to assess; (b) variation arising from the method of measurement used in the assessment (e.g. self-report or problem solving); and (c) random error. Valid tests are those in which the first source of variance in test scores is substantial and the other two are small to trivial. That is, the score obtained should depend principally on differences in the underlying disposition and not on the differences arising from the method being used (and not on errors of measurement). If a person is only an extravert when assessed using a self-report test but not when rated by peers, we have cause to doubt the validity of the assessment, because it depends on the method used. To untangle these two sources of variance, Cronbach and Meehl proposed that we examine simultaneously more than one underlying disposition assessed by more than one method. Tests that use the same method will correlate to some degree because of their shared method variance, but tests of the same construct using different methods should correlate to an even greater extent if the underlying dispositional variance is properly reflected by the tests.

multitrait–multimethod matrix
the pattern of correlations resulting from testing all possible relationships among two or more methods of assessing two or more constructs

Cronbach and Meehl introduced one further idea: a test can be called into question as a measure of a construct if scores on it correlate to any considerable extent with a measure of a different construct. That is, validity is demonstrated not just by the correlation of a test with another measure—the classical view of validity in terms of prediction of a criterion—but also by the lack of correlation of a test with a measure of a theoretically different construct. We need discrimination of measures, as well as their convergence, to demonstrate validity. This was a major step forward in understanding validity, because it uses the counter instance as a method of establishing a claim: in this way a researcher seeks to build confidence in a hypothesis by attempting to demonstrate its falsity.

method variance
the variability among scores on a psychological test or other assessment device that arises because of the form as distinct from the content of the test

For example, there was considerable interest in the 1950s and 1960s in the construct of creativity, and various 'creative' measures of creativity were developed. The research program stalled, however, when it proved difficult to show that the various measures of creativity correlated more highly with each other than they did with measures of intelligence (see, for example, Brody, 1972). That is, they generally correlated more strongly with measures of a supposedly different construct (intelligence) than they did with each other.

The name **convergent and discriminant validity** was given to this method of construct validation. For Cronbach and Meehl, it involved calculating a correlation matrix based on scores for a sample of individuals for whom two or more independent constructs are measured using two or more methods (see Box 5.3). When all possible correlations are calculated and the matrix formed, the researcher can evaluate it in terms of three principal guidelines. First, coefficients in the validity diagonal should be positive, substantial and statistically significant. Second, their magnitude should exceed the magnitudes of those in the same row or column; that is, correlations between different constructs measured by the same or different methods. Third, the pattern of correlations with a measure of a construct should be the same across variations in the method of measurement.

convergent and discriminant validity
the subjection of a multitrait–multimethod matrix to a set of criteria that specify which correlations should be large and which small in terms of a psychological theory of the constructs

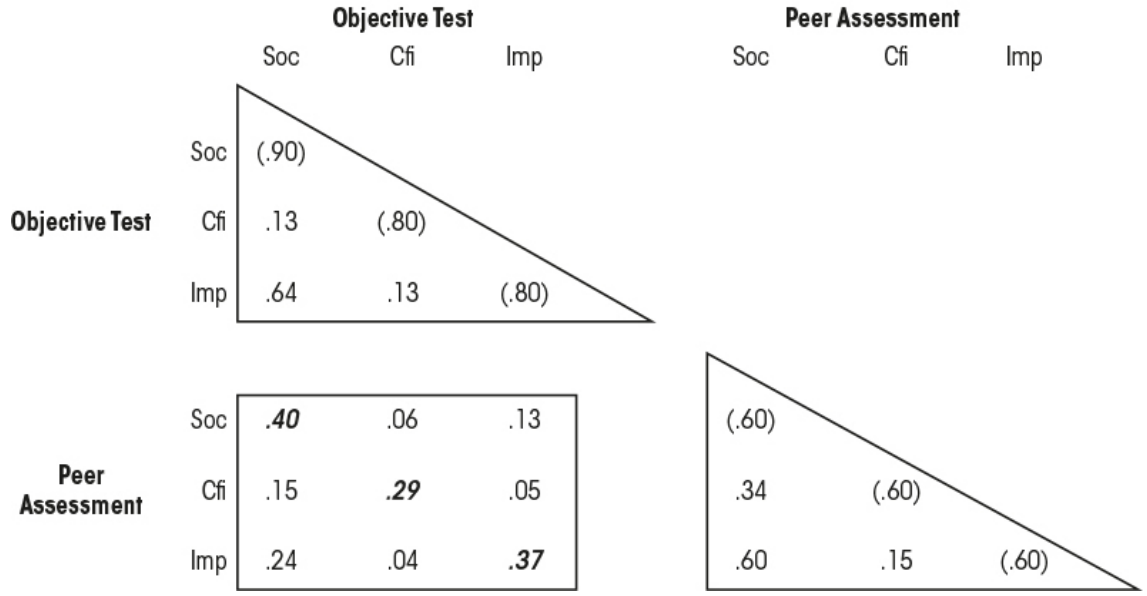
Box 5.3

An example of a multitrait–multimethod matrix

The MTMM matrix requires measures of two or more (hence, 'multi-') constructs (traits) obtained using two or more methods of measurement. The purpose is to examine whether: (a) different measures of the same trait converge (correlate) over methods; and (b) whether the same measures of different traits can be differentiated (discriminated) from each other (fail to correlate). Consider, for example, three traits that according to theory are independent of one another: Sociability (Soc), Cheerfulness (Cfl) and Impulsivity (Imp).

Imagine measuring these traits using two different methods of measurement; for example, objective test and peer assessment. Scores on each trait are obtained using both methods for a reasonable sample of participants,

and the scores are intercorrelated. A hypothetical matrix of intercorrelations is presented in the following:



The coefficients in brackets are reliability (internal consistency) coefficients that are provided to establish the magnitude of the correlations that are possible with the various measures (see Chapter 4). The values in the triangles at the top and right of the matrix are the intercorrelations of different traits measured with the same method (in the case of the triangle at the top, the method is objective test; in the case of the triangle on the right, the method is peer assessment). The values in these two triangles should not be too large because they involve measures of supposedly independent constructs. There will be some correlation, because a method is common in each triangle, but it is a question of their size relative to other correlations in the matrix. The coefficients in bold that lie on the validity diagonal of the square should be large, and certainly larger than the other coefficients in the same row and column.

Cronbach and Meehl's work has been subject to criticism (see Pedhazur & Schmelkin, 1991), not the least because of the difficulties of determining in advance the independence of methods, but the approach provides a valuable addition to thinking about construct validity.

Although the work on MTMM matrices is important, it would be wrong to conclude that this is the only way that construct validity can be examined. In essence, construct validity involves theory testing and this can be done in many ways, which means that there is no fixed set of operations that define construct validity. That said, some procedures are more common than others (Thorndike, 1982). Groups considered to differ in terms of a construct might be compared, or an intervention expected on the grounds of theory to affect a construct could be

introduced to see if the presumptive measure of the construct varies as a consequence. Probably the most widely used method, however, has been factor analysis.

Factor analysis

The correlation coefficient summarises the relation between two variables for a given sample. When there are more than two variables involved, say a number of test scores or a number of items on a test, there are a number of correlations—in fact, $N(N - 1)/2$ correlations where N is the number of variables. The family of statistical techniques termed factor analysis attempts to capture the patterns of relation that can be observed when several variables are related. When there are few variables, visual inspection of the matrix of correlation coefficients might reveal certain patterns of similarity and difference among the variables. Some variables might be highly related and seem to form a set, whereas others might show little relation with these but form their own set. This patterning, or structure, in a correlation matrix can be attributed to the action of latent (unobserved) variables or 'factors' that influence all the variables within a set. The aim of **factor analysis** is to reveal the factors that give rise to this patterning. The Technical Appendix provides some detail on how this is done.

factor analysis
a mathematical method of summarising a matrix of values (such as the inter-correlation of test scores) in terms of a smaller number of values (factors) from which the original matrix can be reproduced

Two methods of factor analysis are distinguished: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Both isolate factor structures, but CFA allows for the testing of hypotheses about the factors and, importantly, the comparison of different hypothesised structures. In the case of EFA, the researcher might have certain expectations about the factor structure, but it is not possible to test for these in a rigorous way. Both techniques have their place and in fact there are hybrid forms that mix the two approaches.

EFA was the earlier of the two methods, originating with the work of Spearman (1927) and Thurstone (1938). Two main types of EFA are recognised. One, principal components analysis (PCA), attempts to provide an economical summary of the relations in a data matrix, whereas principal axis factoring (PAF) is used to reveal the factors that account for the pattern of relations. PCA treats all the variance in the matrix as common variance; PAF analyses only the reliable variance. As the reliability or the number of variables being analysed increases, the results of the two methods converge.

CFA is a more recent addition, stemming from the work of Jöreskog (1969), and is directed to testing conceptual models of relations among variables. A researcher's model will specify the number of factors to be expected in accounting for the inter-relationships of the variables, and which variables and factors are related and which are not. The program for CFA then attempts to generate the model from the data and tests statistically for the fit of the model and data. Importantly, the technique can test for the fit of different models to the data and help the researcher decide which of a number of plausible rival models best account for the patterning in the data set. With CFA it is usual to work with the variance-covariance matrix of the variables being analysed rather than with the correlation matrix, because the variance-covariance matrix provides more accurate information about the estimates made in testing the model.

An example of the application of factor analysis in examining the validity of psychological tests is found in a paper published by Travers, Creed and Morrissey (2015). These researchers were interested in developing a valid and reliable measure of a construct termed Explanatory Style: the way individuals explain unexpected negative events that they encounter (e.g. fail an examination or lose a job). They reasoned from theory and previous research that Explanatory Style consists of three dimensions and they wrote a series of self-report items to capture these dimensions. In an initial study with 320 participants they used PAF to explore the relationship among eighteen of the original thirty items chosen using item analysis (see Chapter 6). Three factors were identified that each accounted for 12–13 per cent of the variance in the matrix of correlations of the item set. In a second study, the eighteen-item set was administered to an independent sample of 396 participants and CFA was used to test the fit of a three-dimension model to the data. Fit was tested using a number of statistics and found to be acceptable, lending weight to their theory of Explanatory Style and their test for measuring it. They went on to make further tests of the convergent and discriminant validity of the new scale using a further sample of participants.

Waschl et al. (2016) used CFA to test hypotheses about the Raven's Progressive Matrices (RPM), a test of general mental ability (see Chapter 1). Because the test involves processing information about geometrical figures, some researchers have argued that the RPM involves visuo-spatial ability as well as general mental ability. Waschl et al. sought to examine the structure of the test using three different samples of participants and three versions of the RPM. CFA with these samples indicated that the best fitting model was a single dimension underlying performance. This outcome rules out the possibility that both general mental ability and visuo-spatial ability are involved in the RPM, but leaves open the question of how this single dimension is best characterised.

A further example is the work of Lovibond and Lovibond (1995a) in developing a test to assess anxiety and depressive states. The Depression Anxiety

Stress Scales (DASS; see Chapter 9) consists of forty-two items with equal numbers of items assessing the three constructs. In developing items for the test, PAF was used to examine the pattern of relations among items. Subsequently, CFA was used with separate samples of participants to test the hypothesis that there were three factors. The test is now widely used for the purposes the authors had in mind and the results of several factor analytic studies testify to its construct validity (see, for example, Crawford & Henry, 2003).

Chapter summary

Tests can be conceived as tools to be used in practical situations such as selection or classification, or as tools of theory. The approach to evaluating validity will vary depending on the focal interest, with predictive validity and utility being of primary concern in practice, and construct validity being of more interest when the theoretical meaning of a test is the concern. But it would be incorrect to see these differences in approach as being absolute. Analysis of a practical problem in terms of theory can suggest appropriate constructs to measure, and identification of such measures can provide the practical solution needed. Alternatively, a theory about a construct might lead to the hypothesis that a measure of it will predict a given criterion; construct validity can be shown by the predictive validity of a test. In all cases, we are interested in reasoning from the test scores to some non-test context, practical or theoretical, and the essential question is: What warrant do we have for going beyond the test?

Questions

1. Compare and contrast reliability and validity.
2. Explain why validity is an important property for a psychological test.
3. Give some examples of criteria commonly used in predictive validity.
4. What is the difference between concurrent and predictive validity?
5. Define content and construct validity.
6. What is a multitrait–multimethod matrix? Discuss its significance.
7. What is factor analysis trying to achieve?
8. How would you validate a test of leadership ability?
9. What is a major obstacle to the development of a screening test for a condition such as schizophrenia?
10. Suggest some strategies for validating a test of emotional intelligence.

Exercises

1. A new test has been developed to predict whether members of a prison population will be diagnosed as psychopathic. Results for a sample of prisoners are as follows:

Test score	Diagnosis
40	Psychopath
20	Non-psychopath
21	Non-psychopath
25	Non-psychopath
35	Psychopath
26	Non-psychopath
30	Non-psychopath
32	Psychopath
25	Non-psychopath
26	Non-psychopath

Looking at this table, what do you think is a good cut-off score for this new test?

a. If this cut-off score is used, what is the validity coefficient for this sample?

b. If this cut-off score is used, what is the valid positive rate?

c. Would a psychologist on the basis of these results be confident in using the test with a sample of adolescents drawn from the community?

2. The selection ratio to be used in a testing situation is set at 0.3 and the base rate for the behaviour in question is known to be 0.3. If the valid positive rate is 20 per cent, what is the valid negative rate? What are the error rates in using the test in this situation?

3. A new test of emotional intelligence has been developed for executive selection. It has an internal consistency reliability of 0.75. As part of the validation process, the test is administered to a sample of 500 managers with a standard test of intelligence (reliability 0.92). Ratings of the general ability level of all managers in the sample and of their emotional intelligence are obtained from their supervisors. Ratings of this sort have a

reliability of no more than 0.45. Draw up a multitrait–multimethod matrix (by using the information provided and coming up with other correlation coefficients) that would point to the validity of the new test.

4. What factors might you expect to find in a factor analysis of the following correlation matrix:

Test	1	2	3	4
1 Vocabulary		0.65	0.07	0.15
2 Reasoning			0.05	0.15
3 Dexterity				0.45
4 Mechanical reasoning				

5. A test purports to measure two personality constructs. How many factors would be predicted to be found in a confirmatory factor analysis of the items of the test? Would it be necessary for the factors to be uncorrelated?

Further reading

Gorsuch, R L (2003). Factor analysis. In J A Schinka & W F Velicer (Eds.), *Handbook of psychology: Vol. 2, Research methods* (pp. 143–64). Hoboken, NJ: John Wiley & Sons.

Kline, P (1994). *An easy guide to factor analysis*. London, UK: Routledge.

Osterlind, S J (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. Upper Saddle River, NJ: Pearson.

Rust, J & Golombok, S (2008). *Modern psychometrics: The science of psychological assessment* (3rd ed.). London, UK: Routledge.

Wasserman, J D & Bracken, B A (2003). Psychometric characteristics of assessment procedures. In J R Graham & J A Naglieri (Eds.), *Handbook of Psychology: Vol 10, Assessment Psychology* (pp. 43–66). Hoboken, NJ: John Wiley & Sons.

Useful website

International Test Commission: www.intestcom.org

6 Test Construction

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. describe the typical steps taken in constructing a psychological test
2. explain similarities and differences between psychological measurement and other types of measurement
3. outline advantages of item response theory that make it an attractive model for the development of psychological tests
4. describe the factors that need to be taken into account in developing the items for a psychological test
5. describe how item analysis is used in the construction of a psychological test
6. explain why a test manual is part of test construction.

KEY TERMS

classical test theory
construct
empirical approach
item response theory
latent trait
model of measurement
plan for item writing
rational-empirical approach
test specification
trace line

Setting the scene

- When medical schools in Australia established postgraduate programs in medicine open to entry by graduates with a range of different degrees, they needed a selection process that could cope with large numbers of applications, but not rely on the information source used for undergraduate medical school selection; that is, high school performance. A new test of aptitude for medical training was required.
- A concern by government that the achievement of Australian school students in mathematics and science was falling behind that of students in comparable countries led to the call for repeated testing for numeracy throughout the years of primary and secondary schooling, and with it the need for age-appropriate tests of ability in mathematical understanding.
- Professional staff in a counselling service dealing with large numbers of clients experiencing grief and loss as a result of crime formed the view that the capacity to accept and forgive was essential to client progress and asked for assistance in developing a test that would track this through the therapeutic encounter.
- A psychology student proposed for her Honours project to test the hypothesis that schoolchildren's altruism is linked to the adequacy of their self-concept, and needed measures of both these characteristics.
- A large business firm wanted to evaluate the morale of its staff and called in a consultant to do this for them systematically and objectively.

Introduction

New ideas in education, health, business and government bring with them the need for more information about human behaviour and experience on which to base decisions, and for new or modified psychological tests. This chapter is about the work that is done in constructing a psychological test. What are the procedures employed? What sorts of decisions need to be made? Is it all based on human judgment, or is empirical evidence brought to bear on the task and, if so, how? In describing the procedures that are typically followed, the intention is not to prepare you for actual test development, but to give you a greater understanding of the processes involved in developing a psychological test. This should make you a more critical user of psychological tests, either as a professional administering or interpreting tests, or as a consumer to whom a psychological test is administered.

The rational-empirical approach

A psychological test is a set of items that allows measurement of some attribute of an individual. The items might be problems to which the individual must find

correct answers—as in the case where the attribute is an ability of the person—or they might be questions about the way the individual typically behaves, feels or thinks, as in the case where the attribute is a personality characteristic. Other types of items will be appropriate for other types of attributes; for example, an expression of a sentiment where an attitude the person holds is the attribute, or a statement of preference where the attribute is an interest. The term '**item**' has been traditionally used as a generic way of referring to the various forms the content of a psychological test can take.

item
the various forms the contents of a psychological test can take

The set of items is in no sense random or accidental, because it must permit some form of measurement of the attribute. Often, one sees in popular magazines collections of items that purport to indicate some attribute of the magazine reader, such as 'Are you a good partner?'. The reader is invited to complete the items and then some 'diagnosis' is offered; for example, 90 per cent correct means you are a 'good partner', 60 per cent means you 'could improve' and 30 per cent means 'there is a lot wrong with the way you are approaching your relationship'. Seldom, however, has there been any rigorous development of the 'test' that permits any reasonable inference being drawn. There is nothing particularly wrong with these 'tests' as long as they are taken as entertainment, which is what the editor of the magazine intends. They have the look of a psychological test in that they are a collection of items, but without the developmental work necessary there can be no claim that any form of measurement of the attribute in question—in this case 'being a good partner'—is possible.

The approach outlined in the sections that follow has been termed the **rational-empirical approach**, as distinct from the **empirical approach**, to test construction (Kline, 1993). In the interval between the two world wars, a number of tests were constructed using the empirical approach. Items were selected on the basis of how well they correlated with a criterion of interest. The constructors of the MMPI, for example, used this approach (see Chapter 8). An item was selected if it was shown to discriminate between a criterion psychiatric group and a normal group (hospital visitors), irrespective of its content. Patients diagnosed as suffering from schizophrenia were found to be more likely than members of the community to endorse an item such as 'I like horseback riding'. As a consequence, this item was included in the schizophrenia scale of the MMPI. Such 'blind' acceptance of item discrimination indices was questioned by a number of commentators who argued for a more rational basis for the development of psychological tests, in which theory about the construct or

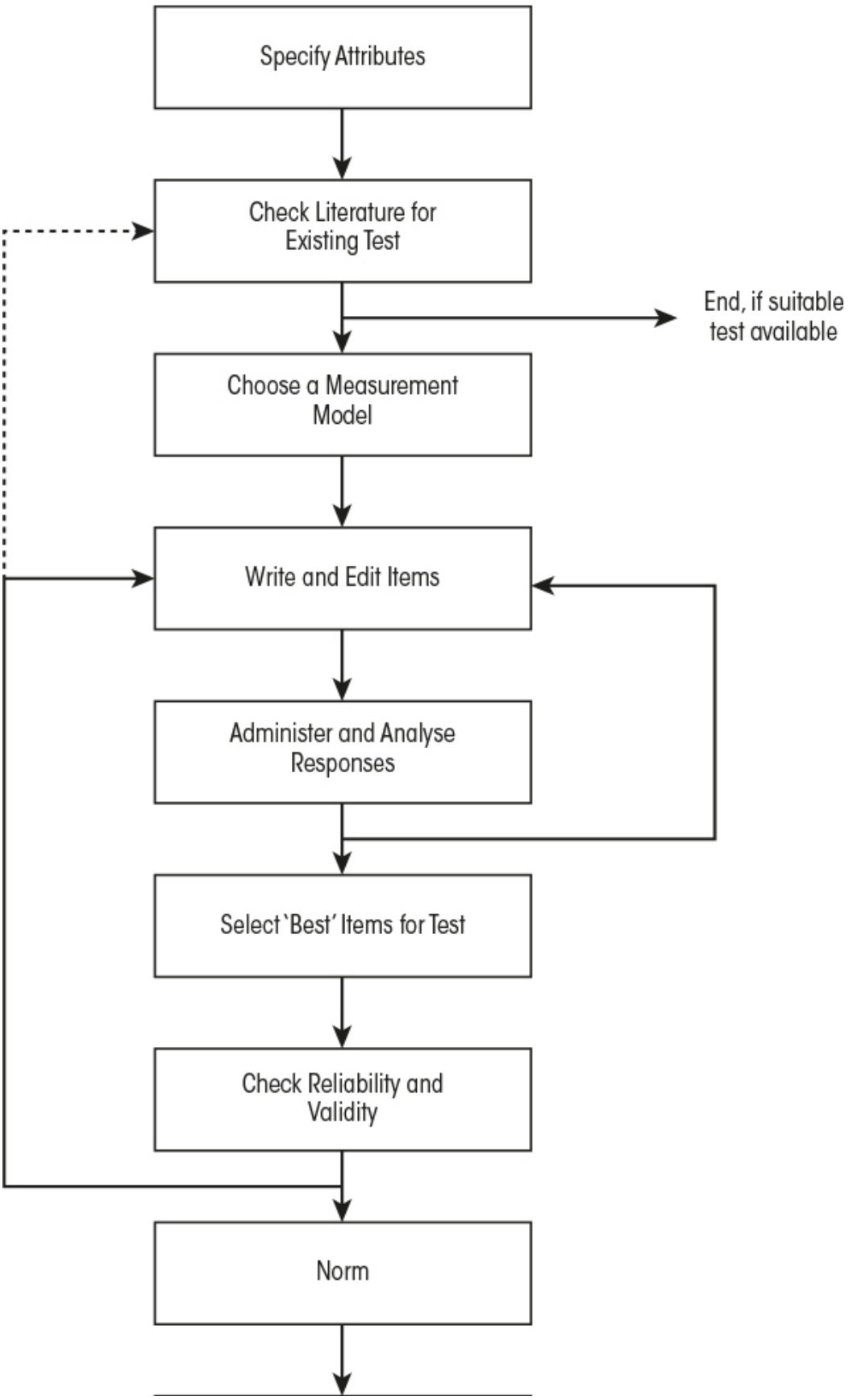
constructs of interest guides the process (e.g. Jackson, 1971). Although the superiority of one approach over the other is itself an empirical question (which one provides the better test?), many test developers today (e.g. Clark & Watson, 1995) opt for the rational-empirical approach, in which both theory and data are used to guide the process of test development.

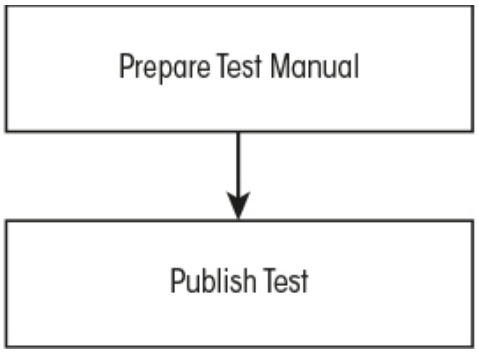
rational-empirical approach
a way of constructing psychological tests that relies on both reasoning from what is known about the psychological construct to be measured in the test, and collecting and evaluating data about how the test and the items that comprise it actually behave when administered to a sample of respondents

empirical approach
a way of constructing psychological tests that relies on collecting and evaluating data about how each of the items from a pool of items discriminates between groups of respondents who are thought to show or not show the attribute the test is to measure; also an approach to personality that relates the reports that people make about their characteristic behaviours to their social functioning and thereby provide tools for personality prediction

For purposes of exposition, the development process is set out in a number of steps. In practice, the process might not always be as linear as this account implies; some steps can be collapsed on to one another and some steps repeated, or the process can be looped at certain points. In general, however, test developers work through a process similar to that described here and briefly summarised in Figure 6.1.

Figure 6.1 Steps in test construction


Specify Attributes
Check Literature for Existing Test
End, if suitable test available
Choose a Measurement Model
Write and Edit Items
Administer and Analyse Responses
Select 'Best' Items for Test
Check Reliability and Validity
Norm



The description of the process of test development is purposely made as general as possible, rather than have it focus on the development of one particular type of test, but examples of particular tests are provided from time to time. The process has itself developed from the earliest work of Binet et al. and has been refined by practice and by developments in psychometrics: the mathematics of psychological measurement. The account given here rests heavily on the work of Nunnally (1967).

Specification of the attribute

The first question to be asked concerns the **attribute** that the test developer is seeking to measure. Test construction requires a clear specification of the attribute or characteristic of the person to be measured and what is known about it. The term **construct** was used in the earlier discussion of validity (see Chapter 5) and that term is useful for locating what is to be measured within a theoretical matrix of other constructs, which specifies what is known about it. The term 'trait' or '**latent trait**' is often used in this context as well, and in theoretical work is often denoted by the lower-case Greek letter theta (θ). Latent trait involves the strong assumption that there is only one dimension underlying the attribute. We will use these various terms interchangeably, although attribute or characteristic is closer to ordinary usage than construct or latent trait. Both definition and understanding of the attribute can change as a consequence of test development, but the test developer needs to begin with as clear a specification as possible. Knowledge of the attribute is needed to generate the items that will form the test —both their format and content—and for testing the validity of the test once it has been developed. Without sound knowledge of the way the attribute is supposed to behave, it is not possible to check the test's validity. Although test development sometimes begins with only a rudimentary theory of the attribute to be measured, attributes embedded in rich theories (i.e. theories with lots of testable implications) make for better starting points for test construction.

attribute
(or characteristic) the consistent set of behaviours, thoughts or feelings that is the target of a psychological test

construct
a specific idea or concept about a psychological process or underlying trait that is hypothesised on the basis of a psychological theory

latent trait
the hypothesised continuously and normally distributed dimension of individual differences that is the sole source of a consistent set of observable behaviours, thoughts and feelings, which is the target of a psychological test

Because more than one person often will be involved in the various stages of test construction, the **test specification** needs to be a written one. This has to include a clear definition of the attribute and the outcome of a literature search that identifies the central theoretical claims about it, and any research findings bearing on it. If the test is to measure more than one attribute, the specification needs to be done separately for each, and a section provided on why and in what ways the attributes are separable (see the discussion on discriminant validity in Chapter 5).

test specification
a written statement of the attribute or construct that the test constructor is seeking to measure and the conditions under which it will be used

Literature search

Once it is clear what it is that a test is supposed to measure, the would-be test constructor needs to establish whether or not a satisfactory test of the attribute exists. There are now a large number (in the thousands) of tests in the psychological literature, as reference to the *Mental Measurement Yearbooks* indicates. It would not be sensible to add to this list, and expend a good deal of effort and money (as these exercises are costly), without being assured that the test is needed. A literature search, beginning with the latest *Mental Measurements Yearbook*, is required to establish what tests of the attribute in question have been published and what their properties are. It may be that no test has been developed, but this is unlikely and can result from a failure to

search the literature carefully enough or because the attribute has not been clearly specified. The would-be test constructor may be calling the attribute X (e.g. intelligence), whereas it has been referred to in the literature as Y (e.g. general mental ability). Ambiguities of this sort are less likely if the attribute has been clearly specified from the outset.

It is more likely that a test has been developed but that its properties are less than adequate for the purpose for which it is required (e.g. there are no norms for the population with which it is to be used). Further work with an existing test (copyright permitting) might be a better investment than development of a new test. The point being made is that the test developer needs to justify the test development project. In so doing, the expected use of the test (e.g. for research or for decision making in the individual case) will be made clear.

Choice of a measurement model

Having decided on the attribute and the theory about it, and having determined that no suitable test is currently available, the next choice is the **type of measurement** required and the model to be used to attain it.

type of measurement
the scales of measurement proposed by Stevens; that is, nominal, ordinal, interval and ratio

Types of measurement

Psychological tests are developed to measure psychological attributes such as cognitive abilities, personality characteristics and social attitudes. This seems straightforward, but what is meant by 'measure' in this context? We measure our weight by stepping on the bathroom scales, our height with a tape measure, and the time it takes us to walk a kilometre with a stopwatch. The process in all these instances of measurement is routine, and we have been familiar with them since a fairly young age. Is the measurement of psychological attributes the same? Is, say, an IQ score we obtain with an intelligence test the result of the same sort of process as the time we take with a stop watch, or the height we obtain with a tape measure?

A moment's reflection suggests that it is not, although it might be difficult to put our finger on why. In the case of height or time, we have a scale (metres or seconds) and we express the person's height or the elapsed time as a ratio of units of the scale. If we use a tape to measure our height as 2 metres, we have a real number, which is the ratio of an attribute of ourselves to the scale of length (with

the unit being metres). In the case of IQ, we do not have such a scale, and there have been authorities since at least Kant who have argued that we never will (Michell, 1999). The Ferguson committee of the British Association for the Advancement of Science (Michell, 2004) was probably the most recent of the authorities to so decide. The committee, in its report in 1940, reasoned that, whereas we can 'concatenate' lengths, weights and periods of time, we cannot do so with any known psychological attribute, and therefore psychological measurement is not measurement as we normally understand it in the physical sciences. 'Concatenate' means 'link together' or '*add* together', and the basis of measurement in the physical sciences involves the discovery of appropriate concatenation operations such as adding two lengths together to measure a third longer length, adding time intervals to measure time, or adding standard weights to measure weight. Concatenation of physical quantities is mirrored in the addition of numbers and facilitates the use of mathematics to express physical theories.

The Ferguson committee's somewhat disheartening conclusion prompted the psychologist S S Stevens (1946) to propose an operational basis for measurement: measurement is the assignment of numbers to objects and events according to rules. The measurement of height entails the use of a tape measure, which allows us to assign numbers consistently to people who differ in height. The property height (length) is thus a function of how it is measured. Stevens' use of the more generic term 'rules' included the idea of concatenation operations as developed by physical scientists, as well as the methods of scaling developed by psychologists as outlined in this book. Stevens' view came to be known as operationism and implied that we cannot really understand something until we can measure it. Indeed, according to the operationists, scientific constructs are *defined* by the process of measuring them: their so-called operational definition. Thus, measurement became a major goal of any scientific enterprise. Operationism was to have a major impact on psychology in the middle years of the twentieth century. For example, Boring (1923, cited in Rogers, 1995) famously defined intelligence as what intelligence tests measure. Critics argued that such definitions were basically vacuous and the age of operationism in psychology gave way to the modern emphasis on construct validity.

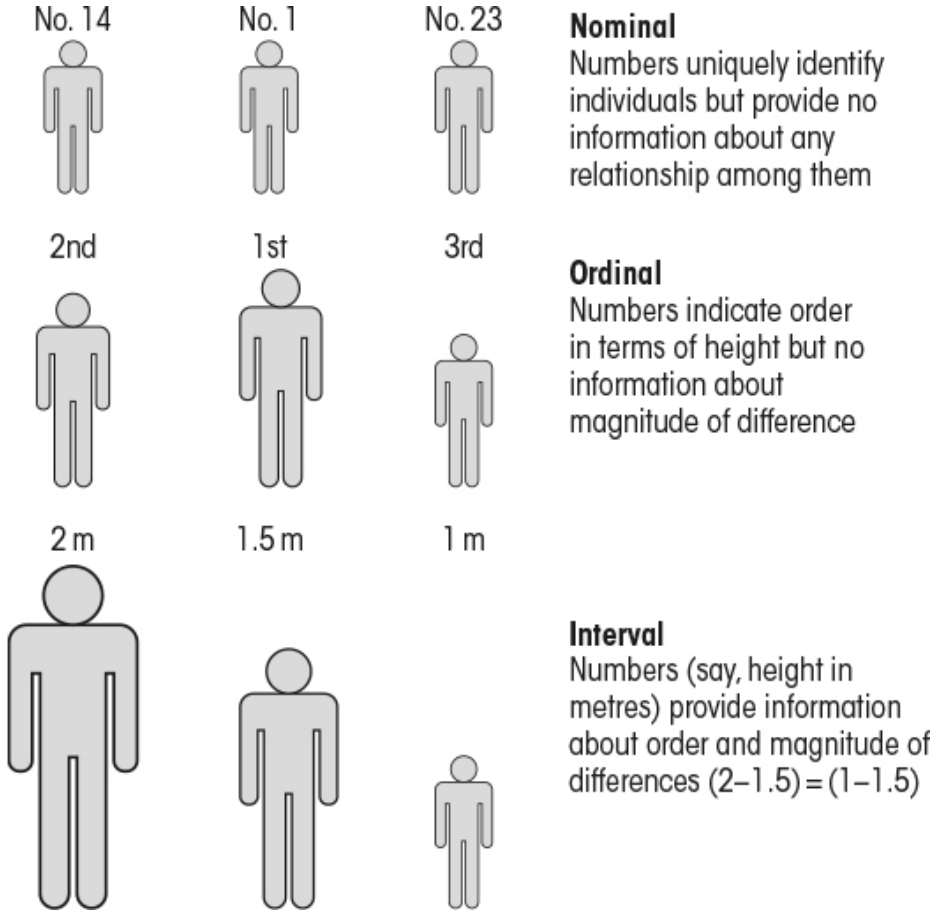
To circumvent the problem that the use of numbers implied addition, which further implied an underlying concatenation operation, Stevens argued that we could classify all **measurement**—be it in the physical sciences or in the behavioural and social sciences—in terms of four categories or scales: nominal, ordinal, interval and ratio. Addition was meaningful only for interval and ratio scales, but the properties of the number system could be used for any of the scales, as long as one did not overstep what was permissible.

measurement
the assignment of numbers to objects according to a set of rules for the purpose of quantifying an attribute

Nominal measurement (see Figure 6.2) is hardly measurement at all, and simply involves naming objects to indicate their discreteness. Players in a sporting team wear numbers on their shirts to identify them for the referee. They could be identified by having their names on their shirts, but if two players happened to have the same name there would be some confusion. Numbers are unique. The rule applied here is simply that each player receives one and only one number, and that no two players can receive the same number.

nominal measurement
the lowest form of measurement that assigns numbers to objects to represent their discreteness from each other

Figure 6.2 Types of measurement



Ordinal measurement improves on this by assigning numbers in a way that permits some inference about relationships among objects. Objects are ranked in terms of the quantity, and numbers are assigned from more of the quantity to less (or from less to more; it does not matter as long as a consistent approach is adopted). Larger numbers mean more of the quantity in question than smaller numbers (or vice versa if ranking has been done in the reverse order), but how much more is unknown. There might be very little difference, for example, between performers ranked first and second in a competition, but a considerable difference between these two and the person ranked third. That is, the distance between 1 and 2 in this case is not equivalent to that between 2 and 3. Ordinal measurement does not carry any information about the distance between objects in terms of the quantity of interest.

Interval measurement, on the other hand, does. We now know the relative positions of two objects with respect to the quantity in question, but we also know how far apart they are in intervals of a certain magnitude. We know that a boy with a height of 70 centimetres is taller than a boy of 65 centimetres, but we also know that the difference in height of these two boys is the same as that for boys of 84 and 89 centimetres. The scale, in this case length, is informative because the intervals are of equal magnitude.

If a scale has the property of equal intervals but also a true zero—that is, there is a point at which the quantity is said not to exist—then this is a **ratio scale**, according to Stevens. Length and mass as we commonly measure them are ratio scales. Temperature as measured in degrees Celsius or Fahrenheit is not a ratio scale because at 0°C or 0°F there is still heat in an object (the molecules of which it is constituted are still moving). The zero on the Celsius scale is the freezing point of water, a useful but arbitrary reference point.

ratio scale
a scale that has the properties of an interval scale but also has a true zero

Stevens' classification had several implications. One was that it fermented a controversy, which is still alive in the psychological literature, about the appropriate statistical methods to be applied to variables at different levels of measurement. If no more than ordinal level measurement is attained, then the arithmetical processes required to compute a mean (or a median) and a standard deviation have no meaning, and many of the statistical procedures used by psychological researchers cannot be employed. Although this may bring a sense of relief to students, the relief might be short-lived because many have been unwilling to accept this stricture, and point to the role of convention in measurement (Cliff, 1982). As long as the conventions are understood, and claims are not made beyond the limits of the conventions, then reasonable

inference is possible. Thus, we can reasonably compute a student's grade point average by summing their grades over different subjects and dividing by the number of subjects, even though we might not wish to defend the proposition that the differences between HDs, Ds and Cs (or As, Bs and Cs) are all equal.

A second implication—as Michell (2009), a trenchant critic of most current approaches to psychological measurement, has pointed out—was that the use of the classification allowed psychologists to 'smuggle in' the idea that most psychological attributes are quantitative. For Michell, 'quantitative' means possessing an additive structure. In asserting that A can be ranked above or below B with respect to some characteristic, psychologists assumed that the characteristic must therefore be quantitative in nature, and capable of measurement. Nunnally (1967) made this explicit in adapting Stevens' definition: 'Measurement consists of rules for assigning numbers to objects in such a way as to represent quantities of attributes' (p. 2). Again, note the use of the term 'rule' rather than 'concatenation'. Interestingly, it turns out that additive structure, and hence valid mapping to the number system, can be established without the discovery of obvious concatenation operations. Luce and Tukey's (1964) development of additive conjoint measurement is an example.

There is much more that could be said on these matters, but for present purposes we will adopt a pragmatic approach and accept Nunnally's definition of measurement.

Models of measurement

Given that the level of measurement likely to be attained in developing a psychological test is, at best, an **interval scale**, there is a further consideration: what **model of measurement** is to be used in test construction?

interval scale
a scale that orders objects in terms of the attribute in such a way that the distances on the scale represent distances between objects

model of measurement
the formal statement of observations of objects mapped to numbers that represent relationships among the objects

A model is a way of representing a phenomenon (McGrath, 2011). It might be close to the phenomenon it seeks to represent—which is the case for a recipe and the meal that it produces—or it might be more abstract, which is the case with the atomic model of matter. Although models do not have to be mathematical,

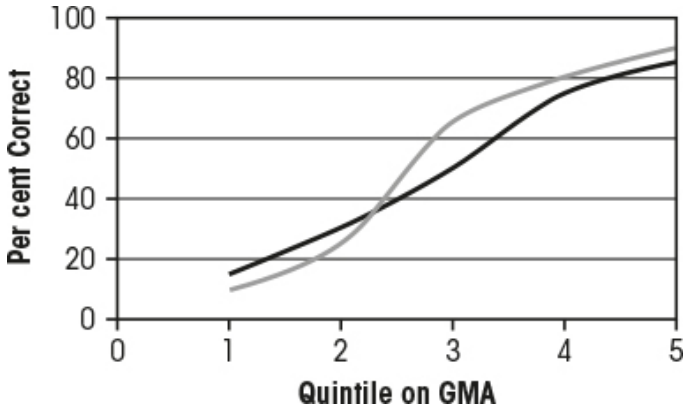
the precision of mathematical models makes them attractive in science. Mathematical models in psychological testing seek to represent the relationship between the attribute to be measured and the response of individuals to the items that make up the test. This is done for a single item, and can be represented by what is called a **trace line** or an **item characteristic curve**. This relates the likelihood of endorsement of the item (in the simplest case, whether the respondent says 'yes' or 'no' to a question, or gets a problem right or wrong) to the person's position on the underlying attribute of interest. Because much of the early work in this field involved general mental ability, it is appropriate to use an example from that literature. Given that we are attempting to measure general mental ability, we can assume that any individual in question will have a position somewhere along the underlying distribution of ability. They might be of high ability, somewhat less than that, below average or anywhere along the assumed underlying dimension. We are attempting to identify their true position, their true score, using the fallible set of items. Each item we use might not be a good marker of this true position and we need to know how individual items behave. If the item is a good one, then (at least most of the time) those high in ability will pass the item and those low in ability will fail.

trace line
a graph of the probability of response to an item as a function of the strength of or position on a latent trait

item characteristic curve
the term for a trace line in item response theory

Suppose we have 100 participants to whom we have administered a ten-item test of general mental ability. If we divide the sample into quintiles on the basis of their total score on the test (i.e. split the sample at the 20th, 40th, 60th and 80th percentile equivalent of total score), so that we have five groups, and then calculate the percentage of each group that answered each item correctly, we can plot a trace line for each item. Figure 6.3 plots imaginary data for two items. Both curves show that the percentage of each group getting the item right increases from the lowest quintile to the highest, but the curves are not the same. One rises more steeply than the other and one is further to the right. These trace lines are based on (fictitious) data, and the question is: how can we best model these curves? Is there a mathematical function that, on theoretical grounds, presents a model of item responding? If there is a theoretical trace line, then we can use it to judge how well any particular item in our test is behaving.

Figure 6.3 Empirical trace lines for two items from a (fictitious) test of general mental ability



For the first half of the twentieth century, test developers were content to use a monotonic increasing function for the trace line in the model without further constraints. In the second half of the century, test developers argued that better measurement was possible if stronger assumptions were made about the trace line. For example, as well as being monotonic (not changing direction), the precise nature of the mathematical function relating probability of responding to position on the underlying attribute could be specified. Some adopted the cumulative normal distribution, but most opted for a simpler function to work with (the logistic function). By specifying a particular mathematical function for the trace line, certain parameters could be fixed, the fit of the model to actual data in any particular case could be tested, and more information about item behaviour provided.

Because the assumptions are stronger, the older approach of **classical test theory** (CTT) is sometimes called 'weak' true score theory to differentiate it from the later **item response theory** (IRT), sometimes referred to as latent trait theory (see Embretson & Reise, 2000). IRT is not a single theory with only one model for the trace line (or item characteristic curve, as it is termed in IRT); it is a family of models making different assumptions and seeking to estimate different parameters of the function. For example, some attempt to estimate only one parameter of the function; that is, where the ICC is positioned above the horizontal axis. Models of this sort are commonly referred to as 1PL (one parameter logistic) models as distinct from those (2PL) that attempt to estimate both position along the X axis and rate of rise of the function. A 3PL model attempts to estimate both position and rate of rise parameters, as well as how far up the vertical axis the ICC begins.

classical test theory
the set of ideas, expressed mathematically and statistically, that grew out of attempts in the first half of the twentieth century to measure psychological

variables; and that turns on the central idea of a score on a psychological test comprising both true and error score components

item response theory
(IRT) a family of theories that seek to specify the functional relationship between responses to a psychological test item and the strength of the underlying latent trait; in this functional relationship, it is expected, for example, that high ability students will have a higher probability of getting a difficult item correct than low ability students; IRT is used to develop and evaluate test items and tests, and underpins computerised adaptive testing and most large scale testing administrations

The Rasch model is an example of a 1PL IRT model that has been used to good effect in test development in Australia and elsewhere. The Undergraduate Medical and Health Sciences Admission Test (UMAT) used to assist with the selection of students into the medicine, dentistry and health science degree programs at undergraduate level at a number of Australian universities was based on the Rasch model. One of the early applications of the model was in testing for selection for the Australian Army (Anderson, Kearney & Everett, 1968). The Woodcock-Johnson Test of Cognitive Abilities III and the Stanford-Binet 5 (see Chapter 13) are tests that are now based on the Rasch model.

IRT models are clearly more sophisticated than that provided by CTT and require a lot more technical expertise for their application and interpretation. One might well ask, why bother when a simpler model is available? There are several answers to this (see Thomas, 2011). IRT approaches promise a better level of measurement (genuine interval measurement) and a means for determining whether this is achieved or only claimed. When test items are known to fit a model such as the Rasch, for example, the results obtained for individuals do not depend on the particular set of items used for the test but stand for all items that can be shown to fit the model. This means that rather than administer the entire test to all testees, a few items can be used to identify the position of a respondent on the underlying trait, and this assessment can then be refined by choosing items appropriate to that trait position. Two individuals can then be compared in terms of the trait without having completed the same items. This is the essence of tailored testing, used to good effect in computer adaptive testing, which brings considerable efficiencies in terms of time saved in practical testing situations.

A further reason is that IRT makes possible a more searching examination of differential validity. Groups (e.g. males and females; younger and older; or ethnic minority and ethnic majority members) can be compared on how they respond to particular items. This can then help to determine if there are important differences, which might mean that the test is inappropriate for use with these groups. A model, for example, that fits the data for European Australians but not

for Indigenous Australians means that the two groups are responding differently to the items (a concept known as **differential item functioning**, or DIF), and that comparing scores from members of the two groups should not be undertaken (see the section 'Cultural differences, testing and assessment' in Chapter 2).

differential item functioning
the possibility that a psychological test item will behave differently for different groups of respondents

With the advent of software to solve the estimation process, IRT models are being used more often in test development and are becoming the standard for test construction in large-scale studies.

This consideration of types and models of measurement allows some choices to be made at this phase of test construction. If a scale with ordinal properties is considered fit for purpose, CTT is the appropriate option. If, however, interval level measurement is sought then IRT and the Rasch model is the better option.

Item writing and editing

At this stage in test development, with the attribute defined and the model of measurement selected, the items for the test need to be written. The time spent in specifying the attribute begins to pay off here. A blueprint or **plan for item writing** is required that stipulates the number of items, the types of items and the areas the items are to be drawn from, all of which requires adequate attribute specification. For example, in developing a cognitive test, the plan would specify the functional areas to be covered (e.g. word knowledge, numerical ability and spatial reasoning). The test may be broader or more limited, but what it will be testing needs to be specified. In the case of a test of depression, the symptom clusters that are to be covered and the relative importance of these would be specified (e.g. thoughts, emotional experiences and problem behaviours). This determines the number of items for each to be included. The blueprint derives from the initial construct specification, and should be an explicit, written document.

plan for item writing
a plan of the number and type of items that are required for a test, as indicated in the test specification

In deciding the item type or types there are a number of options, although these are usually specific to the domain of the test. Figure 6.4 lists examples of item types for ability, personality and attitude tests. In the case of cognitive tests, multiple choice is widely used. The correct answer is to be selected from a number of options, frequently four because a respondent who did not know the correct answer could guess and get the answer right 25 per cent of the time, rather than 50 per cent of the time if there were only two options. Whether people answer completely at random or use what knowledge they have to narrow the possibilities is a moot point. Although providing four options is common for multiple choice tests, a case can be made for three as the optimal number (see, for example, Rodriguez, 1997, 2005). The true/false format is more common in the case of personality tests, and in the case of attitudes, a form of **Likert scale**. (Named after Rensis Likert, who was the first to propose the scale, 'Likert' is always spelt with an initial upper-case letter.) Likert scales with a five- or seven-point response format are most widely used, but again there are arguments about the optimal number of scale points (e.g. Preston & Colman, 2000).

Likert scale
a graphical scale originally with five points used by a respondent to represent the strength of an underlying attitude or emotion

Figure 6.4 Examples of item formats

Ability

- Complete the following series: 2, 4, 8, 16, …

- What is the capital of the Northern Territory? (Tick the correct option)
  a. Alice Springs
  b. Darwin
  c. Tennant Creek
  d. Katherine

Personality

- Indicate which of the response options best reflects your beliefs or feelings:
  I like motorbike racing. True False

- Children should be seen but not heard. Yes No Unsure

- Without any reason,
  I find myself in tears. Frequently Sometimes Rarely Never

- I would describe myself as (circle those that best describe you): warm, shy, anxious, outgoing, moody, light-hearted, solitary, lonely, hard-working, punctual.

Attitude

- Global warming presents a distinct danger to the planet.

Strongly agree	Agree	Neutral	Disagree	Strongly disagree

Irrespective of item type, writing good items requires creativity, and some test developers are better at it than others. Partly for this reason, a panel of item writers is usually employed. These are experts in the sense that they have a good knowledge of the content area of the test (e.g. mathematics teachers for a test of mathematics knowledge or clinical psychologists for a test of depression) and also have some appreciation of what is comprehensible and meaningful for the target audience. The panel is assigned the task, but needs to be oriented to it with a statement about the construct being targeted in the test (based on the construct specification) and some example items.

The example items, and all the items to be eventually included in the test, need to conform to the best standards of expression (e.g. clarity and succinctness), including grammar and punctuation. They need to be logically correct, and to be as brief as is consistent with transmitting the necessary meaning. Good items do not use specialist language, and they focus on the important and not the incidental aspects of the construct. They should not be overly complex, seek to trap or trick the respondent, or trivialise the task by the use of humour. They should not use fashionable terms that can date quickly, or colloquialisms that can restrict their meaning. Readability needs to be appropriate for the education level of those who ultimately take the test. Where multiple choice is used, the options provided should be less correct than the right answer, approximately equivalent in their attractiveness or plausibility, and be of the same or approximately the same length. Negatives (e.g. 'not') and exclusives (e.g. 'except') need to be used cautiously because they can lead to ambiguity. The use of the phrase 'none of the above' as one of the options in a multiple choice test is often criticised, but empirical studies of item options justify its use (Rodriguez, 1997).

Once a pool of items has been gathered, the items need to be edited by the test developer against the principles specified in the preceding paragraph, and a question order established. Items of a particular type or content area are usually, although not invariably, kept together in a test, and the most difficult items on cognitive tests are included towards the end to maintain motivation early in the test. An answer key needs to be developed that specifies which answer to a question signifies the construct in question, and suitable instructions—written and oral, depending on how the test is to be used—need to be formulated to guide the test taker. Comprehensibility and comprehensiveness of the item pool, and correctness of the answer key, can then be checked by an expert panel, preferably not the one used to write the questions in the first instance.

One further step is pilot testing with a sample from the population for which the test is being developed. Here the interest lies in the reactions of the sample to each of the questions, and can be achieved using focus groups, where difficulties with the items—ranging from wording to cultural appropriateness—can be identified.

Item analysis and selection

Item analysis involves a qualitative and quantitative examination of the items in the test following its administration to a relatively large (say, 100+) sample from the population for which the test is intended.

item analysis
the process of studying the behaviour of items when administered to a group of respondents, usually with a view to the selection of some of the items to form a psychological test

The items will have been reviewed and edited prior to administration using local experts and a pilot (i.e. a small-scale sample) study with members of the intended population. The larger sample would be asked, once they have completed the test, to comment on the items in the test in terms of such things as their readability, comprehensibility, clarity and apparent strangeness. This qualitative information can be used in selecting items for the final version along with the quantitative data collected from administering the test.

The focus in quantitative item analysis is on how the items 'behave' when people are asked to complete them. The analysis of item behaviour depends on the measurement model chosen in the earlier step in test construction, but typically the focus is on item difficulty and item discrimination. In the case of the Rasch model, item discrimination is assumed to be constant and the interest is item difficulty. An example of a CTT-based item analysis is presented in Box 6.1 and more information on item statistics is provided in the Technical Appendix. Item difficulty (p) is the frequency with which the correct answer is endorsed in the sample. Item discrimination is the extent to which an item differentiates between those with more of the attribute of interest and those with less. It can be computed in a number of ways, as shown in the Technical Appendix.

There are a number of other indices that have been used from time to time for item analysis. One that needs special mention is the index referred to as item validity.

Item validity usually refers to the correlation between an item and score on an external criterion being used to validate the test. By selecting items with high item validity, the correlation between the score on the final version of the test and the external criterion should be maximised. This was a common strategy when validity was thought of only in terms of criterion validity and external keying was the method of choice for test construction, neither of which is true anymore. If a test is being developed for one highly specific use, then attention to item validity might make sense, but to the extent that one seeks to maximise the correlation with one criterion measure, one might be reducing the correlation of the test with another criterion that is only moderately correlated with it, and thereby reducing the value of the test. To develop an aptitude test to predict success in first-year university psychology by selecting on the basis of the correlation of items with the result of a first-year psychology examination may produce a test with reasonable predictive value against this criterion. However, to the extent that results in psychology do not correlate well with results in, say, first-year cultural studies, then one could be doing a disservice to those seeking

advice on their aptitude for university study. Again, one needs to specify in advance the construct that one is seeking to measure before the exercise begins, and let this specification guide all the decisions that are made along the way. Viewed from this perspective, item validity may have only curiosity value.

item validity
the extent to which the score on an item correlates with an external criterion relevant to the attribute or construct that is the subject of test construction

Box 6.1

Item analysis

Imagine you have prepared a five-item multiple-choice test (four options for each question) and have administered it to a sample of ten participants. The problem is hypothetical and for the purposes of demonstration only. An actual item analysis would involve a larger item set (20+, typically) and a larger sample (100+). The item data from the imaginary exercise are summarised in Table 6.1.

The table lists the responses of each person to each option for each item. The number 1 indicates that an option was selected and 0 that it was not. Thus, Person 1 answered the first item by endorsing option a, the second item by endorsing option a, the third item by option b, and so on. At the bottom of the table is the mean endorsement for each option for each item and the variance (sample value and not population estimate) for the options that are being scored as 'correct'.

The first step in the analysis is to examine the popularity of each of the responses, and, for a multiple-choice test such as this one, the attractiveness of the options that are used as 'distracters'. In the case of a cognitive test, distracters are options that anyone who knows the right answer to the item would not choose but which might seem correct to someone who does not.

For Item 1, the correct answer has proved quite popular (endorsement of 70 per cent) and the participants who did not choose the correct answer have spread their responses over the remaining three options equally. This is a relatively easy item (high proportion of correct answers) and one for which the distracters are working well. Item 2 is a very easy item (the correct answer has a 100 per cent endorsement), but as it stands would be rejected because of its high endorsement. The beginning assumption in test development is that individuals differ. The purpose of development of the test is to assess these differences. An item that does not show differences (as with Item 2) is therefore

not useful. Where endorsement is too high (more than 90 per cent of respondents) or too low (less than 10 per cent), the item is usually rejected.

Item 3 is a more difficult question, with a response rate of 50 per cent for the correct option, but two of the distracters have not been chosen at all. This item needs further consideration, because as it is currently written there is one option that is attractive but wrong. This could be because there is some ambiguity in the wording of the item. Item 4 is a slightly more difficult item than the previous one, but here the distracters are all working well (all options are achieving reasonably equal endorsement). The same could be said for the final item. At this stage, one item (Item 2) would be rejected, three accepted and one held for further consideration.

The next step is to analyse the intercorrelations among the items and the correlations of each of the items with the sum of all of the items. The intercorrelation analysis is straightforward (each item is correlated with every other item), but the item-total correlation analysis deserves some comment. The correlation is between the item and the sum of all the items. This sum is the best estimate of what is common to all the items in the set. It is assumed that while there are likely to be some problem items in the set, the sum over all items is a reasonable first approximation to the attribute to be measured. The item writers, working from a reasonably tight specification, should have been able to generate at least a reasonable number of good items, and summing over the complete set should even out to some degree the limitations of individual items. In calculating the sum, each item is in turn deleted to give what is sometimes called the corrected item total. If the item itself were included when correlating an item with the total of all the items, there would be an artificial inflation of the correlation, because the item would be part of that with which it is being correlated, and this must produce some level of correlation. So, in determining the item total correlation for each of the items, each item is in turn deleted from the total of all the other items.

Table 6.1: Fictitious item data for a five-item test each with four options administered to ten individuals

	Item 1 options	Item 2 options	Item 3 options	Item 4 options	Item 5 options
Person	a b c d	a b c d	a b c d	a b c d	a b c d
1	1 0 0 0	1 0 0 0	0 1 0 0	0 1 0 0	0 1 0 0
2	0 1 0 0	1 0 0 0	0 1 0 0	1 0 1 0	0 0 1 0
3	1 0 0 0	1 0 0 0	0 1 0 0	0 0 0 1	0 1 0 0
4	1 0 0 0	1 0 0 0	1 0 0 0	0 0 0 0	1 0 0 0

	Item 1 options	Item 2 options	Item 3 options	Item 4 options	Item 5 options
5	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0
6	0 0 1 0	1 0 0 0	0 1 0 0	0 0 1 0	0 0 0 1
7	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	0 0 1 0
8	1 0 0 0	1 0 0 0	1 0 0 0	0 0 0 1	0 0 0 1
9	1 0 0 0	1 0 0 0	1 0 0 0	0 0 0 0	1 0 0 0
10	0 0 0 1	1 0 0 0	0 1 0 0	1 1 0 0	0 0 0 1
Mean	0.7, 0.1, 0.1, 0.1	1 0 0 0	0.5 0.5 0 0	0.4 0.2 0.2 0.2	0.3 0.2 0.2 0.3
σ2	0.21	0	0.25	0.24	0.21

Note: option 'a' is the correct option in all cases.

Table 6.2 presents the corrected (item deleted) and uncorrected item total correlation. (The intercorrelations of the items themselves would also be examined, but in the interests of space have been omitted. You may wish to calculate them.) The uncorrected correlations are included simply to make the point that they are different—and, with such a small item set and small N, not surprisingly they are substantially different.

Table 6.2: Corrected and uncorrected item-total correlations for the data in Table 6.1

	Item-total correlations	
Item	Uncorrected	Corrected
1	0.66	0.36
3	0.90	0.75
4	0.23	–0.17
5	0.77	0.52

Note that in Table 6.2, Item 2 has been omitted. Recall that this was the item that all ten participants answered correctly. There is no variance for this item and therefore no correlation is possible with any other variable.

Inspection of the corrected item total correlations in Table 6.2 provides the next significant piece of information in item analysis. Item 2 has already been

discarded. Items 1 and 5, which were easy and more difficult items at the first stage, are found to correlate reasonably well with the sum of all the other items. Item 3, which was suspect because of its pattern of responding across distracters, is found to have quite high item total correlation. This suggests that it should be kept and the distracters reformulated or reworded. Item 4, which appeared a reasonable item at the first stage, is found here to have a negative correlation with total score. This item is not measuring what the other items in the set are measuring and therefore will have to be discarded.

Thus we finish with three items, one of which requires further work. If we wanted a five-item test we will now need to go back to the item writing stage and then do a further item analysis. Items are typically lost in item analysis and hence the pool of items should be made larger at the outset to allow for this.

You may wish to calculate the alpha coefficient for this three-item test and then apply the Spearman-Brown formula (see Chapter 4) to determine the number of items you would need to add to take the alpha to, say, 0.90.

On the basis of an evaluation of the item statistics and qualitative information about the items, a subset will be selected as the most useful for the test. If the subset is smaller in number than the number considered desirable for the final version of the test, then more item writing and item analysis will be necessary. A final step is to examine the distribution of total scores on the test developed. It may be that the distribution is skewed (too many easy or difficult items), which will make the test unsatisfactory for use with the population in mind. Further item selection might be necessary at this point.

Assessing reliability and validity

The subset of items that have been selected is then evaluated for reliability and validity. Two techniques used in this regard—certainly when a CTT measurement model has been adopted—are exploratory factor analysis and determination of internal consistency reliability (see Chapters 4 and 5). If only one construct has been targeted in the test then EFA should show one strong factor. If more than one construct is being examined in the test then more than one factor should emerge with notable factor loadings for items as specified in the construct specification and test plan. Such patterning does not prove that the test is valid, but a lack of such patterning would question its validity. As for reliability, Cronbach's alpha would be calculated and evaluated in the light of guidelines for test use as described earlier (Chapter 4). If the test is found wanting then further development is called for, which might include writing new items or re-specifying the construct.

The data gathered for item analysis can be used for this purpose, but it is important to collect data using independent samples to ensure that chance effects are not confounding the conclusions being drawn about the test. With only a moderate number of items, there is the possibility that, with the number of correlations being calculated, some of these are due to sampling error. Using a number of representative samples allows one to check the replicability of the findings, and provides increased confidence that the decisions being made about the test are sound.

Work on reliability and validity of a test is never complete. As users apply the test and information about it accumulates, a more balanced evaluation of it becomes possible and aspects of validity or lack of it for particular purposes become clear.

Norming the test

With a test of satisfactory reliability and validity for the purpose for which it was devised, the test developer has two further tasks to complete, if the test is to be used professionally. It may be that the purpose of constructing the test is to use it in research, and in this case the next two steps are not required. But if the test is to be used by others for decision making, then relevant and representative norms for the uses to which the test will be put need to be developed, and a manual for using the test needs to be prepared.

The features of norms were discussed in Chapter 3. Representativeness is of course the key consideration, and this depends on a clear answer to the question: representative of what? What is the population to which the test user is likely to want to compare a score on the test for an individual? With some constructs—general mental ability, for example—the comparison is often to the population at large. For other constructs, there may be a particular sub-population that is important. A test of suicide potential, for example, might be developed for use with patients with a diagnosed mental disorder in an in-patient setting, or the test might be for memory in patients with dementia. The norms in these cases should represent the respective patient populations, rather than the general population.

In developing norms, it is often the case that factors known to correlate with scores on the test are explicitly included. Gender and age are two variables that correlate with measures of a number of psychological constructs, albeit only at a modest level, and for that reason are often explicitly included in preparing norms. When we say 'explicitly included', we mean that norms are prepared in such a way that these variables are identified in the tables that are prepared. They will almost always be implicitly included in developing a sampling plan for collecting norms. When age and gender are explicitly included, the test user can

base interpretation of the extremity of a score on its deviation from the mean for the age group most similar to, and the gender of, the individual tested. Although this is helpful for the user, it does increase the work involved in norming the test. It is not one mean that is now of interest but a set of means, one for each of the groups formed by the cross-break of the two variables. For example, if separate norms are to be developed by gender and age, at a minimum the means for four groups will need to be found. Gender is fixed at two levels and age could be reduced to two levels (old and young). Therefore, there are 2 × 2 or four groups. But a split into old and young is a very coarse treatment of the age variable for most purposes, and three to five levels are more realistic, which would make for up to ten groups (2 × 5). It was noted in Chapter 3 that from 200 to 500 participants are needed to estimate the mean with reasonable accuracy; Kline (1993) proposed 300. This means that a total of from 4 × 300 (1200) to 10 × 300 (3000) participants will be needed, depending on how coarse a grouping on age is acceptable. If separate norms are not provided for age and gender, then 300 participants would be sufficient, although one would normally need to sample in such a way that these variables are adequately represented in the norming sample.

The reader is referred to the discussion in Chapter 3 and to more advanced texts (e.g. Pedhazur & Schmelkin, 1991) regarding the issues that need to be considered in developing a sampling plan for the collection of normative data. Careful development of test items and comprehensive work on the validity of the test for given purposes will be compromised by poor sampling for the establishing of norms. Serious users will quickly identify problems relating to poor or inadequate sampling and either use an alternative or wait until these problems are corrected.

With normative data to hand, the decision needs to be made on how best to present it; that is, what form of transformation of the raw scores on the test needs to be made to best communicate the required information (refer to Chapter 4). Frequently, both standard scores (or some whole number transformation of them) and percentiles are provided to maximise the information available to the user. Tables are then prepared with the transformed values for all possible raw scores so that the user can read off the appropriate transformation once the raw score on the test has been computed.

Publication

The final task, if the test is to be used professionally, is to prepare the test for publication. If this is being done commercially, the test publisher will be of considerable assistance at this stage, but even here the test developer remains responsible for the decisions made. Some of these decisions concern how the test

will be made available to potential users. For example, what materials will be used for the test items to maximise their readability, durability and professional format? Will they be included in a kit? If so, what form will it take? Will the test user be able to carry it about easily? In the age of computer testing, a somewhat different set of questions arises about the optimal presentation of material on computer screens and the ways answers are recorded, and security of test material becomes an even more significant issue. These are not psychological decisions as such, but they can have an important bearing on test use and are therefore important.

A second set of questions arise with respect to the test manual to accompany the test. This will outline the way in which the test was developed, indicating the theoretical account of the construct relied on, how items were constructed, the item analysis procedures followed and criteria employed in selecting items, and the data currently available on reliability and validity, as well as, of course, the normative data obtained. The manual must provide instructions for administration of the test, including any time limits that need to be observed and how the test is to be scored. The populations for which the test is appropriate need to be specified, including any requirements of those taking the test; for example, the upper and lower chronological ages for which the test is appropriate, and the reading age necessary to understand items. The qualifications necessary for test users to interpret test scores need to be clearly stated. The limitations of the test should be admitted and caution expressed about any ways in which it could be foreseen the test could be misused. If there are published data on the test, reference to these should be included, or an indication given that summaries of unpublished work are obtainable from the author of the test. Preparation of an adequate manual is a significant exercise in psychological writing and could run to the length of a small book. As Cronbach (1970, pp. 118–19) put it: 'The manual must be clear enough that any qualified user can comprehend it—and clear enough that the reader who is not qualified will realise that he (she) is not. Yet the information must be precise enough to satisfy specialists in test research.'

test manual
the document that accompanies a psychological test and that records the way in which the test was developed, how the test is to be administered (including the groups for which it is relevant), information on the reliability and validity of the test when used for specific purposes, and norms for test interpretation

Box 6.2

A test of retrograde amnesia

Loss of memory has an impact on well-being and the enjoyment of life, and can be an indication of central nervous system (CNS) damage or disease. Diagnostic tests of memory loss have concentrated on identifying what is termed 'anterograde amnesia'; that is, memory loss resulting from insult to the CNS that involves events occurring in the patient's life subsequent to the damage or onset of the disease process. For example, following a motor vehicle accident, a person may have difficulty remembering day-to-day events that happen to them. Retrograde amnesia, on the other hand, involves loss of memory for events prior to the CNS insult; for example, events in the early years of the person's life, well before the accident.

Shum and O'Gorman examined the literature on retrograde amnesia and found no published tests suitable for use in Australia. Although psychological tests can often be used with good effect in countries other than where they were first developed, given a common language and similar culture, in the case of tests of retrograde amnesia the problem of cultural difference becomes particularly acute. Although one could ask about particular events in a person's early life, these will differ from individual to individual and there is frequently no one who can verify the answers the person gives. For this reason, tests of retrograde amnesia commonly use statements of events or faces of people that would generally be known to those who have lived through a particular period. Choice of the events or faces is critical because, if they are too obscure, failure to recognise them might reflect lack of knowledge in the first place rather than loss of memory. If, on the other hand, they are too well known, recall of them reflects general knowledge rather than a specific memory. For example, the face of a past president of the USA might be a useful item for checking memory for previous events in a citizen of that country, but might not be of use for an Australian population.

Having specified the construct of interest and checked the literature thoroughly, Shum and O'Gorman embarked on an exercise in test construction, some of the details of which are reported in Shum and O'Gorman (2001). A pool of 90 famous faces and 90 public events relevant to the decades between the 1930s and 1980s was compiled from a number of sources, chiefly published photographs from newspaper or magazine stories. The item pool was administered to a sample of 47 participants for item analysis. A number of criteria were used, including the difficulty level of the items and their item-total correlations. From this large pool, 54 famous faces and 54 public events were selected for the final version of the test. The Cronbach alpha for the Famous Faces and Public Events part of the test were 0.92 and 0.91, respectively.

Validation of the test relied on two principal criteria. The first was the relationship between age and memory in a group of participants without known CNS damage or disease. It was expected that memory for events in the remote

past would be poorer than that for more recent events, but that this would depend on the age of the person tested. Older compared with younger participants should have better recall of events and faces from the decades through which they had lived but the younger people had not. Shum and O'Gorman (2001) were able to show that this was the case. The second criterion was the sensitivity of the test to CNS disease known from other studies to affect memory for past events. The performance of patients with Alzheimer's disease or Korsakoff's syndrome—disorders different in aetiology—was compared with that of disease-free volunteers of approximately the same age. As predicted, the patient group showed greater memory loss on both parts of the test. Further data on the test has accumulated with its use by other researchers (e.g. O'Gorman & Shum, 2012).

Chapter summary

The steps followed in a rational-empirical approach to constructing a psychological test have been outlined and discussed: justification of the need for the test; specification of the construct or constructs to be tested; selection of a measurement model; item writing and editing to a plan for the test; item analysis and selection; assessment of reliability and validity; norming; and preparation of a test manual. To provide an example of the procedures in practice, a test construction project undertaken by two of the authors is outlined in Box 6.2. This is not meant to illustrate test construction at its very best, but it does serve to show how the steps come together when a real question is posed.

Questions

1. Define and give examples of S S Stevens' four levels of measurement.
2. In what ways is psychological measurement different from physical measurement?
3. What parameters in item response theory does the Rasch model specify?
4. What are the foundations on which test construction builds?
5. What is item analysis? What are some of the indices commonly used in item analysis?
6. Why is it necessary to have a manual for a test?
7. Is the plan for a test the same as a construct specification statement?
8. How would an operational definition of anxiety differ from a construct definition?
9. What variables are important to include in a sampling plan for a test of general ability, and why?

10. What is a distracter analysis and how is it done?

Exercises

1. A psychological test has sixteen items. The mean, SD, and item-total correlation for each is as follows:

0.13, 0.33, 0.03	0.21, 0.41, 0.36	0.11, 0.31, 0.16	0.15, 0.36, 0.01
0.11, 0.32, 0.29	0.08, 0.28, 0.15	0.23, 0.42, 0.15	0.01, 0.13, 0.28
0.11, 0.37, 0.34	0.08, 0.27, 0.02	0.01, 0.12, -0.23	0.11, 0.31, 0.24
0.06, 0.24, 0.21	0.19, 0.39, −0.19	0.10, 0.30, 0.25	0.01, 0.09, 0.03

   a. Compute coefficient alpha. Assume that these are 'true/false' items and perform an item analysis of the test, to the extent that this is possible with the data available.

   b. If the desired reliability of the test is to be 0.90, would you need to add items to it after your item analysis, and if so how many?

2. Write a five-item test of social desirability. How would you set about testing its validity?

3. The following have been offered as items for a test of general mental ability for use in Australia. Would you use them and, if not, why not?

   a. What is the population of the southern-most town in New Zealand? 10,000 people or more than 10,000 people?

   b. The prime minister before the prime minister who was the prime minister before the present prime minister was or was not John Hewson?

   c. It is not the case that a ball is not out in tennis if it is not outside the line. True or false?

   d. Complete the following number series: 20, 30, 40, 50 …

4. In scales of what types of constructs might the following items be included?

   a. I have recently returned from Switzerland where I have been repairing cuckoo clocks. T/F

   b. I never gossip. T/F

   c. I feel tense most of the time. T/F

5. Evaluate the following items for inclusion in a personality test:

   a. I sometimes forget things but most of the time my memory is better than most people older than me. T/F
   b. I need people to make a fuss of me. T/F
   c. Sometimes I find myself doing things that I wouldn't normally do and could not explain why I was doing them if someone were to ask me or to suggest that I should not.

6. Evaluate the following items for inclusion in a general knowledge test for university students.

   a. The Lithuanian coastline is on which sea?
      i. Caspian
      ii. Adriatic
      iii. Dead
      iv. None of the above.

   b. What is the meaning of:
      i. Apocrypha
      ii. lettuce
      iii. hormone

   c. Climate change is caused by global warming and industrial development T/F
   d. Who was the guitarist in ACDC who was born in Glasgow?
      i. Scott McCampbell
      ii. Angus Young
      iii. John Noble
      iv. None of the above

Further reading

Osterlind, S J (2006). *Modern measurement: Theory, principles, and applications of mental appraisal*. Upper Saddle River, NJ: Pearson.
Rust, J & Golombok, S (2008). *Modern psychometrics: The science of psychological assessment* (3rd ed.). London, UK: Routledge.
Schmidt, K M & Embretson, S E (2003). Item response theory and measuring abilities. In J. A Schinka & W F Velicer (Eds.), *Handbook of Psychology: Vol 2, Research Methods* (pp. 429–46). Hoboken, NJ: John Wiley & Sons.

Useful websites

Assessment Psychology Online:
www.assessmentpsychology.com/psychometrics.htm

Institute for Applied Psychometrics (IAP): www.iapsych.com/IAPWEB/iapweb.html

PART 3 SUBSTANTIVE TESTING AND ASSESSMENT AREAS

7 Intelligence

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. identify the important historical steps taken to develop an understanding of human intelligence
2. understand the controversies surrounding the conceptualisation and measurement of intelligence
3. define the content of important tests of intelligence
4. differentiate between aptitude and achievement tests
5. describe and contrast the most recent theories of intelligence
6. explain what aspects of intelligence are measured (and not measured) by modern tests of intelligence.

KEY TERMS

achievement test
aptitude test
CHC theory of intelligence
crystallised intelligence (Gc)
culture fair test
deviation IQ
fluid intelligence (Gf)
Flynn effect
'g' (general mental ability)
IQ (intelligence quotient)
multiple intelligences
's' (specific ability)
triarchic theory of intelligence
two-factor (Gf-Gc) theory of intelligence
Wechsler Adult Intelligence Scale

Setting the scene

- Everyone has an *implicit* theory of intelligence: a personal understanding of how intelligence is structured, how it develops and how it can be defined. In this chapter, you will come to appreciate how these implicit theories differ from *explicit* theories, which are theories of intelligence developed by psychologists.
- Early measures of intelligence assessed very narrow human sensations and reactions, and researchers attempted to relate sensory acuity and speed of responding to intellectual functioning. Unfortunately, these approaches failed. You will find out why.
- For most of the history of intelligence testing, the main contentious issue was whether intelligence should be thought of as a global construct or considered as a cluster of specific cognitive abilities. Reflect for a moment; what is your position? You will find out how this dilemma was resolved.
- The important results obtained from most modern tests of intelligence are measures of broad cognitive abilities, rather than a measure of a single IQ score. Why do think this is so?
- The topics of intelligence and intelligence testing still elicit strong emotions in many people. Do you have strong feelings—negative or positive—about intelligence testing? You will learn about the issues that continue to fuel the controversies in these areas.

Introduction

No construct in psychological measurement generates as much debate and controversy as that of intelligence. Psychologists and lay people alike have few concerns about defining and measuring other psychological constructs, such as personality, attitudes and interests. However, they are challenged, perplexed and even distressed when contemplating intelligence, intelligence tests and intelligence test scores, and how these are applied. The controversies surrounding intelligence and its measurement are rooted in both misconceptions and genuine concerns about how intelligence scores were used in the past and are used today. One important misconception is that an intelligence test score, or **IQ (intelligence quotient)**, is an all-encompassing, stable summation of a person's worth. In reality, IQ scores are narrow measures of specific sets of abilities, which vary across the lifespan because of education, personal experiences and motivation, and are poor global summaries of individual value (Gregory, 1999). Genuine concerns are grounded in beliefs, sometimes well justified, that intelligence tests are discriminatory, and disadvantage some groups in the areas of employment, education and access to civic benefits. Minority ethnic and racial groups, women and people with a disability are at particular risk of discrimination, and can be further disadvantaged by the inappropriate application of intelligence test scores (Jensen, 1980).

IQ (intelligence quotient)
the overall intelligence score obtained from one of the many current intelligence tests; the IQ score is a raw score conversion drawn from the normative sample, which has an arbitrary set mean of 100 and an arbitrary set standard deviation of 15 for each age group

Despite these misgivings, intelligence tests are widely used in Australia and overseas. Several surveys in Australia have found that a large majority of psychologists regularly use tests of intelligence for such diverse purposes as assessing learning difficulties and developmental disabilities, assisting vocational choice, and quantifying day-to-day functioning problems (Meteyard & Gilmore, 2015; Thompson et al., 2004). Intelligence testing also continues to fascinate the public at large. In 2002, an Australian commercial TV channel aired the *National IQ Test* program, which allowed viewers to respond to questions and calculate their own IQ score. This program was the most watched TV show for that year, with approximately 3.5 million viewers. It rates as one of the most watched TV programs in Australia (Stough, 2002).

The concept of intelligence

Everyone has an opinion or theory about the nature of intelligence. These 'everyday' or **implicit theories of intelligence** (also known as lay theories) reflect personal definitions and assumptions about how intelligence is structured, its component parts, the processes underlying intelligence, and how it develops and changes. Implicit theories of intelligence are constructed by individuals, and are affected by culture (e.g. modesty, politeness and respect appear in lay conceptions in India; Baral & Das, 2004), age (e.g. children stress academic skills, such as reading well and doing well in class; Yussen & Kane, 1985) and experience (e.g. university academics from different disciplines have somewhat different perceptions; Sternberg, 1985a). Implicit theories differ from **explicit theories of intelligence**, which are constructed by psychologists and other social scientists, and are based in empirical research that tests hypotheses about the nature of intelligence.

implicit theories of intelligence
models or schema of the construct of intelligence generated by individuals and based largely on their observations and opinions of how the world works

explicit theories of intelligence
theories of intelligence devised by psychologists and other scientists; the theories

grow out of and are validated using scientific methods, although they can be informed by implicit theories

So what is intelligence? The term itself, as it is employed today, did not come into use until after the development of the Binet-Simon test of intelligence in 1905. Baldwin's *Dictionary of Philosophy and Psychology*, published in 1901, for example, did not include an entry for intelligence. Very early thinkers and writers did contemplate the brain, the mind and intellect, but did not differentiate intelligence as a separate construct from other human characteristics, such as the soul, consciousness or willpower (Matarazzo, 1980). The early 'phrenologist', Austrian physician Franz Gall (1758–1828), considered the 'mind' to be the product of the physiological brain, with the strength of the different 'faculties' of the mind able to be determined by 'reading' (i.e. observing and feeling) their development in the contours of the skull (Simpson, 2005). In the second half of the nineteenth century, the Englishman Francis Galton (1822–1911) began administering a series of psycho-physiological tests that included 'sensory acuity' or efficiency tests, but which also included measures of a diverse range of human characteristics, such as length of arm, hair colour, reaction time and hand strength. Galton devised the statistical technique of correlation and used this to test for relationships among the variables he measured. In 1869, and based on his correlational analyses, Galton made reference to a 'general human ability' and 'special human abilities' (in the twentieth century, these would be called 'g', general mental ability, and 's', specific mental ability; Matarazzo, 1980). However, he did not move away from devising more and more sophisticated ways of testing simple human sensations and reactions, which he considered were at the heart of individual differences and ability. Other psychologists of this 'brass instrument' period (so called because many of the testing tools were of brass construction) emulated Galton's approach, but these attempts to understand human intellect and individual differences by measuring basic units of consciousness were unsuccessful, as scores from these basic units correlated poorly with important, real-world criteria, such as academic achievement. Towards the end of the nineteenth century, psychologists moved away from testing simple sensory responses and began testing more complex behaviours, including language and arithmetic proficiency, general knowledge, history and memory functions. This move away from assessing sensory responses to testing more complex behaviours foreshadowed the important developments initiated by Binet, Henri and Simon in the latter years of the nineteenth century.

Binet's revolution

Frenchman Alfred Binet (1857–1911) was educated as a lawyer, but became obsessed with the study of psychology and, being independently wealthy, he was able to pursue these interests for most of his working life. His revolutionary developments came from insights gained when he was testing his two daughters, which he did from their early to teenage years. As well as using tests of reaction time, Binet also assessed them with more 'complex' tasks that tapped language skills, reasoning and memory. These tasks included tests of word generation and recall, memorising written passages, letter cancellation and figure reproduction. Binet observed that performances on the more complex tasks were more variable, and more sensitive to developmental progress than the simple reaction time measures (Goodenough, 1949). In 1895, with his colleague Victor Henri (1872–1940), Binet published a paper, *La psychologie individuelle*, proposing an intelligence test that would assess ten higher-order 'mental faculties'—notably memory, attention, concentration and comprehension—as well as simpler motor functions such as muscle strength and hand– eye coordination. Ten years later, in collaboration with another colleague, Théodore Simon (1872–1961), Binet produced the first practical test of intelligence: the Binet-Simon Intelligence Scale (Binet & Simon, 1905).

Figure 7.1 Alfred Binet (1857–1911)

The Binet-Simon Intelligence Scale was intended to be used to identify children in French schools who required special education. The scale contained thirty individual tasks, which were ranked in order of difficulty, although all tests were meant to tap higher-order mental abilities. An easy task required the child to name various body parts, whereas more difficult tasks were to explain why two things were different, construct a short sentence from words supplied, and repeat from memory a string of digits. Binet and Simon used fifty children across five age groups (3 to 11 years) as their **normative sample** or comparison group.

These children had been identified by their teachers as of 'average' ability. Thus, if an 8-year-old child was able to successfully complete the tasks typical for 8-year-olds in the comparison group, that child would be considered to have an approximate mental age of 8 years. If the 8-year-old could only complete the tasks typical for 6-year-olds, then the mental age was two years behind their chronological age, and so on. The Binet-Simon scale was revised in 1908, and again in 1911. The 1908 version had an increased age range (from 3 to 15 years), and contained fifty-eight higher-order tasks. The real innovation for this version of the scale was the grading of the tasks in terms of age levels, which were based on the 75 per cent pass rate for the different age groups in the 200-strong normative sample. Thus, the mental age on this test was the highest age level that the child approximated on the test. The 1911 edition was extended to eleven age levels, including an adult level. It contained suitable adult tasks, and adults were included in the normative sample. The Binet-Simon test made no claim to assess individual faculties of intelligence, nor did it seek to assess basic sensory acuity or motor speed. Rather, the aim was to measure **global intelligence**, which for Binet and Simon was represented by reasoning ability, judgment, memory and abstract thinking (Matarazzo, 1980).

normative sample
tables of the distribution of scores on a test for specified groups in a population that allow interpretation of any individual's score on the test by comparison to the scores for a relevant group

global intelligence
the overall or summary ability of an individual, which might be represented as the Full Scale IQ in modern intelligence tests; in hierarchical models of intelligence, global intelligence (or 'g') sits at the top of the intelligence hierarchy

Spearman and 'g'

Binet's perception of intelligence as a global construct was consistent with the view of his contemporary, Charles Spearman (1863–1945), the influential English statistician who pioneered factor analysis procedures. Factor analysis is extensively used in scale development and analysis as it has the capacity to summarise underlying dimensions (or factors) that might exist in large data sets. In essence, factor analysis allows researchers to reduce large amounts of data, which could be test items or tests themselves, to more manageable chunks (see Chapter 5 for a more detailed description of this technique). Spearman proposed that intelligence could be represented by a general, underlying mental ability

factor, which he called 'psychometric g'—or **'g'** for short—and which he conceived to be some form of 'mental energy'. This model of intelligence as a global construct resulted from the observation that when multiple, **specific-ability tests** were correlated, they tended to be associated positively and tended to load on one statistical factor: that of 'g'. Thus, Spearman's model of intelligence comprised multiple specific abilities (such as mathematical reasoning, verbal skills and spatial perception), with all these specific abilities sharing common, or general ability, variance. Spearman concluded that about half of the variance in specific-ability tests could be represented by 'g', with the other half being accounted for by specific abilities related to the particular test (**'s'**) and error ('e'). Spearman called his a two-factor model ('g' and 's'), but it is commonly referred to as a one-factor model of 'g'. The implication of conceptualising intelligence as a general ability, or 'g', is that when it is measured, it can be represented as a single score. Early tests of intelligence, including Binet's initial measures, generated a single score to represent intellectual functioning.

'g' (general mental ability)
the common variance when the results of different tests of mental ability are correlated (sometimes referred to as 'psychometric g', 'Spearman's g' or the 'general factor')

specific-ability test
an individual test or test battery that is designed to assess specific or narrow cognitive abilities, rather than generate a measure of broader abilities or 'g'

's'
(specific ability) limited to a single or small number of tasks, as opposed to 'g', which is reflected in all mental ability tasks; all tasks require the application of 'g' and 's', and individuals differ on levels of both

Figure 7.2 Charles Spearman (1863–1945)

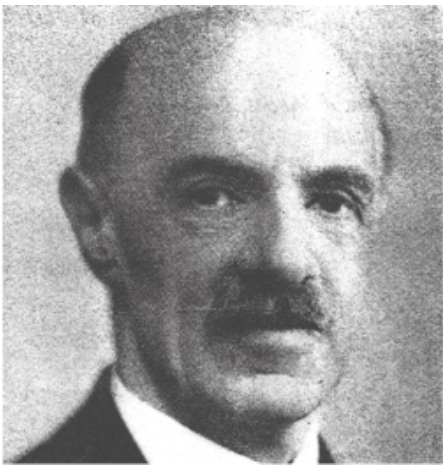
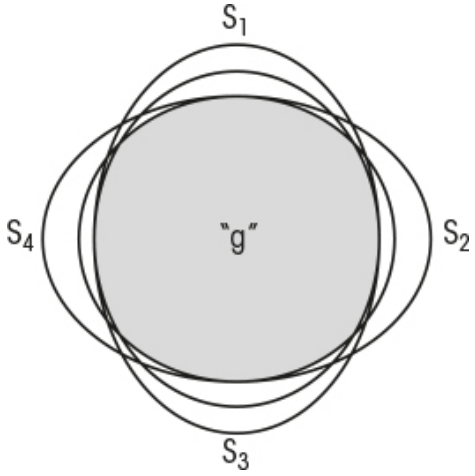


Figure 7.3 Spearman's theory of 'g' and 's'



Terman and the Stanford-Binet Intelligence Scale

Binet's test was widely used throughout Europe and in other parts of the world. It was translated into English in the US in 1908, and was substantially revised in 1916 by Lewis Terman (1857–1956), a psychologist working at Stanford University. This revision, known as the **Stanford-Binet Intelligence Scale**, proved to be very popular, and became the standard against which all other tests of intelligence were compared. The Stanford-Binet also formed the basis for many group-administered intelligence tests (e.g. Army Alpha and Army Beta), which were used initially to select recruits during the First World War, and which were later made available generally, giving a tremendous impetus to intelligence testing in occupational and educational settings. Terman retained Binet's view of intelligence as a global construct (i.e. the test was considered to assess general mental ability), and retained the age differentiation of items (i.e. clusters of items

for each age group, which could be successfully answered by between two-thirds and three-quarters of respondents).

Stanford-Binet Intelligence Scale
Lewis Terman of Stanford University revised the Binet-Simon test for use in the US; released in 1916, the Stanford-Binet Scale has been revised many times and continues to be widely used

Figure 7.4 Lewis Terman (1857–1956)

Terman included a much extended list of items (almost double to ninety items), a considerably larger standardisation sample (of 1000 children and 400 adults, compared with 203 children in the 1908 Binet-Simon test; although the US sample comprised white Californians only, and could thus not be considered representative), and provided detailed administration and scoring instructions. The test was now suitable to assess children and adults across the 'feeble-mindedness' to 'genius' range. An important reason for the popularity of the Stanford-Binet test was the use of the intelligence quotient (IQ) concept. This was based on the ratio between mental age and chronological age, where mental age (MA) was divided by chronological age (CA) and multiplied by 100 to remove decimals. A child with a chronological age of 8 years and a mental age of 10 years, for example, had an IQ of 125 (i.e. $10/8 \times 100 = 125$), whereas the IQ of a child with a chronological age of 10 years and a mental age of 8 years was 80. Being able to summarise intellectual functioning in such a fashion provided many advantages, although the use of this now-defunct conceptualisation of IQ was not sustainable. IQ is not distributed in the same way across all age groups. A 4-year-old with a mental age of 5 (IQ = 125) might not be similarly advanced as an

8-year-old with a mental age of 10 (IQ also = 125), as the variability in IQ at 4 years of age is greater than that at 8 years. This version of the intelligence quotient also cannot be applied sensibly to adults, as intelligence does not progress linearly across the lifespan. The Stanford-Binet test was revised in 1937, 1960, 1986 and 2003, and the 'modern' version remains one of the most widely used tests internationally.

Wechsler scales

The main competitors to the various versions of the Stanford-Binet Intelligence Scale were, and remain, the scales devised by David Wechsler (1896–1981). In the early 1930s, Wechsler, an employee at the Bellevue Hospital in New York City, developed an individual test of intelligence, the **Wechsler-Bellevue Intelligence Scale**, which was used initially to assess adult psychiatric patients. Wechsler (1939) considered intelligence to be 'the aggregate or global capacity of the individual to act purposely, to think rationally and to deal effectively with his environment' (p. 3). Thus, he considered the Wechsler-Bellevue test as a measure of global ability, even though the structure of the test made it possible to obtain scores on specific abilities. For Wechsler, intelligence was an 'all encompassing' facility, with the different specific abilities merely ways intelligence is manifested (Matarazzo, 1972). Wechsler stressed also that many non-intellectual factors, such as persistence and determination, contributed to the expression of intelligence, and while these factors were not formally measured, the administration of the test in standardised situations provided opportunities to observe the test taker's behaviour.

Wechsler-Bellevue Intelligence Scale
the forerunner to the popular Wechsler Adult Intelligence Scale, it was created by David Wechsler and released in 1939 as a test of general intellectual ability; revised many times, it remains the most widely used individual test of ability

Figure 7.5 David Wechsler (1896–1981)

The Wechsler-Bellevue test—and subsequent revisions as the **Wechsler Adult Intelligence Scale** (or WAIS) in 1955, the WAIS–Revised in 1981, WAIS–III in 1997 and the WAIS–IV in 2008—was devised specifically to assess intellectual functioning in adults. See Chapter 9 of this book for details and application of the latest version: the WAIS–IV. Wechsler was critical of many items included in the Binet scales, which were more suitable for use with children, although they were applied to adults. Responses to many of these questions were also overly dependent on speed of performance, disadvantaging older test takers, for example. As mental age norms were not appropriate for use with adults, Wechsler replaced these with point scales. Whereas items in the early Binet scales were grouped together according to age, Wechsler grouped his items according to content area (e.g. all arithmetic questions were grouped together in order of increasing difficulty), and test takers were credited a point for each correct answer they achieved. The 1955 WAIS test generated individual scores for eleven homogenous content areas. Point scales, rather than age scales, are now utilised by all modern intelligence tests. Point scales allow the use of **deviation IQ** scores, which are based on the assumption that intelligence is normally distributed in a population. When tests are normed on different age groups, an individual's score can be compared with others of the same age, and expressed in terms of standard deviations from the mean. The Wechsler scales (and most other tests of intelligence) use standardised scores with a set arbitrary mean of 100 and a standard deviation of 15; thus, an IQ score of 115 is one standard deviation unit above the mean, while an IQ score of 95 is one-third of a standard deviation below the mean. Deviation IQ scores also can be converted to percentiles, which provide additional information regarding the test taker's relative standing; for example, an IQ of 115 is at the 84th percentile, meaning that

the test taker's score is equal to, or higher than, 84 per cent of the comparison group.

Wechsler Adult Intelligence Scale
(WAIS) developed by David Wechsler, and one of the most widely used, individually administered, intellectual assessment batteries; the latest version, WAIS–IV, was published in 2008

deviation IQ
a method that allows an individual's score to be compared with same-age peers; the score is reported as distance from the mean in standard deviation units

The early Binet scales were also criticised for their over-reliance on items that assessed language and verbal skills. Wechsler addressed this problem by including a series of individual scales that tapped non-verbal abilities. These scales required test takers to respond in ways other than by using language, including having them point to a correct answer, copy provided symbols using pencil and paper, and assemble small coloured blocks according to specific instructions. The early Wechsler tests contained approximately the same number of these performance tests as verbal tests. Thus, Wechsler's test generated scores for each scale (e.g. vocabulary, arithmetic, block design and object assembly), a verbal score (i.e. total for all verbal scales), a performance score (i.e. total of all performance scales) and an overall score, or measure of 'g'. As all scales were standardised and normed on the same sample, the Wechsler scales made it possible to assess strengths and weaknesses at the individual scale score level, compare a test taker's verbal and performance abilities, and gain a measure of global intellectual functioning. Wechsler published parallel scales suitable to assess children in 1949 (Wechsler Intelligence Scale for Children, the WISC, revised in 1974, 1991, 2003 and 2014; this latest version is published in both pencil-and-paper and digital formats; Wechsler, 2014) and very young children in 1967 (Wechsler Preschool and Primary Scale of Intelligence, the WPPSI, revised in 1989, 2002 and 2012; Wechsler, 2012a). The most recent versions of the Wechsler scales (i.e. WAIS–IV and WISC–V) generate individual subscale scores and a global measure of intelligence ('g'). These modern versions have done away with the verbal and performance scores, and replaced them with indices; four for the WAIS–IV and five for the WISC–V. For the WAIS–IV, the indices are verbal comprehension (including vocabulary and general knowledge scales), working memory (including memory for numbers and mental calculations), perceptual reasoning (including puzzles and problem solving) and processing speed (i.e. speed of copying). The WISC–V generates indices for verbal comprehension, working memory, processing speed, visual spatial abilities and fluid reasoning.

These indices parallel those generated for the WAIS–IV, with the latter two indices (visual spatial abilities and fluid reasoning) largely overlapping with the perceptual reasoning index from the WAIS–IV. These indices represent more 'pure' measures of the individual scale clusters than would verbal and performance indices; and, as we will see later in this chapter, better reflect contemporary theories of intelligence, such as Carroll's (1993) stratum theory of intelligence. Chapter 9 provides a full description of the four indices for the WAIS–IV and their application; Chapter 13 has the same information for the WISC–V. In concert with the Wechsler scales, and consistent with theory development, the most recent version of the Stanford-Binet Intelligence Scale (Roid, 2003) also produces index scores (fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing and working memory), which are based on a series of ten individual tests, and which sum to give a global rating of intelligence. Unlike the Wechsler test approach, which uses three tests to assess children and adults (i.e. the WPPSI, WISC and WAIS), the Stanford-Binet test retains the capacity in one tool to assess the full age range of individuals, from 2 to 85+ years.

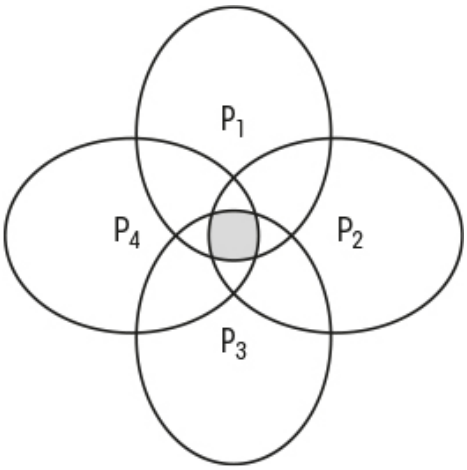
Thurstone and multiple mental abilities

While many early tests of intelligence, such as the Stanford-Binet test and Wechsler scales, generated a single score to represent intelligence, other theorists had proposed alternative models of intelligence. The most notable of these was by the US psychologist Louis Thurstone (1887–1955). When Spearman examined the correlations among different tests of ability, he found sufficient evidence of overlap to argue for global 'g'; when Thurstone did similar analyses, he considered the unique variance explained by the individual ability tests argued for multiple intelligences. Thurstone then proposed a multifactor theory of intelligence and, based on his own empirical work, identified seven main factors, which he labelled **primary mental abilities**. These primary abilities were verbal comprehension, reasoning, perceptual speed, numerical ability, word fluency, associative memory and spatial visualisation (Thurstone, 1938). Thurstone published the *Chicago Tests of Primary Mental Abilities* (Thurstone & Thurstone, 1941) to assess these individual abilities, although these were not widely used, and are not in use today. Thurstone later accepted that his primary mental abilities tests did overlap, and that 'g' did reflect a higher-order factor. In the same way, Spearman also acknowledged the importance of specific abilities, and the later positions of Spearman and Thurstone differed only on emphasis, with one giving more weight to 'g' and the other giving more weight to 's' (Ruzgis, 1994).

primary mental abilities
seven broad ability factors that were identified by Thurstone: verbal comprehension, reasoning, perceptual speed, numerical ability, word fluency, associative memory and spatial visualisation; initially thought to be independent of one another, they were later shown to be correlated, and thus to also contain a 'g' factor

Figure 7.6 Louis Thurstone (1887–1955)

Figure 7.7 Thurstone's model of primary mental abilities



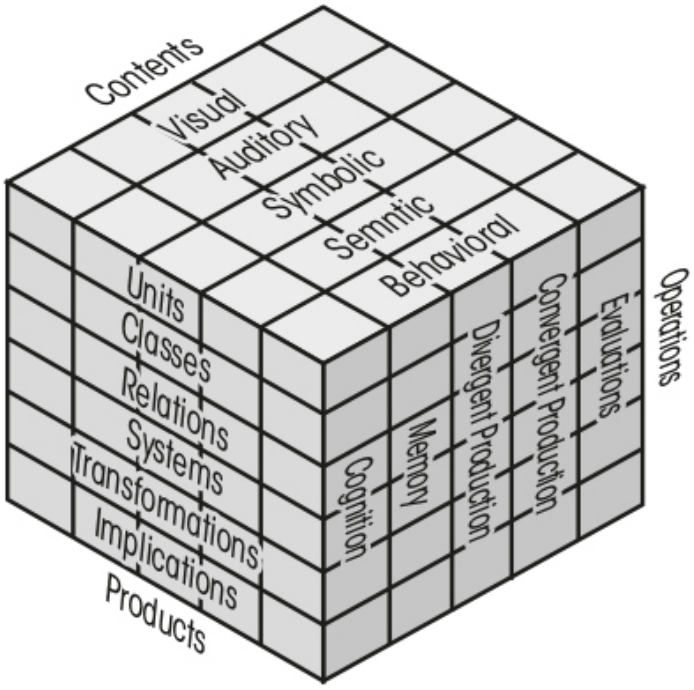
Guilford: A different structure of intelligence

J P (Joy Paul) Guilford (1897–1987), a US psychologist, dramatically expanded the number of factors considered for intelligence (Guilford, 1967, 1985). Guilford rejected the notion of a general factor of intelligence, or 'g'; rather, he proposed that intelligence be viewed along three dimensions: operations, content and product. In his **structure-of-intellect (SOI) model**, 'operations' refers to the type of mental processing required to complete a task (i.e. understanding, memorising, recalling and evaluating); 'Content' refers to the type of stimuli to be manipulated (i.e. visual, auditory, symbolic and affective); and 'Product' refers to the type of information that is manipulated and stored (i.e. a single unit of information, categories of information, information systems, relationships among units, information transformation and predictions). As there were five categories of operation, five of content and six of product, Guilford proposed a three-dimensional matrix of intelligence with 150 individual factors of intelligence (i.e. $5 \times 5 \times 6 = 150$). Later versions included even more factors (Guilford, 1988). While Guilford claimed to have confirmed the existence of a majority of these individual factors, the practical utility of such a complex model was questionable, and Guilford's model had little influence on test construction and theorising. Carroll (1993) uncharitably commented that Guilford's model should be 'marked down as a somewhat eccentric aberration in the history of intelligence models' (p. 60). However, one result of Guilford's model was that theorists and test developers began considering the role of creativity and innovation in intelligence. Early tests of intelligence largely measured 'convergent' thinking (i.e. the use of logical steps to reach one correct answer to a structured question) rather than 'divergent' thinking (i.e. the capacity to creatively generate multiple solutions to a stimulus).

structure-of-intellect (SOI) model
J P Guilford's multifaceted model of intelligence consisting of 150 intellectual abilities arranged along three dimensions of operations, content and product

Figure 7.8 J P Guilford (1897–1987)

Figure 7.9 Guilford's model of intelligence



Vernon's hierarchical view of intelligence

English psychologist Philip Vernon (1905–1987) incorporated Spearman's 'g' and Thurstone's primary mental abilities into an expanded, **hierarchical model of intelligence** (Vernon, 1965). Vernon did little empirical work himself; rather, his contribution was to summarise the large volume of research on intelligence in the years prior to 1950 and propose a plausible, unified structure for intelligence

(Aiken, 1996). At the base of Vernon's hierarchical model were multiple, narrowly defined, specific abilities, which were assessed directly by individual tests of ability. Groups of these specific abilities clustered together (i.e. were strongly correlated) to form 'minor group factors' at the second level of the hierarchy. These minor group factors were similar to Thurstone's primary mental abilities, although Vernon did not indicate how many should be included at this level of the model. In turn, groups of minor group factors clustered together to form two 'major group factors' at the third level of the hierarchy. Vernon labelled these two major group factors as 'v:ed', or the verbal-educational factor (e.g. encompassing verbal comprehension and numerical ability), and 'k:m', or the spatial-motor factor (e.g. encompassing perceptual speed and spatial visualisation). The two major group factors were considered moderately correlated and came together to reflect global ability, or Spearman's 'g' at the top of the hierarchy. Vernon's model is applauded as being a comprehensive hierarchical model of intelligence, as it includes abilities that range from the very narrow, which account for performance in very specific ability areas, to the very broad, including the global summary of intelligence of 'g' (Gustafsson, 1989). The advantage of using such a hierarchy is that it is possible to assess at the level required; that is, it is possible to test to obtain an IQ score or to test specific abilities or clusters of specific abilities. The emergence of hierarchical perspectives on intelligence such as this one defused much of the controversy surrounding the nature and definition of intelligence.

Figure 7.10 Philip Vernon (1905–1987)

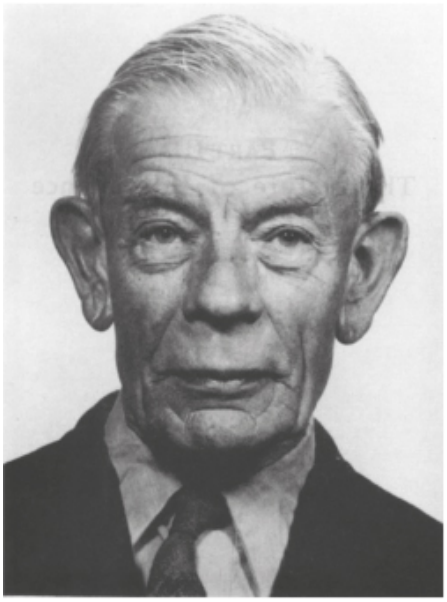
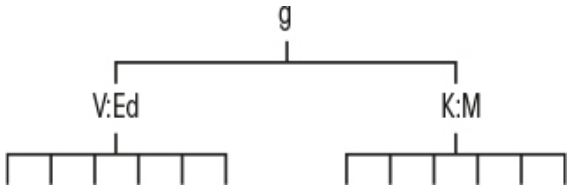


Figure 7.11 Vernon's hierarchical model of intelligence



hierarchical models of intelligence
psychometric models that represent intelligence hierarchically, with many narrow abilities (first-order factors) at the first level, which define a smaller number of broader abilities (second-order factors), and the broader abilities are then represented by a general or 'g' factor at the top

Cattell's two-factor theory of intelligence

Raymond Cattell (1905–1998) was born in England, but is known as an English-American psychologist because he spent much of his adult life in the USA. He argued that general intelligence consisted of two main components, rather than one 'g' factor (as suggested by Spearman) or multiple factors (as suggested by Thurstone). Cattell, with his student John Horn (1928–2006), proposed a two-factor theory of 'fluid' and 'crystallised' intelligences, or the **two-factor (Gf-Gc) theory of intelligence**. Cattell was a factor analyst. He confirmed Thurstone's multiple intelligence factors, but when he analysed the correlations among these primary mental abilities, he produced two second-order factors (rather than a single 'g' factor), which he labelled fluid and crystallised intelligence. Cattell did not dispute the existence of 'g'; rather, he argued that it could be decomposed into these two correlated, but distinct, dimensions. **Fluid intelligence** is the non-verbal, relatively culture-free, basic mental capacity of the individual, which underpins abstract problem solving and reasoning, independent of acquired knowledge. Fluid intelligence is primarily dependent on genetic endowment. In testing situations, for example, the individual uses fluid intelligence to solve unfamiliar tasks, such as identifying a common theme in complex shapes, or solving a visual puzzle. **Crystallised intelligence**, on the other hand, is more dependent on learning. It is the culture-specific fund of knowledge, skills and information that is accumulated through life's experiences and education. This can be assessed by testing the individual's vocabulary, fund of acquired knowledge and social acumen. Successful crystallised intelligence (i.e. the amount of cultural knowledge acquired) is dependent on one's level of fluid intelligence (i.e. it is dependent on one's level of 'raw' intellectual brainpower).

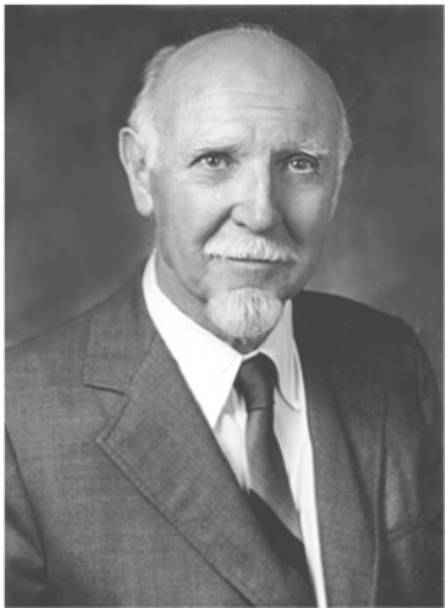
two-factor (Gf-Gc) theory of intelligence
Cattell's original theory, which decomposed 'g' into two component parts: fluid

and crystallised intelligence (Gf and Gc)

crystallised intelligence (Gc)
the accumulated knowledge and skills resulting from educational and life experiences

fluid intelligence (Gf)
the more pure, inherited aspects of intelligence used to solve novel problems and deal with new situations

Figure 7.12 Raymond Cattell (1905–1998)



Cattell's theory can be thought of as a hierarchical theory, as both fluid and crystallised intelligence are made up of lower-level, more specific dimensions of intelligence. Both the Wechsler scales and the Stanford-Binet test are considered to measure both types of intelligence, with fluid intelligence reflected in some performance-based tests (e.g. copying novel designs using three-dimensional blocks and solving puzzles based on complex geometric shapes) and crystallised intelligence dependent on verbal-based tests (e.g. tests of general knowledge and vocabulary). Neither Wechsler nor Terman constructed their tests specifically to assess Cattell's fluid and crystallised abilities (both original tests produced a global 'g' score); rather, the evolution of both tests paralleled the development of the understanding of intelligence.

Cattell's theory stimulated a search for 'culture-free' tests, which might assess the basic, fluid intelligence, uncontaminated by crystallised intelligence. But this

proved an impossible task (Cole, 1999). Sitting at a desk facing an examiner, making eye contact and using a pencil and paper, even to solve non-verbal problems, are all culturally determined behaviours. The best that could be hoped for were tests that were more 'culture fair' than the traditional measures of intelligence. **Culture fair tests** of intelligence typically incorporate few verbal instructions and tap intelligence using images and visuo-spatial puzzles. Cattell himself devised a 'culture fair' test, which he professed assessed fluid intelligence, and which would be useful with individuals, for example, who were deprived of formal education experiences; however, the results of this test, and other culture fair tests, have been shown to still be influenced by cultural experiences (Smith, Hays & Solway, 1977). The culture fair tests most widely used today are the Progressive Matrices series devised by John Raven (see the box 'Raven's Progressive Matrices' in Chapter 2). The series includes the Coloured Progressive Matrices (for use with children and test takers in the lower end of the ability range), the Standard Progressive Matrices (for use with those in the middle-range of ability) and the Advanced Progressive Matrices (for use with individuals with superior ability). All versions are based on analogous problem-solving tasks. The test taker is presented with a complex pattern, which is incomplete, and asked to 'solve the problem' by working out which of a number of provided options (or pattern pieces) completes the complex pattern (Raven, 1939). More details and applications for the Progressive Matrices test are provided in Chapter 10.

culture fair test
a test devised to measure intelligence while relying as little as possible on culture-specific knowledge (e.g. language); tests are devised to be suitable across different peoples, with the goal to measure fluid rather than crystallised intelligence

Cattell, Horn and Carroll extend the 'Gf-Gc' model of intelligence

Over the second half of the twentieth century, Cattell, Horn and others extended the *'Gf-Gc'* model of intelligence to include other second-order factors along with *'Gf'* and *'Gc'*. These additional second-order factors paralleled many of Thurstone's primary mental abilities. They were identified using factor analytic studies, which on the one hand confirmed Cattell's fluid and crystallised intelligences, and on the other identified other factors that needed to be considered. The full list of the Cattell-Horn broad factors is as follows:

'Gf'—fluid intelligence
'Gq'—quantitative knowledge

'Gc'—crystallised intelligence
'Grw'—reading and writing ability
'Gsm'—short-term memory
'Gv'—visual processing
'Ga'—auditory processing
'Glr'—long-term retrieval
'Gs'—processing speed
'CDS'—correct decision speed.

The list is ranked in order of their strength of association with general ability, or 'g'; that is, fluid intelligence has a stronger correlation with 'g' than quantitative knowledge, which in turn has a stronger correlation than crystallised intelligence, and so on (Flanagan et al., 2002).

This expanded theory became known as the Cattell-Horn 'Gf-Gc' theory. In 1993, the 'Gf-Gc' theory was extended in a slightly different manner, when the US psychologist John Carroll (1916–2003) published *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Carroll proposed that intelligence should be viewed as comprising three levels, or *strata* (plural of stratum), which could be differentiated in terms of the breadth or specificity of the ability being assessed. Carroll's 'three stratum theory' of intelligence was the result of a herculean exercise, during which he re-analysed and summarised more than 460 datasets that had been used since Spearman's day. Such was the impact of this piece of work that Horn (1998) called it 'a tour de force' (p. 58), and compared it to the development of the periodic table in chemistry. It was also praised as being 'virtually the grand finale of the era of psychometric description and taxonomy of human cognitive abilities', something that was unlikely to 'ever be attempted again by anyone, [or] could be much improved on' (Jensen, 2004; p. 5). General ability (consistent with Spearman's 'g') sits at the top of Carroll's three stratum, **hierarchical model of intelligence** (Stratum III). This is the broadest conceptualisation of intelligence, and reflects global IQ scores obtained from intelligence tests such as the Stanford-Binet test and Wechsler scales. Stratum II sits below Stratum III, and consists of broad intellectual abilities, including Cattell's 'Gf-Gc' main factors, many of Horn's broad factors, and others reminiscent again of Thurstone's primary mental abilities. Cattell's original eight Stratum II factors were as follows:

'Gf'—fluid intelligence
'Gc'—crystallised intelligence
'Gy'—general memory and learning
'Gv'—visual perception
'Gu'—auditory perception
'Gr'—retrieval ability
'Gs'—cognitive speed

'Gt'—decision reaction time.

Again, those higher on the list are considered more highly correlated with general intelligence (Flanagan et al., 2002). The index scores obtained in the modern versions of the Stanford-Binet test and Wechsler scales can be considered to represent broad Stratum II abilities. At the base of Carroll's hierarchy (Stratum 1) is a large number of narrow abilities. For example, Stratum I abilities reflecting crystallised intelligence ('Gc') include vocabulary knowledge, listening ability, general knowledge, information about culture and communication ability. Stratum I abilities reflecting visual perception ('Gv') include the capacity to form and manipulate mental images, understanding rotated pictures, recognising patterns, length estimation and visual memory. The individual subtests included in the Stanford-Binet test and Wechsler scales were devised to assess narrow abilities, and can be considered to be representative of Stratum I abilities. Clearly, there is overlap between the Cattell-Horn 'Gf-Gc' model and Carroll's three stratum theory. Both include multiple, broad factors (Stratum II), which subsume specific, narrow abilities (Stratum I). Many of the Stratum II factors are very similar in both models, although some at this level do differ (e.g. the Cattell-Horn model includes a reading/writing factor 'Grw' at Stratum II, whereas Carroll has this as a specific ability at Stratum I and feeding into crystallised intelligence, 'Gc'). Finally, Carroll's model specifically includes the broad, general ability factor of 'g' (Stratum III), whereas Cattell and Horn downplayed this level (Alfonso, Flanagan & Radwan, 2005).

The Cattell-Horn-Carroll (CHC) model of intelligence

During the 1990s, US psychologist Kevin McGrew proposed an integration of the 'Gf-Gc' and three stratum theories (McGrew, 1997). This integrated theory is known as the Cattell-Horn-Carroll theory of cognitive abilities, or the **CHC theory of intelligence** (Alfonso, Flanagan & Radwan, 2005). McGrew's 2012 version of the CHC theory (Schneider & McGrew, 2012) retains general ability, 'g', at the top at Stratum III, but has an expanded Stratum II that includes sixteen broad factors, compared with the ten in the 'Gf-Gc' theory and eight in Carroll's three-stratum model. McGrew's additional Stratum II abilities are as follows:

CHC theory of intelligence
the Cattell-Horn-Carroll model; a merging of Cattell and Horn's Gf-Gc theory and Carroll's three stratum theory, which proposes three levels or strata of abilities: narrow, broad and general (or 'g')

'Gkn'—domain specific knowledge
'Gps'—psychomotor speed
'Go'—olfactory abilities
'Gh'—tactile abilities
'Gk'—kinaesthetic abilities
'Gp'—psychomotor abilities.

McGrew also arranged his broad Stratum II factors into clusters of associated abilities of domain-independent general capacities (e.g. fluid reasoning, memory and speed of processing), acquired knowledge systems (e.g. general knowledge, reading and writing, and quantitative knowledge) and sensory/motor-linked abilities (e.g. visual processing, auditory processing, olfactory, tactile and motor abilities). Whether all sixteen broad factors will be retained in the CHC model, or whether more will be added, will depend on theory development and the results from empirical research. Schneider and McGrew (2012), when commenting on the CHC model, stated: 'The end goal, however, has always been for CHC theory to undergo continual upgrades so it would evolve towards an ever-more accurate summary of human cognitive diversity' (p. 137). At the base of this latest version of the CHC are seventy-four specific, narrow abilities at Stratum I. Quantitative knowledge ('Gq'), for example, subsumes the two specific abilities of mathematical knowledge and mathematical achievement, while long-term storage and retrieval ('Glr') subsumes twelve specific abilities (e.g. word fluency, associative memory and figural retrieval fluency).

The CHC model and modern tests of intelligence

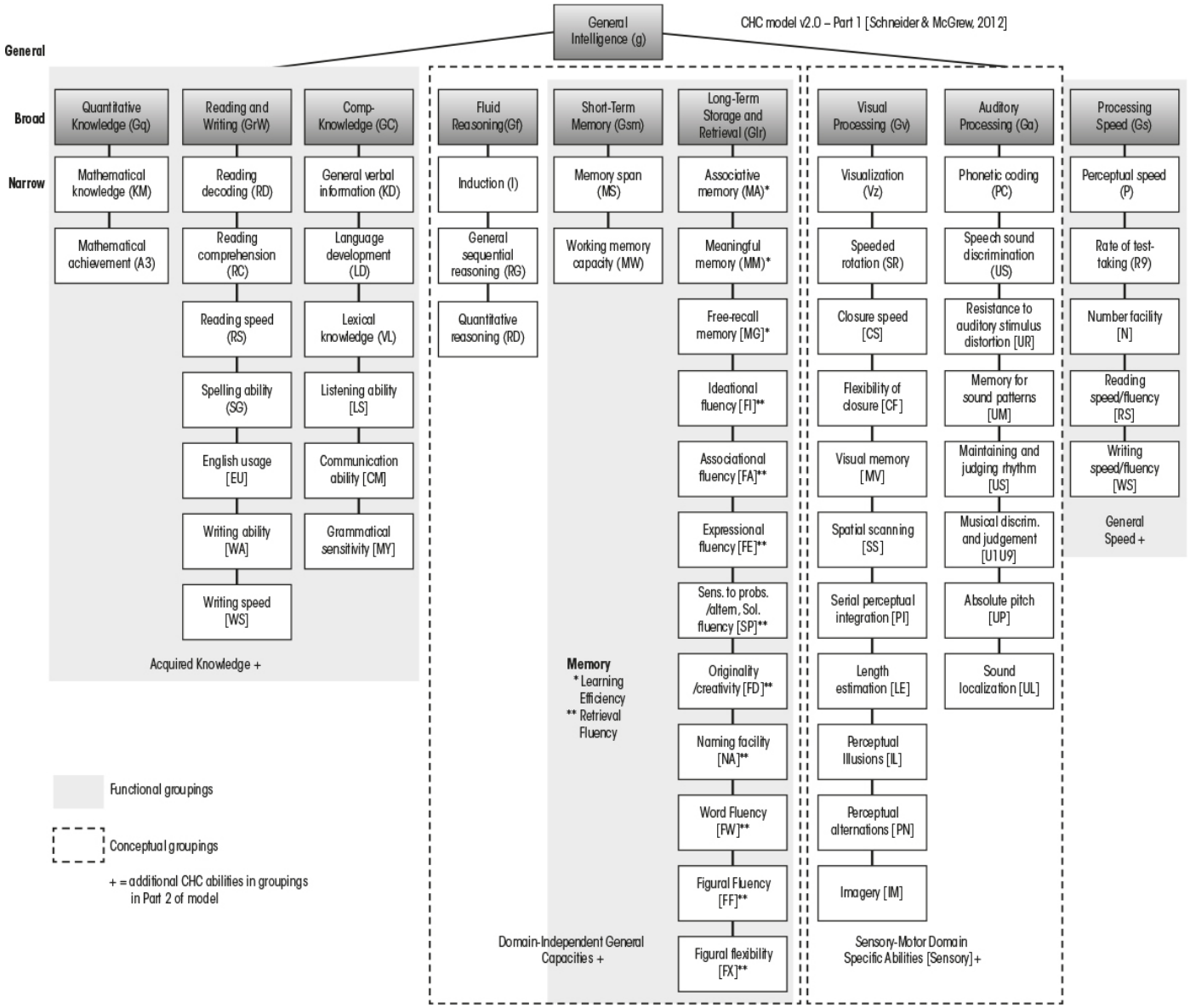
Most tests of intelligence published before the year 2000 assessed few (typically only two or three) broad Stratum II abilities. The Wechsler Adult Intelligence Scale–Revised, published in 1981, had adequate measures to assess crystallised intelligence ('Gc'; using four subscales) and visual processing ('Gv'; also using four subscales), but, for example, had insufficient measures of fluid intelligence ('Gf') and no way to assess long-term retrieval ('Glr'). Typically, pre-2000 tests did not adequately assess the Stratum II abilities of 'Gf', 'Gsm', 'Glr', 'Ga' and 'Gs', as these early tests of intelligence were based on historical precedents and/or idiosyncratic conceptions of the structure of intelligence (Alfonso, Flanagan & Radwan, 2005). In recent years, the Cattell-Horn-Carroll (CHC) theory, which has generated considerable support for being a comprehensive account of human intelligence, has been the main influence on the development and revision of the major tests of intelligence (Keith & Reynolds, 2010). The CHC theory, either implicitly or explicitly, has been used as the foundation for nearly all contemporary, comprehensive and individually administered tests of intelligence published since 2000, including the Stanford-Binet (Fifth Edition), published in

2003, the Wechsler Adult Intelligence Scale (Fourth Edition), published in 2008, and the Wechsler Intelligence Scale for Children (Fifth Edition), published in 2014. Other important tests now based substantially on the CHC model are the Differential Abilities Scales (Second Edition), the Kaufman Assessment Battery for Children (Second Edition) and the Woodcock-Johnson Battery (Fourth Edition). Most of these contemporary tests of intelligence are structured to assess four or five broad, Stratum II cognitive abilities, as well as generate both an overall, global score ('g' representing Stratum III) and scores for individual subtests (Stratum I narrow abilities). See Table 7.1 for the broad abilities assessed in the Stanford-Binet (Fifth Edition) and the Wechsler Adult Intelligence Scale (Fourth Edition).

Both the Stanford-Binet (Fifth Edition) and the Wechsler Adult Intelligence Scale (Fourth Edition) now include improved measures of fluid intelligence ('Gf') and short-term memory ('Gsm'), as do many other of the recent tests; however, no tests, including the Stanford-Binet test and the Wechsler scales, assess the full range of broad (Stratum II) or narrow (Stratum I) abilities. This has led some to propose an 'across-battery' approach to the assessment of a broader range of Stratum II abilities (McGrew & Flanagan, 1998). This approach suggests that practitioners might augment an individual test (e.g. the Stanford-Binet test or the Wechsler scales) by selecting measures from other individual tests, cognitive batteries or neuropsychological tools to allow a more comprehensive understanding of a person's strengths and weaknesses. The selection of the additional tests should reflect CHC theory, be guided by history taking of the client, and be used to answer specific hypotheses in the assessment (Flanagan, Alfonso & Ortiz, 2012).

The Psychology Board of Australia (2016b) has mandated that psychologists, for general registration, must 'demonstrate competence in the administration, scoring and interpretation' (p. 5) of the WAIS–IV and the WISC–V intelligence tests, and 'demonstrate general familiarity with the use and purpose' (p. 5) of other tests of intelligence (i.e. Wechsler Preschool and Primary Scale of Intelligence, Stanford-Binet Intelligence Scales, Kaufman Adolescent and Adult Intelligence Test, Wechsler Abbreviated Scale of Intelligence, and Woodcock-Johnson Test of Cognitive Abilities). Knowledge of these tests is assessed in the national psychology examination conducted by the Psychology Board of Australia. Where psychologists seek endorsement for 'speciality' areas of practice (e.g. clinical or organisational), more in-depth knowledge regarding test selection, administration, scoring and interpretation is expected, as these psychologists seek to claim advanced knowledge and proficiencies.

Figure 7.13 The CHC model of intelligence


CHC model v2.0 – Part 1 [Schneider & McGrew, 2012]
General
General Intelligence (g)
Broad
Narrow
Quantitative Knowledge (Gq)
Reading and Writing (GrW)
Comp-Knowledge (GC)
Fluid Reasoning (Gf)
Short-Term Memory (Gsm)
Long-Term Storage and Retrieval (Glr)
Visual Processing (Gv)
Auditory Processing (Ga)
Processing Speed (Gs)
Mathematical knowledge (KM)
Mathematical achievement (A3)
Reading decoding (RD)
Reading comprehension (RC)
Reading speed (RS)
Spelling ability (SG)
English usage [EU]
Writing ability [WA]
Writing speed [WS]
General verbal information (KD)
Language development (LD)
Lexical knowledge (VL)
Listening ability [LS]
Communication ability [CM]
Grammatical sensitivity [MY]
Induction (I)
General sequential reasoning (RG)
Quantitative reasoning (RD)
Memory span (MS)
Working memory capacity (MW)
Associative memory (MA)*
Meaningful memory (MM)*
Free-recall memory [MG]*
Ideational fluency [FI]**
Associational fluency [FA]**
Expressional fluency [FE]**
Sens. to probs. /altern, Sol. fluency [SP]**
Originality /creativity [FD]**
Naming facility [NA]**
Word Fluency [FW]**
Figural Fluency [FF]**
Figural flexibility [FX]**
Visualization (Vz)
Speeded rotation (SR)
Closure speed [CS]
Flexibility of closure [CF]
Visual memory [MV]
Spatial scanning [SS]
Serial perceptual integration [PI]
Length estimation [LE]
Perceptual Illusions [IL]
Perceptual alternations [PN]
Imagery [IM]
Phonetic coding (PC)
Speech sound discrimination (US)
Resistance to auditory stimulus distortion [UR]
Memory for sound patterns [UM]
Maintaining and judging rhythm [US]
Musical discrim. and judgement [U1U9]
Absolute pitch [UP]
Sound localization [UL]
Perceptual speed (P)
Rate of test-taking (R9)
Number facility [N]
Reading speed/fluency [RS]
Writing speed/fluency [WS]
General Speed +
Acquired Knowledge +
Memory
* Learning Efficiency
** Retrieval Fluency
Domain-Independent General Capacities +
Sensory-Motor Domain Specific Abilities [Sensory] +
Functional groupings
Conceptual groupings
+ = additional CHC abilities in groupings in Part 2 of model

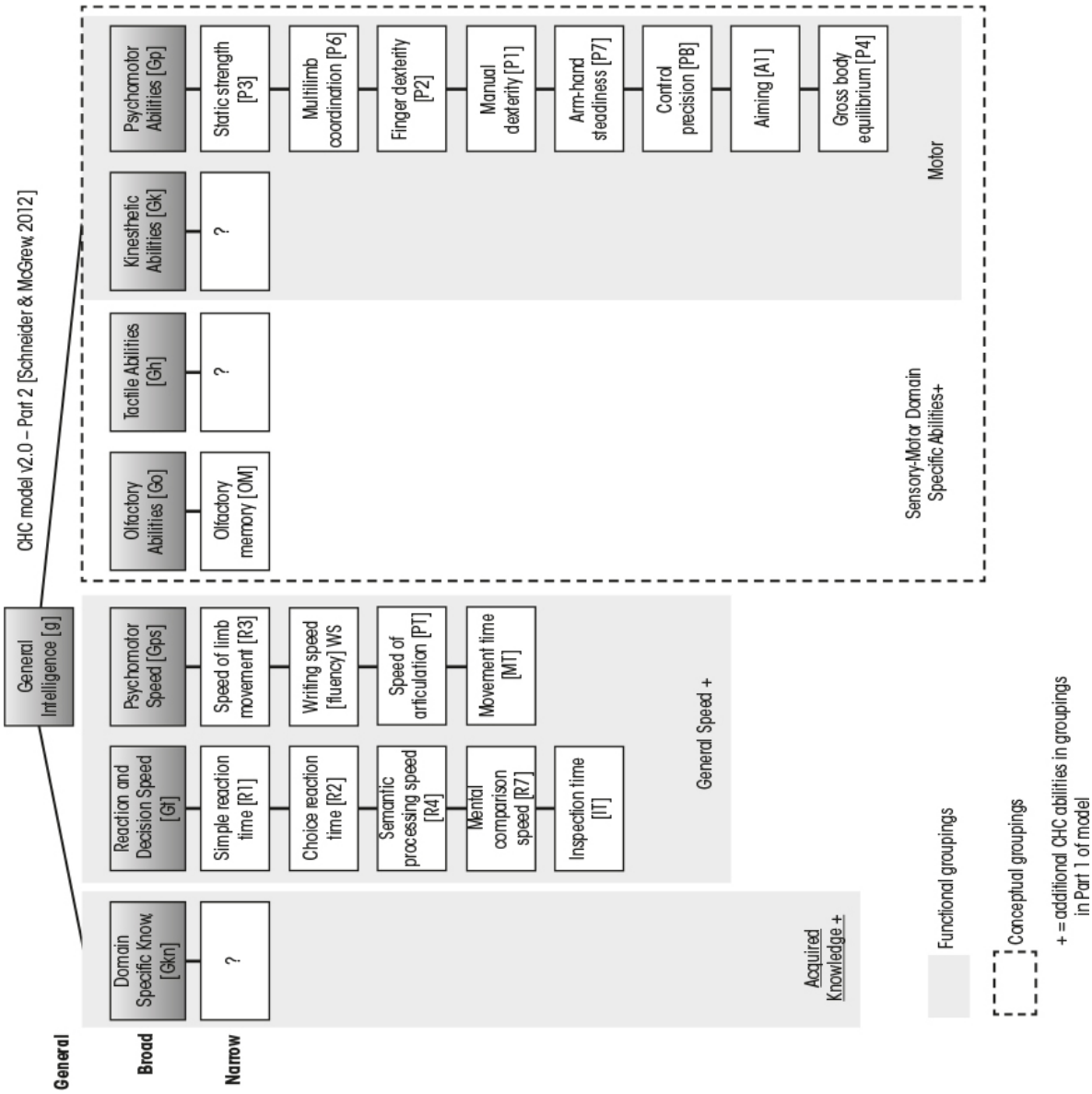


Table 7.1: Broad (Stratum II) abilities assessed by the Stanford-Binet and Wechsler Adult Intelligence Scale

Stanford-Binet test (Fifth Edition)
Fluid reasoning (reflecting fluid intelligence, 'Gf') Knowledge (crystallised intelligence, 'Gc') Quantitative reasoning (quantitative knowledge, 'Gq') Visual-spatial processing (visual processing, 'Gv') Working memory (short-term memory, 'Gsm')
Wechsler Adult Intelligence Scale (Fourth Edition)

Stanford-Binet test (Fifth Edition)
Verbal comprehension (representing crystallised knowledge, 'Gc') Perceptual reasoning (representing visual processing, 'Gv', and fluid intelligence, 'Gf') Working memory (representing short-term memory, 'Gsm') Processing speed (representing cognitive speed, 'Gs')

Roid (2003); Benson, Hulac and Kranzler (2010)

A developmental conception of intelligence

The early models of intelligence (e.g. Spearman's one-factor model of 'g' and Thurstone's multifactor model of primary mental abilities) and the more recent models (i.e. the Cattell-Horn 'Gf-Gc' model, Carroll's three stratum model and the merged Cattell-Horn-Carroll CHC model) were based largely on factor analytic procedures. These factor analytic, or hierarchical, models are known as '**psychometric theories**', as they seek to explain the structure of intelligence by understanding the relationships among individual tests. Intelligence has been considered from other, non-psychometric, perspectives. Jean Piaget (1896–1980), a Swiss psychologist, proposed an important 'developmental theory' of cognitive abilities. Piaget proposed that intelligence in children develops as a result of the interaction between their biological endowment and their experiences in the environment. Children come into the world with a few simple schemata (plural of schema), or cognitive structures, that are required, for example, for grasping and sucking. As children interact with the environment, these schemata are continuously organised and reorganised to incorporate more sophisticated understandings of the world around them. Thus, Piaget's view of intelligence was synonymous with 'adaptation' to the environment, resulting from schemata development in the face of environmental experiences.

psychometric theory
a theory concerned with the measurement of psychological constructs (like intelligence); the two main theories underpinning test development are classical test theory and item response theory; psychometric techniques typically include factor analysis and its variants

For Piaget, these schemata are developed across four main stages during the years from birth to early adolescence. These four stages are: 'sensorimotor' (birth to 2 years), during which the child integrates sensory input and motor abilities; 'preoperational' (2–6 years), which is characterised by egocentrism, magical

(illogical) thinking and the development of language; 'concrete operational' (7–12 years), when logical thinking emerges, but is still concrete, and egocentrism declines; and 'formal operations' (from 12 years), when abstract and logical thought develops, and the emerging adult can consider information not yet personally experienced. Piaget showed that children's thinking was qualitatively different from that of adults, and while his theory has been influential in developmental and educational psychology, little of it has been incorporated into the field of intelligence testing, even though the theory has important implications for the assessment of children.

An information-processing view of intelligence

Intelligence has also been conceptualised in terms of how material is processed by the brain. The 'planning, attention-arousal, simultaneous and successive' (PASS) cognitive processing theory (Naglieri, Das & Goldstein, 2012) proposes that there are four main cognitive processing units, which have biological counterparts in cortical structures. This theory reflects much of the work undertaken by Russian psychologist Alexander Luria (1902–1977), who worked with, and studied, brain-injured soldiers returning from the Second World War (Luria, 1973). Luria suggested that the brain comprised functionally independent areas, but that in order to function, these areas needed input from other areas, and needed to interact with one another. The four PASS cognitive processing units are:

1. planning, which reflects important aspects of executive functioning, involves abilities associated with goal-setting, planning and monitoring behaviours associated with meeting those goals, and implementing self-regulatory and adjustment strategies to keep on track
2. attention-arousal, which reflects abilities associated with maintaining sustained attention and focus, and resisting distraction
3. simultaneous processing, which is essential for integrating different stimuli into a coherent whole, as is required in speech comprehension, when individual words, sentences, inflections and non-verbal cues are integrated to give meaning to the interaction, or when multiple visuo-spatial components are integrated to allow appreciation of a piece of art or a natural vista
4. successive processing, which is required when dealing with information that is sequential or serially ordered, such as when spelling a word, working through the steps to solve an arithmetic problem, or arranging files in alphabetical order.

Naglieri and Das (1997) devised a cognitive assessment system to assess these four main components of cognitive functioning in children and adolescents (age range 5–17 years). The test produces scores for the individual subtests (twelve in the standard battery), PASS scale scores (i.e. for planning, attention-arousal, simultaneous and successive processing) and a global 'full scale' score. This test has not challenged the standard tests of intelligence, such as the Stanford-Binet test and Wechsler scales, but does provide a theoretically based assessment of cognitive development in children. The test is used in educational settings with children (e.g. for the assessment of learning difficulties and reading problems) and with adults who have suffered cognitive impairment.

Gardner and multiple intelligences

US psychologist Howard Gardner (1943–) proposed a theory of '**multiple intelligences**', which, he argued, was also based on functional areas of the brain, but which find expression within a cultural context (Gardner, 2006). These two notions are reflected in Gardner's (1999) definition of intelligence as 'a biopsychological potential to process information that can be activated in a cultural setting to solve problems or create products that are of value in a culture' (pp. 33–4). Gardner suggests that in addition to the three types of intelligence typically assessed by standard intelligence tests (i.e. linguistic, logical-mathematical and visuo-spatial intelligence), there are five other types of intelligence: bodily-kinaesthetic, inter-personal, intra-personal, musical and naturalistic (Chen & Gardner, 2012). Bodily-kinaesthetic intelligence refers to the abilities used by sports people, dancers and surgeons, who need to master physical expression and control; whereas, inter- and intra-personal intelligences reflect abilities for understanding of others and personal insight, respectively. Musical intelligence reflects abilities with rhythm, harmony, pitch and so on, which are exploited by musicians; whereas, naturalistic intelligence refers to abilities related to understanding and managing the natural world, which are needed by farmers, foresters and biologists.

multiple intelligences
a theory usually associated with Howard Gardner, who proposed that intelligence comprises multiple, discrete modalities that are not aggregated to 'g'

Consistent with Gardner's view of intelligence, these eight intelligences find different expressions in different cultures. Gardner rejects the notion of an over-arching measure of 'g' and argues that individuals can be differentiated by different profiles on the eight individual intelligences. The study of 'savants'—

individuals who experience mental deficiencies but who excel in one highly specific skill or ability—lends support to this multiple intelligences theory. One famous savant, Kim Peek, portrayed in the movie *Rain Man*, was born with severe brain abnormalities and had trouble with many facets of everyday life. However, he had the ability to read and memorise extraordinary amounts of information, including the 12,000 books and manuscripts he had read. His parents had to stop taking him to live performances, as he would stand up and correct the actors and musicians when they made an error in the dialogue or musical score. Gardner's theory has become very popular with educationalists, as every child can be identified as having strengths in some area or areas, and it is useful for identifying areas for remediation. However, the theory has been criticised on a number of grounds (e.g. Waterhouse, 2006), including that it has not been able to account for the intercorrelations among the different intelligences, that some of the intelligences (e.g. musical and bodily-kinaesthetic) might more reasonably be considered as talents rather than intelligences, and that some of the intelligences can be considered more personality-based than reflecting intelligence (e.g. inter-personal and intra-personal). A number of tests have been devised to assess Gardner's different intelligences in children, adolescents and adults. For example, the Spectrum battery (Chen, 2004) was devised to assess children's abilities in seven areas—language, mathematics, music, art, social understanding, science and movement—and other individual tests available in the literature have also been proposed (e.g. Visser, Ashton & Vernon, 2006).

Sternberg's triarchic theory of intelligence

Robert Sternberg (1949–), also a US psychologist, proposed a third information-processing theory: the **triarchic theory of intelligence** (Sternberg, 1985b). 'The essence of intelligence', according to Sternberg, 'is that it provides a means to govern ourselves so that our thoughts and actions are organised, coherent, and responsive to both our internally driven needs and to the needs of the environment' (1986, p. 141). In this theory, intelligence is reflected in three main cognitive processes:

triarchic theory of intelligence
a theory proposed by Robert Sternberg in which intelligence comprises three components: analytical abilities ('componential'), creative abilities ('experiential') and practical abilities ('contextual'); it suggests that individuals high on the three components should experience real-life success

1. Componential processes, also known as analytical processes, reflect intellectual abilities traditionally considered to be related to intelligence, including higher-order, executive functions (e.g. planning, monitoring, evaluating and problem solving), learning processes (i.e. processes associated with knowledge acquisition) and abilities needed to perform tasks (e.g. the ability to store information, see relationships and use inductive reasoning).
2. Experiential processes, or processes associated with creative intelligence, reflect abilities associated with dealing with novel and unusual situations, such as generating new ideas, devising new ways to carry out a task, and coming up with innovative ways to solve a problem.
3. Contextual processes, or practical intelligence, refer to abilities associated with adapting to one's environment (i.e. changing oneself), shaping one's environment (i.e. changing the world around you) and selecting an environment' (e.g. relocating to a new, more satisfying environment) (Sternberg, 2012).

All three strategies reflect the individual's capacity to develop an agreeable 'fit' with the environment. In this context, a person with a good environment fit can be considered to be more 'streetwise'. Like Gardner's theory, Sternberg's theory is intuitively sound, and resonates with the way that lay people view intelligence; however, Sternberg's theory has not greatly influenced the development of intelligence tests or the way intelligence is measured generally. The theory's contribution (mirroring Gardner's contribution) is to highlight that there are other forms of intellectual worth apart from 'academic intelligence', as measured by standard intelligence tests. Sternberg argues that his three broad processes (i.e. analytical, creative and practical) are independent of 'g', but research has not supported this contention. The theory also has been criticised for its overlap with other individual differences, such as social acumen, motivation and personality (Gottfredson, 2003).

So, 'what is intelligence?'

There is clearly no final agreement among researchers and thinkers as to what constitutes intelligence. Many of these differences of opinion on intelligence reflect, in part, the different research traditions taken by researchers (i.e. psychometric, information-processing or cognitive-developmental approach). Some researchers and practitioners tend even to shy away from the use of the term 'intelligence' and use, instead, terms such as 'cognitive abilities' and 'aptitudes'. Wasserman (2012), who listed thirty-one definitions of intelligence

dating from 1855, bemoaned the fact that psychologists, after more than a hundred years of studying intelligence, were still unable to agree to a definition; asking: 'How much longer must we wait?'

However, the picture may not be as dire as suggested by Wasserman. In 1986, when twenty-five eminent scholars were asked about their conception of **intelligence**, there was not 100 per cent agreement, but there tended to be considerable overlap in their views of intelligence, which reflected domains such as higher cognitive functions (e.g. reasoning and problem solving), executive functions (e.g. planning and monitoring functions), basic cognitive functions (e.g. perception and attention) and activities reflecting success in one's culture (e.g. career success in Western cultures; Sternberg & Berg, 1986). In 1994, Linda Gottfredson, a professor of educational psychology at the University of Delaware, and a group of fifty-one other university professors, famously published a statement in the *Wall Street Journal* summarising what was known about intelligence, and concluded that there was considerable consensus in the academic community on how to define intelligence; and when existing tests are used appropriately, they all measure more or less the same thing. These fifty-two experts stated:

> Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings—'catching on', 'making sense' of things, or 'figuring out' what to do. (Gottfredson, 1997, p. 13)

intelligence
cognitive abilities such as problem solving and learning, although some definitions include other aspects of the individual, such as personality and creativity

Aptitude versus achievement tests

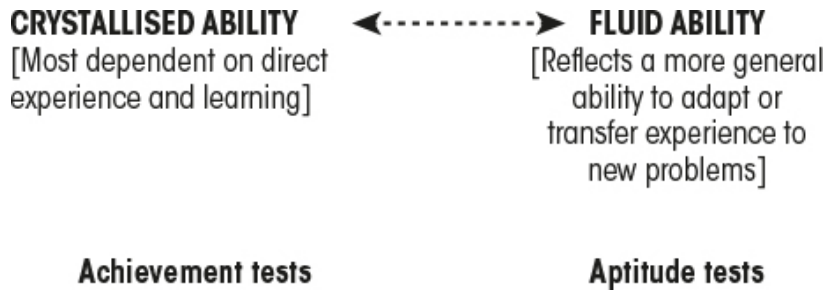
Intelligence testing seeks to identify cognitive differences among individuals. Testing today is less likely to focus on assessing global abilities (i.e. Stratum III level in the CHC model) and more likely to be structured to assess specific cognitive abilities found at Stratum II. Identifying strengths and weaknesses at this level can be more useful, for example, to aid in a diagnosis, to assist in formulating a rehabilitation plan or to select someone to work in a particular area. These tests can be considered as measures of 'cognitive ability'. Tests at this

level also can be thought of as 'aptitude-based' or 'achievement-based', although the distinction between these two types of measures is blurred, and there can be much overlap between them. **Aptitude tests** are measures of an individual's potential, and the results on these types of tests should be correlated with a later performance. For example, a test designed to assess a student's potential to study for a law degree might include tasks such as problem-solving, reasoning and abstract abilities; tasks that are more likely to assess fluid rather than crystallised abilities (see Figure 7.14). **Achievement tests**, on the other hand, seek to tap an individual's understanding and knowledge that are dependent on past experiences and specific learning. University and school examinations are achievement tests as they seek to identify what has been learned in the classroom, as are tests of vocabulary and general knowledge, which are included in current intelligence tests (e.g. the WAIS–IV). Achievement tests assess crystallised more so than fluid abilities, although, as argued by Cattell (1987), the capacity to acquire knowledge is largely dependent on one's fluid ability. Both aptitude and achievement tests can be broad or narrow in what they are assessing. The WAIS–IV vocabulary subtest assesses language competence (e.g. word knowledge and comprehension) acquired over a lifetime, whereas a specific mathematics achievement test might assess what was learned in class during one school term.

aptitude test
a test used to assess future learning potential

achievement test
a test used to assess past learning

Figure 7.14 The fluid–crystallised dimension



Group (rather than individual) testing

The tests referred to so far have been mostly individual tests of intelligence; that is, one administrator assessing one test taker. However, for many reasons, chief among them being efficiency and cost-saving, intelligence (aptitude and achievement domains) can be assessed in groups with specifically devised group tests; that is, one administrator assessing multiple test takers. In fact, the number of people who are assessed in group situations is far more than the number assessed individually. The Australian Armed Forces, for example, use group tests of ability to screen all applicants as to their suitability for service, and most medium-sized and large Australian companies include tests as part of their job selection procedure. These two sectors alone account for hundreds of thousands of Australian adults being tested each year. Apart from job selection, group-based ability tests are widely used in school settings (e.g. screening to identify children in need of special programs), when selecting candidates for entry to professional training programs (e.g. the Graduate Medical School Admissions Test, or GAMSAT, which is used as part of the selection procedure for entry to medical schools in Australia), and in much social science research.

Many group tests are still constructed as paper-and-pencil tests, but more and more they are being adapted for computerised administration, where, in some cases, the test taker can log on to the test and sit it without having to leave their home. Some group tests are structured to assess a broad range of cognitive abilities, while others are developed to assess single, specific aptitudes. An example of a group test devised to assess a broad range of abilities is the Otis-Lennon School Ability Test (OLSAT; Otis & Lennon, 2003—first published in 1979 and now into its eighth edition). This battery contains twenty-one individual tests, which are grouped into five broad abilities of verbal comprehension, verbal reasoning, pictorial reasoning, figural reasoning and quantitative reasoning. The test then generates scores for verbal ability and non-verbal ability (based on the five broad abilities), and a total score, which equates to general ability (i.e. Stratum III 'g'; Harcourt Educational Measurement, 2003). Group-based tests such as the OLSAT have very similar content to individually based tests, although, for obvious reasons, they cannot include tests of manipulation (e.g. arranging blocks) or working memory (e.g. showing images that have to be copied from memory). Group tests are also devised to assess very specific abilities. There are too many of these to list, but they assess a diverse range of capabilities, including mechanical reasoning and computer aptitude (e.g. for selecting trade apprentices), clerical ability (e.g. for selecting office workers), verbal reasoning (e.g. for selecting sales personnel) and spatial ability (e.g. selecting for entry into architectural programs). Most group-based tests, whether comprehensive or narrow, have an aptitude focus, predicting, for example, how well the individual will do at school, at work or in the military.

Group differences in intelligence

Much of the controversy over intelligence testing revolves around the real and perceived differences in ability that emerge across different groups of people, as well as the explanations that are sometimes given for these differences. The history of intelligence testing includes many examples where the differences found among individuals were extrapolated to highlight differences among groups of people, which were then used to justify different treatments for different groups. In the first part of the twentieth century, the eugenics movement—based on the notion that humans can control their own evolution—proposed that the human species could be improved by restricting reproduction by the 'feeble-minded' and facilitating the reproduction of those of superior ability and status. These attitudes reached an awful climax under National Socialism in Germany during the Second World War, when so-called 'lesser humans' were persecuted and exterminated; unfortunately, the attitudes still present themselves in some circles today (see Stephen Jay Gould's 1981 book, *The Mismeasure of Man*, for excellent counter-arguments).

The controversial book, *The Bell Curve*, by Richard Herrnstein and Charles Murray, was published in 1994. This book generated much debate worldwide, which remains ongoing today. Simply put, Herrnstein and Murray's thesis was that wealth and social advantage in US society was more and more going to an intellectual elite, and that affirmative action programs would not be helpful in boosting intelligence for those outside of this elite—particularly those from other racial groups—as intelligence is not readily modifiable through environmental actions. In a rapid response to the publication of this book, the American Psychological Association determined 'that there was urgent need for an authoritative report' on issues surrounding the controversies, and set up a task force to prepare a report on the 'knowns' and 'unknowns' about intelligence. This task force reported in 1996 (Neisser et al., 1996, p. 77) and its findings are discussed below.

However, before considering group differences on intelligence in more detail, it is important to reflect on two issues. First, as group differences found in intellectual functioning have been used to support discriminatory attitudes and support discriminatory policies, the evidence around these differences, and the explanation for them, need to be carefully assessed. Controversial results need strong supporting evidence. Second, even where group differences in intelligence are identified, they reflect differences in mean scores, while the underlying distributions of the two groups still substantially overlap. Consider two groups, say Group X and Group Y, where Group X has a higher mean intelligence score than Group Y. In this case:

- Each group will have substantial variability around the mean.

- The variability of each group around the mean will exceed the difference between the two means.
- There will be many individuals in Group Y who score above the mean for Group X.
- There will be many individuals in Group X who score below the mean for Group Y.
- Even knowing this information, it will not be possible to make a sensible comment about the intelligence level of any individual in either group; that is, it is not possible to extrapolate from group data to an individual.

Now, back to the controversies generated by *The Bell Curve*. Much of what was considered to be agreed upon by mainstream researchers in relation to group differences in intelligence, and intelligence testing, was summarised by Neisser et al. (1996), who produced the APA task force report, and Gottfredson (1997), who coordinated the expert opinion on intelligence that appeared in the *Wall Street Journal* in 1994. Regarding differences in intelligence among groups of people of different racial backgrounds: the mean IQ for distributions of scores for white people (based largely on US and European participants) centres on 100 IQ points. The mean IQ for distributions of Asian groups centres above that for whites, while the mean IQ for black and Hispanic groups centres somewhere below that for whites. Note, however, that members of all groups can be found at every level of IQ (see Box 7.1 for intelligence and ability assessment with Australian Indigenous peoples). The observable differences in intelligence among individuals are due to both genetic endowment and environmental influences, with heritability accounting for between 40 per cent and 80 per cent of the variability. The observable differences in intelligence among different groups are less well understood, and the reasons for these differences might not be the same as for the differences among individuals, even though genetic and environmental influences will be involved. Importantly, the differences cannot be explained by **test bias**: intelligence tests are largely accurate measures of intelligence.

test bias
the systematic favouring of one group over another in test outcomes; this can be due to more than one cause

Group differences also have been identified along other dimensions. By and large, men do not differ from women on global measures of intelligence, although some group differences are present for some specific abilities. Males tend to do better than women on some visuo-spatial tasks, for example, whereas women tend to score higher on some verbally based tasks. Intelligence varies for age

groups across the lifespan. The growth in intellectual functioning is very rapid up to the age of 10–12 years, continues to increase (though not as dramatically) until the age of 20–25 years, remains very much stable until about 60 years of age, and then declines noticeably after this age, with more striking declines for very old individuals. This progress, however, shows considerable variation among individuals. Intelligence also varies across generations. The average IQ score, for example, is significantly higher today than it was 50 years ago. This remarkable finding, that intelligence levels have been rising over the 100 years or so that intelligence tests have been used, is known as the **'Flynn effect'** (Flynn, 1987), and has been confirmed by many researchers. The average IQ increase is about 3 IQ points every decade, suggesting, for example, that the average IQ today of 100 IQ points would approximate an IQ score of 115 in your grandparents' day (based on two generations of 25 years each). As new intelligence tests are devised, or old ones revised, they are re-standardised against current populations, and this effect is not obvious; but nonetheless, it does exist. The causes of this increase in intelligence over time are unknown. Explanations for it include improvements in nutrition and the fact that people live in increasingly more complex societies, which might reflect, in turn, increases in body and brain size, and brains more stimulated by better educational, TV and life experiences.

Flynn effect
refers to a steady increase in scores on IQ tests since about the 1930s; first drawn to the public's attention by James Flynn

Box 7.1

Mental ability and Indigenous Australians

The early treatment of the Aboriginal and Torres Strait Islander peoples of Australia, like minority people elsewhere, was influenced to their disadvantage by social Darwinism and the eugenics movement. Scientific investigators viewed Indigenous Australians as being at 'an early stage of development', especially as they were considered to have been isolated on the Australian island-continent where they had existed in an 'untouched environment' for millennia: these were seen to be ideal conditions for the study of human evolution.

Psychologists and other social scientists had very early contact with Indigenous Australians, and included in their interests were mental abilities and how similar or different Indigenous Australians were from Europeans. In the late 1800s, an English expedition to the Torres Strait conducted 'brass

instrument' tests of sensorimotor function with the local Murray Islander people. These scores were compared with results from testing with English people, with the result that few differences were identified, and where they were found, some favoured the local people and some favoured the European sample.

Early in the twentieth century, Stanley Porteus (1883–1972) devised a supposedly culture-free intelligence test, the Porteus Maze (which is still used by neuropsychologists today for other purposes than measuring intelligence), and applied it to schoolchildren to identify those in need of special education classes. Porteus also used this test with Aboriginal mission children and Aboriginal adults, and found that both groups did poorly compared to Europeans. Despite Porteus finding a relationship between ability as measured by the test and exposure to Western cultures (indicating that the test was not culture-free), he went on to develop a 'racial hierarchy of intelligence', which no doubt fed prejudice within Australia and affected government policy towards Indigenous peoples.

Prior to the Second World War, a group of psychologists from Perth tested Aboriginal men and women in central Western Australia (see Kearney, deLacey & Davidson, 1973). These researchers found: (a) great variability in Aboriginal intelligence, suggesting that, as found elsewhere, ability distributions for Indigenous peoples overlap with distributions from non-Indigenous peoples; (b) that the difficulty levels for items in the test were equivalent across the Indigenous and non-Indigenous samples; and (c) performances for Indigenous people were associated with their level of contact with Western culture.

In the years following the Second World War, and based on a large sample of more than 1000 Aboriginal children and adults tested using the Queensland Test (an individually administered test of ability where the tasks are explicit and there is no need for any direction), McElwain and Kearney (1973) reported that the means for the Aboriginal groups were lower than for the European comparison group, but that the differences were largely in proportion to the Indigenous groups' contact with Western society. Aboriginal children were also tested for development on Piagetian constructs (e.g. conservation) and were found to be behind non-Indigenous children, although, again, contact with white society and Western schooling was associated with higher scores for the Aboriginal children (de Lemos, 1969). More recent testing with Aboriginal children of desert origin has suggested they might have advantages in some areas of cognitive functioning compared with European samples, such as visual memory and visual strategies (Kearins, 1981).

This account of ability testing with Indigenous Australians suggests: (a) that the results on the tests used were heavily dependent on exposure to Western culture; (b) that the results of the tests were sometimes used inappropriately (cf. Porteus) to the disadvantage of Indigenous Australians; (c) that we do not know from this testing where the mean for ability for Aboriginal and Torres Strait

Islander peoples stands in relation to other groups; and (d) that the distributions of ability overlap with distributions from non-Indigenous peoples. In addition to these testing issues, there were (and are) great differences within Indigenous groups as to their health and education status, community arrangements and cultural contexts, and great differences between Indigenous and Western people on these variables that must be taken into consideration.

Rickwood, Dudgeon & Gridley (2010)

Chapter summary

Intelligence is a difficult psychological construct to define for lay people and experts alike. Early experts dealt with this problem by conceptualising it as a unitary construct, and then devising tools to measure global mental ability. Only later, when sophisticated statistical software programs became available, was the construct of intelligence able to be decomposed into its component parts, and tests devised to measure these various aspects. Much fine-tuning remains to be done; however, measures of intelligence are now very sophisticated, and they generate confidence that what is being measured does meaningfully reflect an individual's general functioning. Despite this confidence, we need to remain vigilant that tests of intelligence, and test scores, are used for the good of all, and not used in ways to advantage any one group of people over another.

Questions

1. What was Binet's important insight that led to a 'paradigm shift' in the way human abilities were measured?
2. Early factor analysts can be loosely categorised into two camps based on their view of intelligence. What criteria would define these groups?
3. The CHC theory of intelligence offers a hierarchical account of intelligence. How does this hierarchy 'fit together'?
4. How well do the modern tests of intelligence (in our case, the Wechsler scales and the Stanford-Binet test) assess the three strata of ability reflected in recent theories of intelligence?
5. The identification of differences in intellectual functioning among various groups in the community has thrown up many challenges. How might two important groups, scientists and civic leaders, contribute to the general debate on these differences?

Further reading

Gottfredson, L S (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*, 13–23.
Neisser, U, Boodoo, G, Bouchard Jr, T J, Boykin, A W, Brody, N, Ceci, S J & Urbina, S (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101.
Schneider, W J & McGrew, K (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, NY: Guilford.
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. 459–83). New York, NY: Guilford.

Useful websites

Australian Psychological Society: www.psychology.org.au
Psychology Board of Australia: http://www.psychologyboard.gov.au
The MindHub: www.themindhub.com
US Association for Psychological Science: www.psychologicalscience.org

8 Personality

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. list the major personality paradigms that have influenced psychological testing and assessment
2. list the major techniques employed in personality testing and assessment
3. identify some of the strengths and weaknesses in the various approaches that have been adopted to personality testing and assessment
4. describe a way of integrating at least some of the various approaches.

KEY TERMS

clinical interview
empirical approach
interpersonal approach
multivariate (trait) approach
paradigms in personality assessment
personological approach
positive psychology
psychoanalytic approach
social-cognitive approach

Setting the scene

- An organisation is interested in developing its senior staff and considers that successful performance involves more than intellectual ability.
- A non-profit organisation wishes to identify ways in which newly incarcerated prisoners might best be assisted while in prison and in their transition to civilian life. This requires knowledge of their characteristic ways of adjusting to the world and their personal strengths.
- A careers counsellor is interested in the match between a client's skills and the demands of different occupations but may see the need for a broader assessment of the individual.

Introduction

We differ among ourselves in height, weight, skin colour, strength of grip and—more importantly for a personality psychologist—in terms of the characteristic ways we think, feel and behave. In spite of these differences we are all recognisably members of the same species making similar adaptations to our biological and social environment. That is, there are common mechanisms that underlie the differences in thought, feeling and behaviour that characterise different individuals. Personality is about both the individual differences and the common mechanisms (e.g. Cervone, 2005). This chapter is concerned with the ways psychologists have sought to open personality and individual differences to scrutiny both theoretically and in terms of practical methods of assessment.

The study of personality is probably as old as humankind. As intelligent creatures we try to make sense of our own and our fellows' behaviour. Certainly, from the Middle Ages onwards the idea of 'the person' became a topic of interest (Baumeister, 1987). The 'modern' study of personality, however, is only some hundred years old, with Freud's landmark work, *The Interpretation of Dreams*, appearing at the beginning of the twentieth century. Clinical investigation of patients with mental illness and later empirical studies with samples of people gathered in therapeutic, organisational, academic and community settings provided a store of information for theorising about personality as well as the development of a variety of means for assessing it.

Summarising such a broad literature is difficult without doing some disservice to both what is included and what is not. As a convenient expository device, the framework adopted by Jerry S Wiggins (1973, 2003) is used in this chapter. Wiggins proposed there were five basic **paradigms in personality assessment**, identifiable in terms of the different communities of scholars and practitioners writing about personality, the issues they each see as central, the ways in which they collect data, and the criteria they accept for resolving theoretical issues in the light of data. He termed these paradigms the

psychoanalytic approach, the interpersonal approach, the personological approach, the multivariate (trait) approach and the empirical approach. To these we add two approaches that came to prominence in US psychology in the latter part of the twentieth century (the social-cognitive approach) and the early twenty-first century (positive psychology).

paradigms in personality assessment
approaches to personality assessment that share: assumptions about how personality is best studied; methods for collecting personality data; and criteria for making judgments about what constitute adequate statements about personality

We offer a brief outline of each, including the major tools employed in assessment. In closing, we consider a more eclectic approach that attempts to combine several of the separate approaches. Before we begin, however, it is well to consider that personality assessment unless done professionally can result in plausible but superficial nonsense (see Box 8.1).

Box 8.1

Acceptance of personality assessments

Ulrich, Stachnik and Stainton (1963) had university students complete a personality test, then prepared personality assessments for each student apparently based on the test results, and finally asked the students to rate the accuracy of the assessments as descriptions of their personalities. Of the class (N = 57), 93 per cent rated the report as 'excellent' or 'very good' and only 2 per cent thought it a 'poor' assessment. The catch was that everyone received exactly the same personality assessment. The assessment that purported to be of them as a person described the overwhelming majority of the class just as well. How was this so?

The report was made up of statements that were ambiguous ('You are generally cheerful but sometimes depressed'), vague ('You enjoy a certain degree of order in your life') and favourable ('You are well-liked by others'), or were statements with a high base rate of being true ('You do not always find studying easy'). Paul Meehl (1956) termed the acceptance of assessments of these sorts of personality statements the Barnum effect, after the US impressario P T Barnum, who was known for giving the public what it wanted—to his considerable profit.

The demonstration by Ulrich et al. was a replication of an earlier study by Forer (1949) and the essentials of the demonstration have now been replicated

many times (Dickson & Kelly, 1985; Synder, Shenkel & Lowery, 1977). There is still some doubt about the factors influencing the effect but not of the effect itself.

Are these demonstrations of the Barnum effect little more than a parlour trick, momentarily diverting but of little long-term interest? No, they reveal some essential points about psychological assessment. First, they show that acceptance of personality assessments by those whom they purport to describe cannot be taken as evidence of their validity. The assessments can be convincing even though they fail to provide information that is specific to the person. Second, the demonstrations underline the need for useful statements about personality to go beyond what is superficially true for many and be specific to the individual case. If uniqueness is not being captured, what is their purpose?

This latter point is important in an age when computers are used to administer and score personality tests and to prepare assessment reports based on the results (Butcher, 2012). Computer generated reports can be useful in the assessment process but the Barnum effect alerts us to a potential limitation.

The psychoanalytic approach

An important source of information about personality is the thoughts, feelings and behaviour of those suffering mental disorder. Sigmund Freud, in his attempts to relieve the distress of patients he was seeing in his medical practice in Vienna at the turn of the twentieth century, developed a set of ideas about personality that were to have great influence on applied psychology, and on the humanities and social sciences more broadly (Ellenberger, 1970). These ideas became known as the **psychoanalytic approach**. For Freud, the starting point of understanding was that our behaviour, thoughts and emotions are the result largely of processes of which we are unaware (e.g. Brenner, 1974). These unconscious processes operate by binding and discharging a particular form of energy (libido) that is characterised principally by the sexual drive. As the infant develops, personality structures form to manage the investment and transformation of psychic energy as, first, the practical demands of reality and, later, the moral demands of society limit its expression.

psychoanalytic approach
an approach to personality that originated in the work of Sigmund Freud on the role of unconscious motivational processes in normal and abnormal personality functioning; it was elaborated on by a number of researchers during the course of the twentieth century

Freud's ideas, developed over 40 years, were revised on more than one occasion, and were initially supported and subsequently rejected by fellow practitioners (Monte & Sollod, 2003). Carl Jung, for example, saw the importance of the central idea of unconscious motivation in behaviour, but doubted the validity of the ways in which Freud elaborated the idea. In time, Freud's thinking was extended by his daughter, Anna Freud, who systematised his work on the so-called defence mechanisms, the characteristic ways individuals have of discharging impulses. Later, Erik Erikson broadened the account of development from that focused on sexual interest to one that included the response to challenges posed by the culture in which the individual matured, which he termed psycho-social development. There were to be further extensions. The personality structure that Freud proposed to handle the conflict between the requirements of reality and the internal demands for drive discharge—the ego—came to be seen as having a much more important 'conflict-free' role in the psychic life of the individual. The ego psychologists, as they were called, were joined by the object relations theorists who drew on Freud's observations about the significance of the relationship between the child and the caretaker in the early years of life, and by the self-theorists who stressed that aspect of the ego to do with phenomenological experience of personhood.

What we have now, 120 years since Freud began writing, is a storehouse of concepts, many with clinical relevance but few with experimental support. Together, they do not constitute a coherent theory of personality or of treatment, and a single unified theory of psychoanalysis (Freud's term) is not a goal for many who see value in this approach. It is more a matter of sifting the psychoanalytic literature for concepts that offer useful perspectives for practice, perspectives that are unlikely to be provided by other approaches to personality. The practice of psychological assessment, particularly in clinical settings, has been influenced in this way by psychoanalytic thinking. What to look for, how to interpret it and how these observations can provide an account of the person being assessed are questions that a psychoanalytic perspective can inform.

Assessment practice

Given the magnitude of the literature, it is impossible to catalogue in the space available the relevance of psychoanalytic concepts to assessment. Instead, a particular example of the approach is taken as illustrative and used to introduce some of the methods of assessment peculiar to this approach. The example is that offered by Drew Westen (1995, 1998), a psychologist and a contemporary advocate of psychoanalytic thinking.

Westen (1995) argues that there are three broad questions to be answered in personality assessment:

'What psychological resources [in the form of] cognitive, affective and behavioural dispositions, does the individual have at his or her disposal?'
'What does the person wish for, fear and value, and how do these motives combine and conflict to produce conscious experience and behaviour?'
'How does the person experience the self and others, and to what extent can the individual enter into intimate relationships?'

Answering the first question provides the context against which answers to questions two and three take on meaning. These are questions directed to the early concerns of Freud about unconscious motivation (question two) and the concerns of the more recent psychoanalysts with object relations and self (question three).

Answering these questions in any particular case takes some considerable time, and in classical psychoanalysis, where the patient was seen several hours a week for two to three years, there was the necessary time. Cost considerations mean that today such extensive contact is not possible, except for the very wealthy. Shorter methods of assessment are required. These include the **clinical interview**, in which developmental periods as well as current concerns are examined (Williams, 2011), and a battery of psychological tests, modelled on the original recommendations of Rapaport, Gill and Schaefer (1946).

clinical interview
a technique for collecting information about a client; it may take many forms, for example, a psychoanalytic perspective includes detailed exploration of the personal and family history of the client, particularly with respect to psychosocial development, conflict, and defence, self and interpersonal processes

The battery can include a variety of tests, but generally has as its core the Wechsler Adult Intelligence Scale (WAIS), the Rorschach and the Thematic Apperception Test (TAT). The WAIS is described in Chapters 7 and 9, and the TAT is described in 'The personological approach' section later in this chapter. The Rorschach is considered here. The details of the specific tests aside, the general point to appreciate is that assessment with the clinical battery is purposely wide-ranging to gather information that, together with the clinical interview, can answer broad questions about personality of the sort that Westen poses.

The Rorschach is the most widely used of what are termed the projective techniques (some authors refer to them as projective tests, but their status as tests is controversial). First published in German in 1921 and in English in 1942, it was developed by a Swiss psychiatrist from whom the name comes. Hermann

Rorschach (1844–1922) was interested in Jung's theory of personality types. He devised the technique by spilling ink on a page of plain paper and folding it in half, producing a blot symmetrical about its midline (see Figure 8.1). Ten such inkblots were produced, some with colour. The blots are purposely meaningless and the patient is asked to describe what they see in them. The account the patient produces is thought to say something about their perceptual process; that is, the way they see the world. Later (Frank, 1939), the 'projective hypothesis' was formulated to suggest that the person in responding to the inkblots is drawing on their unconscious to give the ambiguous stimuli meaning and is thereby revealing something of their unconscious mental life.

Figure 8.1 An example of a blot similar to that used in the Rorschach

http://en.wikipedia.org/wiki/File:Rorschach_blot_01.jpg

Although initially treated with great enthusiasm by clinical psychologists in the USA as a form of 'X-ray' of the unconscious mind, the Rorschach proved controversial. One source of this controversy was the difficulty experienced in capturing in a reliable way the yield of a Rorschach examination. Several scoring systems were tried, with the Exner system generally considered the best (Widiger & Saylor, 1998). A second source of controversy was the predictive validity, or rather the lack of it, achieved by Rorschach indices. One of the more recent critiques of the technique's validity was provided by Lilienfeld and his colleagues (Lilienfeld, Wood & Garb, 2000; Wood et al., 2003) who reviewed the large literature on the Rorschach and concluded: 'With a few exceptions, projective indexes have not consistently demonstrated incremental validity above and beyond other psychometric data' (Lilienfeld et al. 2000, p. 27). Not surprisingly,

this conclusion is not accepted by proponents of the technique (e.g. Hibbard, 2003). Further work by advocates (Mihura et al., 2013) and critics (Wood et al., 2015) of the Rorschach has led to some agreement that indices of cognitive functioning derived from the Rorschach have reasonable validity, but doubt remains about non-cognitive indices of emotionality and negative affect.

In seeking to make sense of the apparently irrational behaviour of neurotic patients, Freud reasoned that there were causes of their behaviour of which they were not aware and could not consciously control. The unconscious mind was a hypothesis that a century of research has neither confirmed nor refuted, but one that research teams, such as the one headed by Westen, continue to take seriously. It is a hypothesis that has given rise to several approaches to personality assessment, all of which continue to be controversial.

The interpersonal approach

Personality exists only in the personal interactions among people. This is the arresting proposition at the heart of interpersonal theory, or the **interpersonal approach**. Arresting, because we usually think of personality as something that the individual has; something that they carry around with them and express in different situations. For Harry Stack Sullivan (e.g. Monte & Sollod, 2003), however, personality is not a property of the person but of the interpersonal situation. The concept of personality is the observer's way of attempting to capture what it is that is happening when two people interact.

interpersonal approach
an approach to personality that proposes that personality exists only in the interaction between people and that the study of interpersonal processes is therefore central to personality assessment

Having stated the idea in its strongest form, it is necessary to add that the two people do not have to be physically present to each other—or even that both need to exist. The memory of a person (e.g. a dead father) or a fantasy person (e.g. the woman I thought I married) can give rise to a dyadic relationship in which personality is expressed and reinforced. It is the relationship that is critical and the ways in which individuals respond to and elicit responses from others that bring about the enduring characteristics that we unhelpfully 'locate' in them.

Sullivan was a psychiatrist and was concerned with understanding and relieving the symptoms of his patients. He was influenced by Freud but also by the Chicago school of sociology. For Sullivan the role of the psychiatrist was that of a participant observer, in the same way an anthropologist would work in trying to understand a new culture. Although the psychiatrist may seek to be simply an

observer, the very presence of the psychiatrist alters the situation and this must be recognised as part of the process. Sullivan came to see his patients' illnesses as exaggerations of patterns of responding to be found in 'normal' behaviour. These patterns or habits of relating—or 'dynamisms' as he termed them—develop early in life in the interactions with 'significant others' (a term introduced by Sullivan but now almost indispensable for psychologists talking about the influential people in a client's life). These patterns become recurring features of interpersonal behaviour that can distort the interactions with others. Problems in living (another phrase for which we have Sullivan to thank) arise out of these dynamisms.

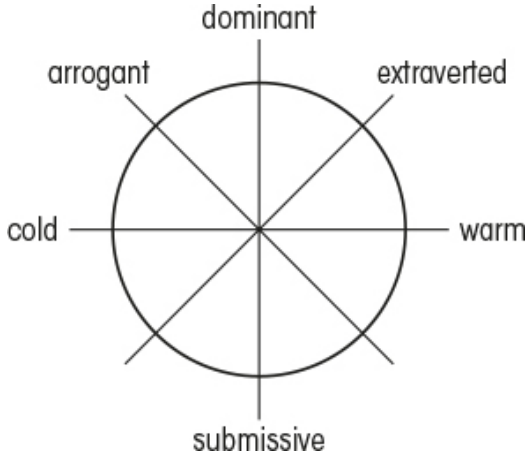
An important dynamism is that related to the self. It acts to protect against loss of self-esteem and to maintain a sense of security. From an early age the child is sensitive to praise and blame from the caregiver. Behaviours that lead to praise become part of the personification of the 'good me' and those that lead to blame part of the 'bad me'. There is also a personification of the 'not me': behaviours that are too awful to contemplate and that are removed from awareness through dissociation or selective inattention, similar in some respects to the Freudian concept of denial. Because of the importance of the self-dynamism, actions in interpersonal situations are often attempts to reduce anxiety associated with loss of self-esteem or to increase security in the relationship. These dynamisms can be self-defeating and generally work to prevent changes in behaviour that are necessary for successful adaptation.

Sullivan's work has had a direct effect on psychiatry through his discussion of the psychiatric interview and less directly through the development of methods of psychotherapy that are linked to but not directly derived from his work, such as the techniques of Kiesler (1996) and Klerman (Klerman & Weissman, 1993). In psychology, the major influence has been on the development of methods of assessment through the pioneering work of Timothy Leary on what has come to be known as the interpersonal circumplex (Wiggins, 2003), a way of describing interpersonal behaviour in terms of a circle of relationships.

The circumplex was proposed by Guttman and has been used to describe the interrelationship of emotions (e.g. Posner, Russsell & Peterson, 2005). Leary was the first to see it as a useful way of describing the components of social behaviour and their interrelationship. For example, some behaviours (e.g. dominance and submission) appear to be the opposite of each other but independent of others that are themselves opposites (e.g. warmth–coldness in interpersonal relations). Some are blends of others; for example, extraverted behaviour can be described as a blend of warm and dominant behaviour, and arrogant behaviour as a blend of dominant and cold behaviour. A model with some subtlety is needed to capture these varying relationships, and so the circumplex suggested itself to Leary. Any line through the centre of a circle joins points on the circumference that are polar opposites; that is, 180 degrees from each other. A line through the

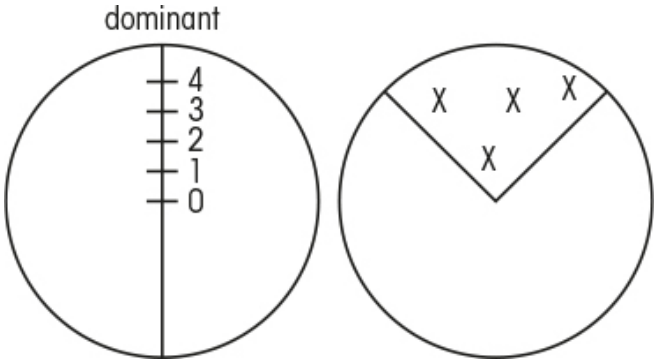
centre at 90 degrees to the first line is maximally different from it; angles smaller or larger than 90 degrees bring the second line into closer relation to the first. A circle can thus map the pattern of relationships (or lack of them) among behaviours. An example appears in Figure 8.2, which maps the pattern proposed earlier in the paragraph.

Figure 8.2 Example of an interpersonal circumplex



As well as mapping the location of dimensions of interpersonal behaviour, the circumplex can indicate the intensity of a particular behaviour in terms of the distance of a point from the centre of the circle. The centre can be taken as indicating the point of indifference between the polar opposites of the dimension, and increasing distances towards the circumference as indicating stronger or clearer expressions of the characteristic, with the point on the circumference indicating the strongest expression. The circle on the left of Figure 8.3 shows notional units of dominance increasing from the centre to the circumference. Given that there is more than one dimension and that dimensions can show varying degrees of relationship, it is possible for a segment of the circle rather than a single line to best describe a pattern of interpersonal behaviour. For example, an effective manager might be high on dominance but also high on related dimensions such as decisiveness, competitiveness and the capacity to influence others. High scores on these dimensions might cluster, as in the circle on the right of Figure 8.3, to reveal a managerial 'type'.

Figure 8.3 Intensity (left figure) and clustering (right figure) of behavioural dispositions can be represented in the circumplex



Leary was the first to show empirically that observations of behaviour in interpersonal situations, collected in studies of university students and patients undergoing therapy, were a good fit to a circumplex model. Subsequent research using other data sets replicated the finding, although the number of dimensions (lines through the circle) that are needed is somewhat contentious. Some have suggested eight and some sixteen, with the number of segments of the circle that these make possible varying as a result. Wiggins (2003), for example, argued for eight dimensions and sought to assess a person's position on each with eight rating scales.

Although there is some disagreement about the number and nature of the dimensions necessary, there is reasonable consensus about the reference dimensions of dominance–submission and warm–cold. Horowitz (2004) has argued for a broader interpretation in terms of agency and communion, with dominance–submission reflecting a broad human tendency to seek control over one's environment, including the social environment, and warm–cold reflecting the need for belongingness. Certainly, dimensions similar to these, although not necessarily so named, can be found in the pattern of interrelationships of measures from a number of personality dimensions, and not necessarily constructed using the interpersonal approach as the starting point.

Assessment practice

The interpersonal approach has influenced more than the development of the interpersonal circumplex and the scales that have been based on this method, such as the Interpersonal Adjective Scales (IAS; Wiggins, 1979) and Revised IAS (Wiggins, Trapnell & Phillips, 1988). The IAS is a 128-item self-report instrument that uses an eight-point response format ranging from 1 (very inaccurate) to 8 (very accurate) on trait-descriptive adjectives. Example items from the Gregarious-Extraverted (NO) scale are 'cheerful' and 'outgoing'. The Revised IAS consists of 64 adjectives with the same response format. Morey (2003), in developing the Personality Assessment Inventory (a broad-band assessment of adult psychopathology), included the two major dimensions of

dominance–submission and affiliation–rejection. Scores on these scales have been shown to relate to the dimensions assessed by the interpersonal circumplex (Ansell et al. 2011). The PAI is described in Chapter 9. A further example of the use of the circumplex but with vocational interests rather than personality dimensions is discussed in Chapter 13.

Leary's work based on Sullivan's theorising is now more widely discussed in the literature on personality assessment than Sullivan's; however, as Sullivan maintained, the importance of interpersonal behaviour and the need to assess it in clinical and other settings is not disputed.

The personological approach

The term 'personology' was coined by Henry Murray to describe the study of the person (or the **personological approach**), for which 'personality' is now the much more usual term. Murray, with Gordon Allport, pioneered the academic study of personality in non-clinical samples (e.g. Monte & Sollod, 2003). In Murray's case, the sample was the rather atypical one of Harvard undergraduate students who, given the nature of Harvard University at the time, were male, academically able and likely to be from well-to-do families. Murray and his colleagues studied a sample of forty-three young men over a period of three years, using self-report tests, clinical interviews and specially devised performance tests.

personological approach
an approach to personality that began with the work of Henry Murray who sought to study personality in terms of the (principally) psychogenic needs of the individual and the extent to which the environment promoted or inhibited these needs

The project gave rise to a number of innovations. One was the development of a new theory of personality directed at normal, as distinct from abnormal functioning—the domain that Freud, Jung and the psychoanalytic movement had explored extensively. Murray's theory stressed the motivational basis of behaviour, just as psychoanalytic theory had, but broadened motivation to include social or 'psychogenic 'concerns, as he termed them, as well as the viscerogenic or biologically based concerns to do with food, sex and elimination. The concept of 'need' was introduced to account for the motive force for behaviour and Murray identified twenty-seven psychogenic needs, based on his work with the sample of Harvard undergraduates.

Murray went beyond attempting to catalogue needs to describe their important features. One of these features was that needs do not operate in a

vacuum but are part of the psychological environment of the person. Some environments may be highly conducive to the expression of a particular need and some environments may frustrate its expression. The 'press' of the environment, Murray contended, must be considered with the person's needs. Press can be of two sorts. Alpha press is the environment as the individual perceives it and beta press is the environment as it appears to observers. These two forms of press are usually in reasonable harmony, but for Murray it is alpha rather than beta press that is crucial.

A second important characteristic of needs is that they can be conscious or unconscious. That is, a person may be aware of and articulate the presence of a particular need, but needs that are not recognised may still exert important influences on behaviour. Conscious and unconscious needs may be aligned, but when they are not, motivation is more complex to understand. Conscious needs can be identified by asking the person about their existence—for example, by having them complete a questionnaire—and Murray's catalogue of needs has been used to develop a number of self-report tests of personality, including the Adjective Checklist (Gough & Heilbrun, 1983) and Douglas Jackson's Personality Research Form (PRF; Jackson, 1984). Unconscious needs, on the other hand, cannot by definition be arrived at in that way and require a more subtle form of assessment.

Just as Freud turned to fantasy to explore the unconscious mind through the meaning of dreams, Murray and his colleagues turned to projective techniques to examine unconscious needs. One of those colleagues was Christiana Morgan, who had considerable artistic talent, as well as an interest in the role of the unconscious in human affairs. Murray and Morgan knew of the Rorschach, but developed their own projective technique using a set of ambiguous pictures, several of which were drawn by Morgan. These were depictions of people and places rather than meaningless inkblots, but allowed for more than one interpretation. In using these they asked respondents to tell them what was happening in the picture, how it came about, and what was the likely resolution. Because the pictures were ambiguous and permitted several different meanings, Murray argued the respondent had to draw on their own motivational life to provide an account. The Thematic Apperception Test (TAT), as they titled it, came to be the second most widely used projective technique after the Rorschach.

Figure 8.4 An example of a picture similar to that used in the TAT

A TAT-like photograph. What is happening here? Who is involved? What led up to this? What will the outcome be?

A further contribution of Murray to personality assessment was the introduction of the diagnostic council, which was to have an impact on assessment practice—although not under that name. In reviewing the data provided on each member of the sample of Harvard undergraduates, Murray had all those involved in assessments meet to discuss the findings. The approach was similar to the case conference in clinical medicine where the data on a patient is reviewed by the treatment team and opinions offered and weighed in formulating a diagnosis and a treatment plan. So, too, Murray's diagnostic council considered all the information in assessing each of the research participants.

In doing this, Murray used the life history as the essential framework, because to Murray 'the history of the person is the personality' (Hall & Lindzey, 1978, p. 211). To make sense of what is happening now, one needs to look at what happened in the past, which again is not so surprising for a psychodynamic theorist, although Murray was not thinking particularly of what happened in early infancy. In examining the life history, Murray looked for what he called 'proceedings' and 'themes'. The former are particular periods when significant events occur that are important later in life. They could be transition points, or times of crisis or particular accomplishment. Themes are ideas that recur in the life of the person and help to give it some structure or coherence.

Murray used the idea of diagnostic council when, as with many psychologists, he joined the war effort with the entry of the USA into the Second World War. He worked in the Office of Strategic Services (OSS), the forerunner of the CIA, and was engaged in assessment of those whose task it was to work behind enemy

lines gathering intelligence or engaging in sabotage. The danger presented in these circumstances called for particular qualities, and Murray and his team devised a number of tasks to assess these. The work was described in the *Assessment of Men* (Office of Strategic Services Assessment Staff, 1948). To bring all the information together and determine the suitability of a candidate for a dangerous assignment, Murray used the diagnostic council.

Murray's approach was adopted postwar in the form of assessment centres used for executive selection and promotion in a number of US corporations, but beginning with Bray's work in American Telephone and Telegraph (Bray & Grant, 1966; see also Chapter 10). At a location away from the corporate headquarters, a small group of staff—five to ten—would be put through a number of tasks over two to three days while being observed by senior staff of the firm. At the end of the time the observations would be brought together in a meeting of those concerned, and the assessments and recommendations jointly made by the senior staff adopted within the organisation. True to the diagnostic council, there was an emphasis on judgment by those who had observed the participants over a period of time and who were familiar with the business the participants were seeking to pursue.

Murray's work was not the only influence on assessment centres. The idea had originated in the German Army before the First World War and had been used successfully by the British Army in their War Office Selection Boards for selecting officers for service (Jeanneret & Silzer, 1998). Murray's work provided a richer theoretical matrix for what was found to be a practically useful method for selection and staff development.

Assessment practices

Murray's work has been used to develop multi-scale self-report measures of personality such as the PRF noted earlier. As well the motives of achievement, power and affiliation that Murray identified have been explored in detail in major research programs using the TAT and related materials by McClelland, Winter and others (see, for example, Brody & Ehrlichman, 1998) and a good deal of information gathered on the issues of reliability and validity. Lilienfeld, Wood and Garb (2000) in their critique of projective techniques were not as harsh in their criticism of the TAT as they were of the Rorschach. Smith (1992) has provided details for the reliable scoring of the TAT for a number of motivational indices, and Schultheiss' (2008) review indicates that many practical correlates of motivational indices have been established on the basis of the work of Murray and the TAT. Spangler (1992), for example, in a review of the literature found evidence of incremental validity of the TAT indices of achievement motivation in that they increased the validity provided by self-report measures of the construct;

and McClelland and Burnham (2003) showed how these concepts could be used in executive selection.

The multivariate (trait) approach

The oldest approach to personality, with origins in Greek and Roman times, is the type or trait approach. Asked by a friend or a potential employer seeking a reference to describe someone, we find ourselves listing attributes we perceive them as having; for example, conscientious, easy going, careless, punctual, tense or good fun to be with. The longer we know them, the richer the description is likely to be, with qualifications and examples. We use the language of trait theory because it seems so natural, and some would argue that indeed it is. Our language over time has coded important characteristics of others, characteristics that we need to know about if we are to work and live with them—an idea that originated with Galton in 1884 (Lubinski, 2000). Others would argue that we need to be very cautious in making inferences of this sort because what we understand from our language may not be an accurate reflection of how others behave. The language seduces us into seeing consistencies in others' behaviour where none exist (e.g. Shweder & D'Andrade, 1980).

In earlier times, types rather than traits were the preferred basis of description, with Galen proposing four temperament types (melancholic, phlegmatic, choleric and sanguine) based on the distribution of hypothetical bodily 'humours' (Kagan, 1994). The typology, with variants, served reasonably well, but in the twentieth century theorists began to talk more of traits as dimensional concepts; that is, continua that encompassed a number of points between extremes. A continuum accommodates a more flexible system of description than the dichotomies afforded by type concepts (e.g. a person may be *somewhat* phlegmatic rather than having to be either phlegmatic or not).

Allport was the first to formalise a trait theory of personality (which would eventually lead to the **multivariate (trait) approach**), although he saw certain traits as unique to individuals rather than being common to all. He and his co-worker Odbert (Allport & Odbert, 1936) used the English language dictionary in common use at the time to list all the words used to describe people. In so doing, he provided a pool of terms for later researchers to develop into taxonomies of traits. RB Cattell, for example, used the Allport and Odbert item pool to construct a number of rating scales with which judgments of personality could be made by peers, parents or teachers (Cattell, 1946). A pair of terms from the list (e.g. timid versus adventurous) was used to anchor each end of a horizontal line on a page. The task for the rater was to indicate where on the line the person being rated should be located. Cattell developed 173 such scales and administered them to a number of groups and then used factor analysis (see

Chapter 5) to sort the similarities and differences among the scales. He identified twelve factors that appeared repeatedly in his analyses of the different groups and labelled these with his own set of trait terms. He later examined these dimensions using self-report rather than peer ratings; that is, sets of short questions about typical ways of behaving, to which the person had to respond 'true', 'false' or 'unsure', depending on their personal relevance. He found the same twelve dimensions plus four others that had not appeared in the rating data. Together, the sixteen factors constituted for Cattell the primary personality factors and he developed a personality questionnaire (the 16 PF) for their assessment (Cattell & Mead, 2008). Cattell subsequently sought the expression of these factors in a number of behavioural and physiological tests.

multivariate (trait) approach
the oldest approach to personality that in its modern form proposes that there are a number of dimensions of individual difference that people have in common and that serve to specify the individual's personality

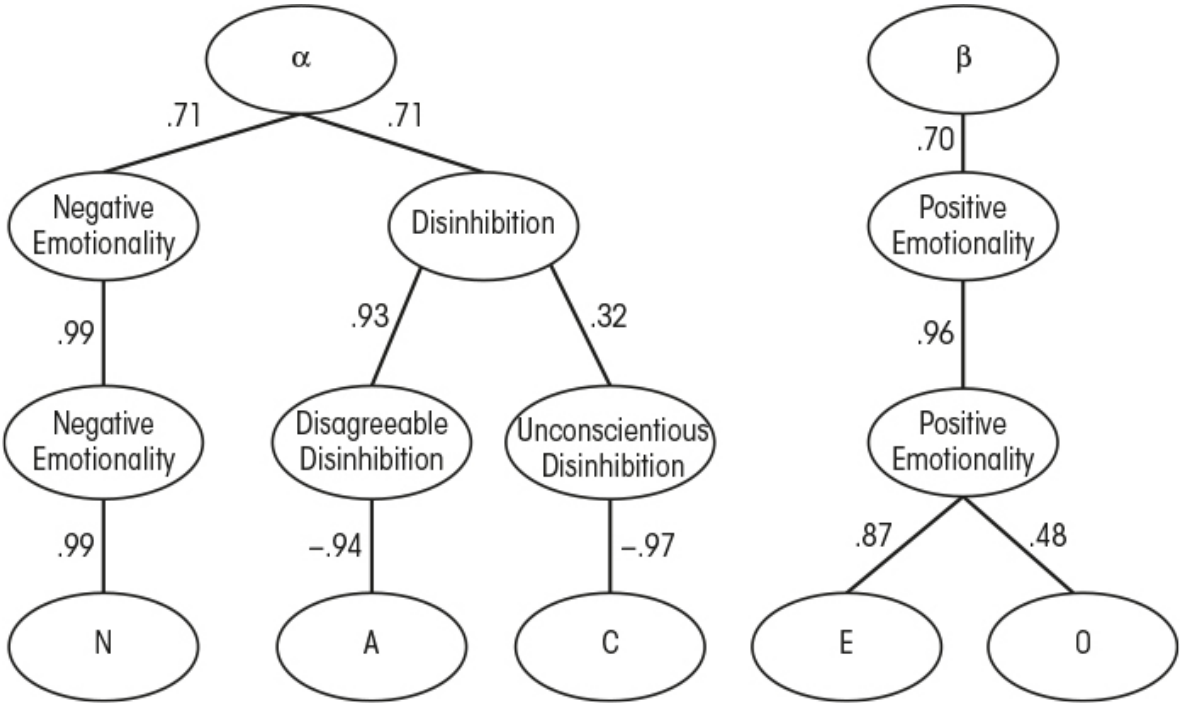
H J Eysenck (e.g. Eysenck & Eysenck, 1975a) used factor analysis to explore the personality trait domain, but rather than allow whatever pattern there may be to emerge from the data, he set out to test specific hypotheses about personality. He drew these hypotheses from the writings of the descriptive psychiatrists whose work preceded Freud's, and who made different starting assumptions from those of Freud. For example, rather than considering neurosis and psychosis on a continuum of seriousness, as Freud had—with the neurotic destined for psychosis if their symptoms were to decompensate—the descriptive psychiatrists saw neurosis and psychosis as separate disorders with no necessary connection between them. Eysenck used factor analysis and batteries of self-reports, peer ratings and behavioural tests with samples of patients and demonstrated that there were three major dimensions of personality consistent with the accounts of the descriptive psychiatrists. He labelled these neuroticism (a dimension that separated neurotic patients from healthy control participants), psychoticism (a dimension that separated anxiety patients from those with psychotic disorders such as schizophrenia) and extraversion (a dimension that separated the patients with neurotic disorder into those with hysterical symptoms and character disorders from those with dysthmic symptoms such as anxiety and depression). He then showed that these three dimensions could be found in samples of individuals without clinically definable disorders. That these were the same dimensions was evidenced by the fact that the response patterns that defined the dimensions in the non-clinical samples were the patterns that gave rise to the discriminations among the clinical groups. His Maudsley Personality Inventory

(later the Eysenck Personality Questionnaire, EPQ; Eysenck & Eysenck, 1975b) was used to measure the three dimensions.

Eysenck thus argued for three major personality dimensions whereas Cattell argued for sixteen. Some resolution of the apparent contradiction came about when it was shown that three factors or a larger number could be identified in the one data set, depending on the strategy of data analysis adopted and the stage of the analysis at which a result was declared. Permitting correlations among factors made for a greater number, as Cattell had shown, but these in turn could be further analysed to produce the three that Eysenck had claimed.

The differences between the supporters of Cattell's and Eysenck's approaches were overshadowed by the research program of Costa and McCrae (1992a; McCrae & Costa, 2008) that showed that rather than three or sixteen factors, the optimal number was five. Although not the first to demonstrate the value of a five-factor solution—that honour should go to Tupes and Christal (1961/1992)—McCrae and Costa showed that, irrespective of what self-report personality test or set of rating scales was used as a data source, the same five factors could be extracted. Further, these five factors were the same across sexes, ages and, to a large extent, cultures (see, for example, John & Srivastava, 1999). These then were the Big Five, the largest and most ubiquitous factors of personality. Table 8.1 lists these and some examples of the traits that coalesce to determine them.

Figure 8.5 The structure of personality



Markon, Krueger & Watson (2005, p. 23)

Table 8.1: The Five-Factor Model (FFM)

Factor	Characteristics
Neuroticism	Worried versus calm
	Insecure versus secure
	Self-pitying versus self-satisfied
Extraversion	Sociable versus retiring
	Fun-loving versus sober
	Affectionate versus reserved
Openness	Imaginative versus down-to-earth
	Preference for variety versus preference for routine
	Independent versus conforming
Agreeableness	Soft-hearted versus ruthless
	Trusting versus suspicious
	Helpful versus uncooperative
Conscientiousness	Well organised versus disorganised
	Careful versus careless
	Self-disciplined versus weak-willed

Burger (2000, p. 186)

The Five-Factor Model (FFM) is now the dominant position in trait psychology, although it has its critics and although it is clear that there are more than five factors in the factor space of personality tests (Block, 1995; Paunonen & Jackson, 2000; Zuckerman, 2002). Seven factors or nine have been proposed, but those identified after the first Big Five (e.g. optimism) are not as large or as frequently observed. The FFM can be reconciled with a number of other models of personality factors, such as Eysenck's three-factor model, by paying attention to the level at which factors are extracted. Markon, Krueger and Watson (2005), in a thorough exploration of the factor space, showed this, as depicted in Figure 8.6. At the second level of factor extraction, the factors of positive and negative emotionality are Eysenck's extraversion and neuroticism, and the disinhibition factor corresponds to Eysenck's psychoticism. At the fourth level the disinhibition factor appears as the separate FFM factors of Conscientiousness (C) and Agreeableness (A), such that the person high on psychoticism has low scores on both C and A.

Assessment of personality using the trait approach is most commonly done using the personality questionnaire in which the person being assessed is asked to report on their own behaviour using a series of short statements (see Box 8.2). There have been several criticisms made of the personality questionnaire, including that it requires a good deal of insight and honesty on the part of the respondent. The opportunity that it provides to dissimulate, to fake a particular personality profile or to provide a socially acceptable view is seen as a major impediment to its use in practical assessment situations (see Box 8.3). A person seeking help for psychological disorder may present their situation as worse than it actually is to solicit support, whereas a person seeking a job may present themselves in as favourable a light as possible. These problems are real, as any number of studies of faking on personality tests has shown, but can be exaggerated (e.g. Topping & O'Gorman, 1997).

Box 8.2

A short personality questionnaire

The Dirty Dozen

Respond to the following statements as honestly as you can by circling the option that applies to you.

1	I tend to manipulate others to get my way.	True/False
2	I have used deceit or lied to get my way.	True/False
3	I have used flattery to get my way.	True/False
4	I tend to exploit others towards my own end.	True/False
5	I tend to lack remorse.	True/False
6	I tend to be unconcerned with the morality of my actions.	True/False
7	I tend to be callous or insensitive.	True/False
8	I tend to be cynical.	True/False
9	I tend to want others to admire me.	True/False
10	I tend to want others to pay attention to me.	True/False
11	I tend to seek prestige or status.	True/False

12	I tend to expect special favours from others.	True/False

Jonason and Webster (2010) developed this set of questions, which they termed the 'Dirty Dozen', as a brief assessment of the Dark Triad. This is a constellation of personality traits that reflect a 'socially malevolent character' originally identified by Paulhus and Williams (2002). The traits comprise Machiavellianism, psychopathy and narcissism, which involve coldness and manipulativeness, thrill-seeking and lack of empathy, and grandiosity and entitlement. Although these characteristics in exaggerated form can give rise to a diagnosis of personality disorder, they can be exhibited by normally functioning individuals and may in fact lead to success in organisational settings (Board & Fritzon, 2005).

Box 8.3

Self-presentation and response sets

The origins of the term 'personality' lie in Greek and Roman usage where, for example, the *per sonare* was the mask used by an actor in a Latin play to conceal the true self while representing a character in the drama (Monte & Sollod, 2003). The term evolved over time and we now think of personality as the essence of the authentic self. A moment's reflection, however, brings to mind a number of instances where what we see of the person is possibly not what is true but a presentation for a particular purpose. On a first date, we are wanting to put our best foot forward and may not appear the person we are, say, two years into a relationship. Likewise, in applying for a job we are keen to have, it is understandable that the self on display is the best we can make him or her. On the other hand, we might exaggerate psychological problems if we consider that will assist in obtaining the treatment we need, or cognitive deficits in a personal injury insurance claim. This may be deliberate or may result not from a conscious desire to distort but from a lack of insight into our own behaviour. Personality assessment has to contend with the fact that both impression management and self-deception may be at work so that the self as publicly presented is not the 'authentic self'.

Self-presentation effects have been studied for well over 50 years, both as the result of the context in which the assessment is being made and as individual difference characteristics in their own right (see Paulhus & Trapnell, 2008). We know that participants can 'fake good' and 'fake bad' in responding to

personality tests; that some items in these tests are rated by participants as higher in social desirability than others; and that some people more than others are likely to endorse items that present them in a favourable light but which are highly improbable ('I have never told a lie') and that this characteristic shows reasonable stability over time. Counter measures have been proposed such as the use of instructions to limit distortion or the use of 'lie' scales to detect when people are making improbable claims about themselves, but none is foolproof.

The use of a lie scale began with the MMPI (the L scale), but has been used in a number of other personality tests such as the EPI and EPQ (Furnham, Eysenck & Saklofske, 2008). Cut-off scores on the lie scale for discounting a respondent's scores as subject to dissimulation are usually proposed by the test developer, but it is not clear in any particular case whether this practice is effective. As well as the lie scale, developers of the MMPI offered a defensiveness (K) scale to identify those who may have psychological difficulties but are unwilling or unable to admit to them. Some level of ego defensiveness is normal, but when greater than this the respondent's answers to other questions are called into doubt. The F scale of the MMPI was an attempt to assess over-claiming about difficulties. It consisted of items that suggested psychopathology but which patients with mental disorder did not in fact endorse. It is sometimes referred to as malingering scale. The 'cannot say' scale was simply a measure of the number of questions that were left unanswered. As well as these, infrequency scales are sometimes used to check that the respondent is conscientious in their approach to answering a questionnaire. The infrequency scale on the PRF includes a number of items that are unlikely to be true ('I have recently returned from Switzerland where I have been repairing cuckoo clocks').

A more subtle effect on the accuracy of personality tests is to be found in the idea of acquiescence response set. This is the tendency to endorse items positively rather than negatively independent of the content of the item (Winkler, Kanouse & Ware, 1982); the opposite effect is also possible but has been less studied. The solution originally advocated for this problem was to develop balanced scales in which an equal number of items indicating a particular trait were worded to elicit positive and negative responses. A high score on the trait can then not be obtained by simply endorsing all items in a favourable direction. Some have argued, however, that this simple solution can compromise the validity of the scale (Schriesheim & Hill, 1981). Acquiescence is but one example of a number of response sets that can influence the results of personality assessment.

Assessment practices

The multivariate (trait) approach has led to the development of a number of personality scales. One of the most widely used currently is the NEO set of tests for assessing the five-factor model. The NEO Personality Inventory-3 (NEO-PI-3) is the most recent version of the NEO-PI-R (Costa & McCrae, 1992b). It consists for 240 items and takes about 35 to 45 minutes to complete. It can be used with adolescents (12 years +) and adults. It measures thirty facet scales, six for each of the five factors (or domains, as the authors refer to them) listed in Table 8.1. The facet scales are theoretically derived rather than empirically based. A full listing is available in McCrae (2009). The NEO Five Factor Inventory-3 (NEO-FFI-3), a revision of the NEO-FFI, provides a quick measure of the five factors, but not of the facet scales. It consists of 60 items and takes no more than 15 minutes to complete.

Reliability of the factor scales is good. For the NEO-FFI, the shorter and therefore less reliable of the scales, internal consistencies of the N, E, O, A and C scales are reported as 0.89, 0.79, 0.76, 0.74 and 0.84, respectively, and test-retest stability over a two-week period as 0.89, 0.86, 0.88, 0.86 and 0.90, respectively (Costa & McCrae, 2008), Validity is based on factor structure and the convergence and discrimination of the scales across different methods of measurement. For example, the five factors of the NEO-FFI were shown to correlate across like factors in self-ratings, peer ratings and spouse ratings, and at substantially higher levels than they correlated with unlike factors using these different methods (see Costa & McCrae, 2008). There are also public domain versions such the Big Five Inventory (BFI; John & Srivastava, 1999) and the International Personality Item Pool (Goldberg, 1999).The Australian Personality Inventory (API) is a fifty-item public domain version developed for use with Australian samples (Murray et al., 2009).

The 16 PF is now in its fifth edition (Cattell et al, 1993) and has an extensive research literature to support it. The 16 PF measures sixteen primary factors (one of which is intelligence), five global (second order) factors, and has in addition three response bias scales. It consists of 185 items that take about 35 to 50 minutes to complete, is suitable for 16 years and older, and is available in thirty-five languages. There is a short version, the 16 PF Express (Gorsuch, 2006) that measures the sixteen factors. Internal consistency varies from 0.66 to 0.86 across the sixteen factors and is estimated, as composites of the primary scales, at an average 0.87 across the five global scales. Validity is demonstrated in terms of factor structure from large scale (N = 10,000+) exploratory and confirmatory factor analyses. There is good evidence for the predictive validity of the 16 PF when used in organisational and educational settings (e.g. Cattell & Mead, 2008).

The empirical approach

The **empirical approach** bears some similarities to the trait approach in that personality questionnaires figure largely in both, but the essential difference is that whereas the trait approach is concerned principally with the dimensions that make for human individuality (i.e. personality structure), the empirical approach is concerned with personality description in the service of predicting socially relevant criteria; that is, aspects of behaviour that the society has an important stake in, such as mental illness, criminality, academic achievement and work performance. The empirical approach asks what the relationship is between measures of individual differences and measures of socially relevant criteria, irrespective of what might be the basis or cause of such a relationship. For the adherent of the empirical approach to know that personality measure x predicts at better than chance level criterion A (where x might be a self-report measure of adherence to social norms and A is the likelihood of being charged with delinquent behaviour) is sufficient justification for research effort.

empirical approach
a way of constructing psychological tests that relies on collecting and evaluating data about how each of the items from a pool of items discriminate between groups of respondents who are thought to show or not show the attribute the test is to measure; also an approach to personality that relates the reports that people make about their characteristic behaviours to their social functioning and thereby provide tools for personality prediction

The development of the MMPI (Minnesota Multiphasic Personality Inventory) followed this strategy. Starke Hathaway, a psychologist, and John McKinley, a psychiatrist, sought a way of assisting the differential diagnosis of patients presenting with a range of symptoms at a large psychiatric hospital (Hathaway & McKinley, 1943). They developed a questionnaire, much as Woodworth had done in the First World War, to screen recruits for adjustment difficulties, but rather than screen for a single disposition 'adjustment' they sought to screen for nine psychiatric diagnoses: schizophrenia, hypomania, psychasthenia, paranoia, psychopathy, hysteria, depression and hypochondriasis. To these they added items for assessing masculinity–femininity and social introversion. They assembled more than 500 items for this purpose by borrowing items from previous questionnaires and by including questions that psychiatrists would ask in making diagnoses. The content was purposely diverse and included questions such as 'I enjoy horseback riding', 'Christ performed many miracles such as changing water into wine' and 'I am seldom troubled by constipation'.

Hathaway and McKinley administered this item pool to patients with known diagnoses—that is, diagnoses about which there was agreement among consulting psychiatrists in the hospital—and tallied the frequency of endorsement of each of the items by these different groups. To provide a basis for

comparison they recruited relatives who were visiting patients in the hospital and tallied the frequency of endorsement of items among this presumptive 'normal' group; that is, without diagnosed mental illness. They then checked the total item set looking for items where the frequency of endorsement for a particular patient group was substantially greater or smaller than it was for the 'normal' group. Where the difference was what they judged to be substantial, they had a possible 'predictor' of membership of the patient group. It was 'possible' because the capacity of the item to discriminate had to be checked in further samples.

The content of the item was not important, only the fact that its frequency of endorsement differed between patient groups and community samples. The observation of a difference was what mattered and not any preconceived idea about whether or not the item should discriminate. For example, the item 'I like horseback riding' was added to the schizophrenia scale because it discriminated schizophrenic patients from normals. The reason that it 'worked' was not of concern, nor was the question of whether those who endorsed the item actually rode horses and liked doing so. The empirical fact was that, faced with the question, patients with the diagnosis of schizophrenia were more likely to say 'true'. Collecting a number of such questions allowed the formulation of a schizophrenia scale. The process, termed 'criterion keying', was repeated with the other diagnostic groups and the set of nine diagnostic scales developed, plus the masculinity–femininity scale and the social introversion scale.

With the development of the questionnaire, psychologists had a means of assisting in diagnosis. An incoming patient would be given the complete set of scales (550 items in the original form) and their pattern of endorsements compared against that of the various diagnostic groups and where a match was found a possible diagnosis was made. This would be provided to the psychiatrist to aid diagnosis based on clinical interview. The 'diagnosis' from the MMPI was not always (in fact, seldom) clear-cut; where it was clear-cut, it was not necessarily agreed to by the psychiatrist interviewing the patient. Under favourable conditions, the success rate for the identification of a particular diagnostic group was about 70 per cent, but this was achieved with the incorrect identification of some who were 'normal' (see Chapter 5 on the decision-theoretic approach). The authors themselves were modest in the claims they made for it. If it did not provide the key to differential diagnoses they had been looking for, it provided a good deal of information and, when this was capitalised on, the success of the instrument was assured.

To quantify differences in endorsement, scores for each scale (the number endorsed in the direction indicating psychopathology) were expressed as T scores (see Chapter 3). A T score of greater than 70 (i.e. two standard deviations above the mean) was taken as indicative of the diagnostic category to which the scale applied (e.g. depressive or schizophrenic). With ten scales and T scores for each, a profile could be drawn up and patterns rather than individual scale scores

compared for various groups. A shorthand way of doing this was to take the three scales for which scores exceeded 70 and then describe the features of the patients with these scores. Initially an atlas (a book of profiles) was prepared, but in time computer scoring became possible.

The empirical approach followed by the authors allowed for the proliferation of scales because it was only necessary to find groups that differed in some way for the item set to be used to derive a new scale. The original authors seized on an obvious difference between people in terms of gender to make a masculinity–femininity scale, although this was at a time when this distinction had more importance than it does today. It was a time, too, when psychiatric opinion still favoured an interpretation of homosexuality as a disorder. But the ease of scale construction meant that in time over hundreds of different scales in addition to the original scales were derived from the MMPI item pool.

There were problems with the test recognised from its earliest usage. There was item overlap among scales so that the one item might indicate the likelihood of belonging to more than one diagnostic category, and in some instances several. Not surprisingly, there were also correlations among the scales, so that likelihood of being diagnosed as depressive was associated with an increased likelihood of being diagnosed as schizophrenic. Complete separation of categories is unlikely, but it was soon recognised that many items were detecting a general feature of being a hospitalised patient (called by some 'demoralisation') rather than any specific disorder. Lack of control over one's life, in whatever regard that had led to being hospitalised, gives rise to responses to questionnaire items of the sort used in the MMPI that indicate demoralisation. There were also problems associated with the unrepresentative nature of original samples used, both patient and community, given that they were relatively small and limited in terms of geographical representativeness.

Assessment practices

The versions of the MMPI now in use have been developed with recognition of these problems in mind. The MMPI–2 (Butcher et al., 1989) restandardised the test using more extensive norms and with Reconstructed Clinical Scales (Tellegen et al., 2003), which aimed to rectify the psychometric deficiencies of the scales. Most recently, the MMPI–2-RF (Tellegen & Ben-Porath, 2008) added further psychometric sophistication to the restructured clinical scales of the MMPI–2.

An important feature of the MMPI from the outset was the inclusion of what the authors called validity scales; that is, scales with the purpose of identifying whether the responses provided could be accepted at face value as predictors. The problems with self-report have been noted earlier (see Box 8.3). It is fairly

obvious in a situation where many of the items refer to events or feelings that are not normal that the respondent might 'fake'. There are now a dozen validity scales that can be used with the present version of the MMPI.

The MMPI is not the only test to be based on the empirical approach. Construction of the California Psychological Inventory (CPI; Gough, 1987) used the approach to develop a questionnaire for the assessment of normal personality functioning. The Strong Vocational Interest Blank (SVIB; Strong, 1959) was developed using the responses of workers in different occupations to assess suitability for different jobs (see Chapter 10). In both cases criterion groups were formed, and the differences among them in terms of item endorsement provided the basis for the different scales making up these instruments.

As well as providing a number of tests still widely used in personality assessment, the empirical approach fermented a long-standing controversy in assessment over the value of human judgment (see Box 8.4).

Box 8.4

Who makes better psychological predictions: clinician or statistician?

Paul Meehl was a great advocate for the use of the MMPI in personality assessment, but he was also a clinical psychologist, philosopher of science, and statistician. He became interested in the question of whether the exercise of clinical judgment (the subjective combination of data based on intuition and experience) leads to better predictions of socially relevant criteria (e.g. recidivism or therapy outcome) than the combination of data that relies on set, empirically based rules (statistical method). He reviewed the literature, such as it was, in the early 1950s and found that nineteen studies favoured the statistical method and only one favoured clinical judgment (Meehl, 1954), and the success of that study he qualified on further analysis. His conclusion was controversial and strongly rebutted by, among others, Holt (1958) who found several reasons to question it. Subsequent reviews by Sawyer (1966) and Grove and Meehl (1996) and meta-analyses by Grove et al. (2000) and Ægisdóttir et al. (2006) of a literature of now over 100 studies, however, point to the same conclusion: the statistical method is superior to clinical judgment when combining information to predict real-life outcomes such as likelihood of subsequent violent offending, suicide or doing well in a psychology training program.

If the outcome is so clear, why do clinicians persist in using an inferior method (Vrieze & Grove, 2009)? There are several possible reasons for this. Many judgments clinicians make do not lend themselves to analysis in terms of

strict criteria and many of those that do have not been researched sufficiently to allow the application of statistical method. Second, human memory is fallible and we are likely to remember our successes more than our failures unless there is immediate feedback on our decisions that cannot be ignored, which is seldom the case unless we arrange for it. And then there is what Dawes called 'cognitive conceit' (1976): it is hard to accept that humans cannot outperform a computer, particularly in a domain such as clinical psychology where interpersonal sensitivity is considered so important.

The lesson is clear. Training in psychology, even in clinical psychology, does not, as Meehl (1954) demonstrated, result in miraculous powers of human judgment; it's better to rely on formal prediction rules wherever possible.

The social-cognitive approach

Walter Mischel spent a good deal of time in his early years (Mischel, 1968) criticising the then dominant approaches of psychodynamic and trait theory for making assumptions about the person that could not be supported by the empirical evidence. His critique was particularly useful for trait theory, in clarifying its claims about personality. In his later years, Mischel (1973; Mischel & Shoda, 1995) has moved his theorising closer to a personality approach in proposing how cognitive processes can stamp behaviour in ways that make it individualistic. His colleague Albert Bandura (e.g. 1982, 1986) has not gone so far and his position may still be thought of as less like a mainstream approach to personality than the others discussed in this chapter. We chose to include them and their antecedents here in a discussion of personality because they represent a major approach to thinking about personality among English-speaking psychologists: the **social-cognitive approach**.

social-cognitive approach
an approach to personality that examines the relationships between people's behaviour, the situations in which these behaviours occur, and their cognitions about them

Mischel coined the term 'person variables' to characterise the consistencies in behaviour and thought that make for differences among individuals. He did not see these as expressions of motivational forces within the individual or as genetically determined dispositions but as resulting from the unique experiences of individuals. Table 8.2 presents his first list of person variables, which was modified slightly in later work (see, for example, Mischel & Shoda, 1995). It is tempting but wrong to think of these as quasi-traits because they are more

flexible than that, undergoing change with experience, with environmental demands, and with the results of cognitive processing. A person may lack a competency in formal writing, for example, and this may be blocking pursuit of an important goal for them (e.g. a career opportunity). They may therefore invest time and effort, and develop a competency sufficient for their purposes. Competencies do not have the same degree of fixity as the trait theorist's concept of mental abilities.

Table 8.2: The person variables identified by Walter Mischel

Mischel (1973)	Mischel & Shoda (1995)
Competencies	Competencies
Encoding strategies	Encodings
Expectancies	Expectancies and beliefs
Values	Affects, goals and values
Self-regulatory systems and plans	Self-regulatory plans

Encoding strategies—that is, ways of perceiving the world or processing information about it—may develop as a result of particular experiences, and once developed are not fixed for all time. George Kelly (1955) argued that the way we construe people or events can be changed through an active process.

Expectancies are even more significant in understanding people's decisions and actions. Originating in the learning theory of E C Tolman, the concept of expectancies has been used in a number of theories of motivation that together are classed under the title of expectancy-value theory (MacCorquodale & Meehl, 1954). For example, Julian Rotter (1954) argued that the choices people made were determined by the sum of the expectations they had about the outcomes of particular actions and the values they applied to those outcomes. Whether a student, for example, will spend a Sunday studying rather than skiing or surfing with friends depends on the outcomes they expect of these different activities and the importance they attach to them. A student who sees academic success as likely to result from concentrated study, and who sees such success as important, will choose to study if the sum of likelihood and importance for study is larger than the sum of the importance and the benefit of having fun with their friends. Other theorists would see the product rather than the sum of expectations and importance as the important determinant. (The essential difference is that with a product but not with a sum, if either expectation or value is zero, a failure of behaviour is the result.)

Bandura saw expectations—outcome expectations as he called them—as quite important but argued that there was an even more important set of expectations and these have to do with the expectations people have about the behaviours necessary to bring about the outcome. Behavioural expectations

relate to beliefs people have about whether they can perform the actions necessary to bring about a particular result. Thus a student may believe that regular study is instrumental in bringing about academic success and this may be valued as important, but the student may doubt that they can engage in regular study. Their doubts may be well founded, having spent several supposed study periods being distracted by different events. Without a sense of self-efficacy—Bandura's term for the belief that one can actually perform the required task—action will not occur. These behavioural expectations are important for the initiation of action and for its maintenance in the face of obstacles. People are likely to quit earlier if they believe that they are not really up to the task and find that progress is not being made. They are unlikely to choose an action in the first instance if self-efficacy is low. For example, if a career choice (say, engineering) is thought to require considerable competency in mathematics, the person lacking self-efficacy in mathematics is unlikely to include that as a career option.

Values are often thought of in terms of the amount of reward or reinforcement potential actions produce. As noted earlier, value combines with outcome expectancy to guide action. Although some types of reward have universal application (e.g. food when one is hungry, and money), adult social behaviour is guided by a wide range of rewards that vary considerably in their force from one individual to another.

Self-regulating systems and plans refer to the ways people learn to control their behaviour, and the strategies they employ and the goals they set in guiding their behaviour. A person may have learned that they are quick to anger and that they need to take steps to defuse it early, and as a consequence have developed ways of responding in anger-provoking situations. Because the goals we set ourselves are often difficult to achieve and lie sometime in the future, we need to plan to achieve them. The goals and strategies we adopt characterise us as much as any of these other person variables.

Assessment practices

Cervone, Shadel and Jencius (2001) sought to characterise a distinctly social-cognitive approach to personality assessment in terms of a set of principles or goals of assessment, rather than in terms of particular methods of assessment. Without considering these in detail, it is important to note the role attached to the situation in assessment in the social-cognitive approach. The extent to which a person's standing on any particular person variable generalises across situations is a matter for empirical inquiry. For example, for one person a sense of self-efficacy in speaking in a university tutorial may not generalise to a similar degree of self-efficacy in making a speech at a party, whereas for another person it may. Situations, the cognitive and affective responses they evoke, and the actions that

are initiated in them, vary from individual to individual and give rise to profiles across situations—or 'behavioural signatures' (Mischel & Shoda, 1995). The implication for assessment is that it must be directed to evaluating the behaviours and cognitive-affective experiences of individuals in the context of specific situations. Claes, Van Mechelen and Vertommen (2004) have attempted to translate the implication into a practical assessment task, in which the person's reports about their behaviour are used to construct scenarios for assessing their behavioural signature. The development cost in using the method is greater than that of the trait or empirical approaches to personality assessment and the payoff is as yet unclear.

Positive psychology

Maslow and Mittelmann (1941), in a then authoritative text on abnormal psychology, proposed that normal behaviour could be characterised in a number of ways. More interesting than the actual characteristics they listed was their attempt to describe normality as distinct from abnormality. It is abnormal behaviour that has been the major interest of personality and clinical psychologists, and a great deal has been written on the topic. Far less has been said about normal behaviour (Henry Murray and those who have followed his lead are an exception). The situation changed with the emergence of humanistic psychology and its heir, the **positive psychology** movement.

positive psychology
a relatively recent approach in psychology that stresses the behaviours, thoughts and feelings that characterise optimal functioning rather than dysfunction

Abraham Maslow, one of the founders of humanistic psychology, began work in experimental psychology but moved to the study of personality and abnormal psychology. Interest in the characteristics of career mentors whom he greatly admired led to the study of self-actualisation. Maslow proposed a pyramid of human motivations. At the base of the pyramid are physiological needs, above those are needs for security, higher still are needs for self-esteem, and at the highest point the need to actualise the self; that is, to fulfil one's potential to the maximum. Although the hierarchy has many critics, the concept of self-actualisation has attracted considerable research interest, particularly with the publication of a questionnaire to measure the construct (Shostrom, 1980).

What Maslow and his contemporary humanistic psychologists such as Carl Rogers were seeking to emphasise was that people exhibit a variety of positive characteristics, such as creativity, that warrant research in their own right. The

humanistic psychologists were keen to assert the positive side of human nature and sought an understanding of what made humans truly human. The self was, not surprisingly, the centre point of study.

It was the work of Martin Seligman that moved humanistic psychology into a new frame. In his presidential address to the American Psychological Association in 1998 (see Fowler, Seligman & Koocher, 1999) he coined the term 'positive psychology' to characterise the study of what was right with people rather than what was wrong. In a subsequent paper with Csikszentmihalyi, he defined the field:

> [A]t the subjective level [it] is about valued subjective experiences: well-being, contentment, and satisfaction (in the past); hope and optimism (for the future); and flow and happiness (in the present). At the individual level, it is about positive individual traits: the capacity for love and vocation, courage, interpersonal skill, aesthetic sensibility, perseverance, forgiveness, originality, future mindedness, spirituality, high talent, and wisdom. At the group level, it is about the civic virtues and the institutions that move individuals towards better citizenship: responsibility, nurturance, altruism, civility, moderation, tolerance, and work ethic. (Seligman & Csikszentmihalyi, 2000, p. 5)

Similarities and differences between humanistic and positive psychology were explored in an article by Waterman (2013) and a series of rejoinders to it in the *American Psychologist* (2014, 88–94).

Assessment practices

The agenda of positive psychology is certainly different from that of the classical theorists in personality, and already a good deal of research has been undertaken on some of the constructs Seligman and Csikszentmihalyi listed. The work on well-being and happiness is the most advanced (see, for example, Lucas & Diener, 2008), but there are also interesting programs underway on hope (e.g. Snyder, 1995: Synder et al. 2000) and gratitude (McCullough, Emmons & Tsang, 2002). A better understanding of constructs such as these will balance what we already know about anxiety, conflict and the like. What is not so clear is whether positive psychology will provide new methods for assessing personality in the way most of the approaches considered to this point have. One possibility in this regard is the refinement of the life history methods that Murray proposed. Dan McAdams (2008), for example, has proposed ways of studying the personal narratives of people to reveal identity, meaning and purpose.

Eclectic approaches

The early work in modern personality research was concerned with the building of 'grand' theories—theories that attempted to be all-encompassing in the phenomena they addressed. Freud's theory was the first and grandest of these grand theories, purporting to provide explanations for all manifestations of behaviour and personality. Difficulties in testing these grand theories, and the poor yield from attempts to do so, led to the theories being used more as storehouses for ideas or as a means of sensitising clinicians and researchers to what to look for or how to spot what was missing. In turn, the assessment methods that most of these theories gave rise to came to be used by practitioners who would not necessarily identify closely with a particular theory, but who saw merit in a particular method.

A 'mix and match' approach to personality theory and methods of personality assessment has developed into something of an orthodoxy, but it has its critics, who point to the theoretical incoherence of eclecticism. By knitting together bits from different theories, we run the risk of an explanation that has no internal consistency which, as such, is unlikely to help understanding 'real-world' phenomena. Looking for, say, an unconscious basis for self-efficacy is to seriously misunderstand the matrix of ideas from which each of these concepts is drawn and will not help the practical task of assessment. Caution is necessary in cherry picking personality theory.

That said, there have been sophisticated attempts to build approaches to personality from selections of concepts from the grander theories considered to this point. McCrae and Costa (1999) have proposed a theory of personality that is more extensive than their FFM. Their theory includes the five factors of personality as 'basic tendencies', but it also includes 'characteristic adaptations' to the social environment (of which the self-concept is a significant part), as well as the objective biography of the person; that is, the instances of behaviour that result from the operation of the basic tendencies and the characteristic adaptations. All of these are in turn influenced by the biological factors of genes and brain function, on the one hand, and cultural norms and situational demands on the other.

Costa and McCrae's five-factor theory shares much with McAdams and Pals' (2006) five-component framework. Component one refers to the basic design of the human system and the strengths and constraints that places on adaption to the circumstances of everyday life. Component five refers to the cultural context in which adaptations occur. Components two, three and four are the 'personality' components that are shaped to various extents by the biological and cultural systems in response to situational demands. Component two includes relatively enduring traits often thought of in terms of the FFM but not exclusively so. Component three includes the mental concerns and strategies that are

characteristic of the individual. Component four is the integrative life narrative that weaves together what has happened in the past with what we expect or would like to happen in the future. It is the story that brings meaning and purpose to life and confers an identity on us.

Although there is much in common between the two accounts, McAdams is wary of arguing for as close a link among their personality components as McCrae and Costa propose among the basic tendencies, characteristic concerns and objective biography. Traits may not influence mental concerns and, in turn, the life narrative in any necessary or systematic way according to McAdams and Pals. For the present, the matter is unresolved, but the broader ideas provide a useful way of summarising essential concepts and assessment methods from the approaches reviewed in the present chapter (see Table 8.3).

Table 8.3: Concepts and methods in personality assessment based on McAdams' possible levels of knowing another person

Level	Component	Concepts	Methods
1. Knowing at the level of the stranger	Dispositions or action tendencies	Cognitive, affective and behavioural resources Traits	Questionnaires Objective test
2. Intermediate knowing	Characteristic adaptations or mental concerns and strategies	Mechanisms of defence Needs and press Dynamisms Expectancies Values Self-regulating systems	Projective methods Self-report Ratings
3. Intimate knowing	Life narrative or objective biography	Psychosocial stages Proceedings and themes	Clinical interview Personal history

The first column is taken from an earlier paper by McAdams and attempts to characterise how much we can know of people from the different approaches to personality. Level 1 is the depth of knowing we have of people we have not previously met, but who are described to us or describe themselves relatively briefly. Information at this level can be useful but it is necessarily superficial. At the next level, termed intermediate by McAdams, we begin to understand something about what is important to the person: how they see the world, their

likes and dislikes. Beyond that is a level that McAdams describes as intimate to highlight the depth of understanding such as we think we have after working or living with someone for a considerable period of time. More recently McAdams (McAdams & Olson, 2010) has described the three levels as layers and placed them in the context of human development.

At each of the levels different concepts become relevant and different methods are used to capture them. At Level 1 we are dealing with traits captured by self-report or in the case of cognitive constructs by objective tests. At Level 2, we have a more fine-grained situational level of analysis in which concepts such as needs, defences and expectancies are used, with the approach determining the type of concept and the method of assessment considered appropriate. At Level 3, the personal or psychosocial history of the individual is the focus, and here the interview is used but again the way it is used differs with the approach.

Assessment of the 'total' personality, to the extent that such a venture is possible, involves all levels and the use of a number of methods of assessment, but in practical assessment situations this is seldom necessary. Selection for employment may be directed to the first level, and trait assessment may be sufficient in many instances. An exploration of the person's psychosocial history in such a context would be a waste of time and an intrusion of the individual's privacy. In some clinical situations, however, a more extensive exploration is called for at least at the level of characteristic adaptations and may call for assessment of the life history.

Chapter summary

This has been a brief review of personality theory as it bears on psychological testing and assessment. The approach taken is that of Wiggins, who identified five major paradigms or approaches to personality that have influenced assessment: the psychoanalytic, the interpersonal, the personological, the trait and the empirical. Each makes different assumptions about the characteristic ways people think, feel and behave, and each has generated a particular approach to assessment. To Wiggins' list we have added the social-cognitive approach and the positive psychology approach, both of which have influence in contemporary personality theory. Although neither has produced new techniques of assessment, each has brought a different and helpful perspective to the task of assessment. Social-cognitive theory stresses the important role of situational factors in the actions people take and is a useful corrective to the idea that personality is an unchanging feature of the individual. Positive psychology, on the other hand, draws attention to the optimal functioning person, as distinct from the dysfunctional personality that the psychoanalytic paradigm, for example, has stressed for much of its history. No one paradigm is the 'right one' in that each has certain weaknesses, but by using more than one in personality assessment it is possible to obtain a more rounded picture of the individual.

Questions

1. Is it possible to describe a person's personality without invoking a particular theoretical view of personality, at least implicitly?
2. Does the psychoanalytic approach warrant serious attention in the twenty-first century?
3. Given that we can see the origins of trait theory in the time before the modern era, why is it that the theory has survived for so long?
4. Is personality assessment just a matter of fitting people into categories?
5. Positive psychology is a 'nice idea', but has it contributed to our understanding of personality?
6. What concepts from positive psychology would you add to Table 8.3 and how would you assess them?
7. What are response sets and why might they be important?
8. What is the 'Barnum effect' in personality assessment?
9. Psychologists are trained to have a special insight into human personality. Discuss.
10. What is a 'behavioural signature' and how would you assess it?

Further reading

Boyle, G J, Matthews, G & Sakloske, D H (Eds.). (2008). *The Sage handbook of personality theory and assessment, Vol 2. Personality measurement and testing.* Thousand Oaks, CA: Sage.

John, O P, Robins, R W & Pervin, L A (2008). *Handbook of personality* (3rd ed.) New York, NY: Guilford.

Monte, C F & Sollod, R N (2003). *Beneath the mask* (7th ed.). New York, NY: Wiley.

Wiggins, J S (2003). *Paradigms of personality assessment.* New York, NY: Guilford.

Useful websites

All about personality (Psychology Today):
www.psychologytoday.com/basics/personality

Personality theory and research (The Personality Project): http://personality-project.org/personality.html

PART 4 AREAS OF PROFESSIONAL APPLICATION

9 Clinical and Mental Health Testing and Assessment

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. define a referral question and explain why sometimes it is necessary to clarify a referral question
2. identify what information needs to be collected during history taking
3. explain the nature, purpose and steps of a clinical interview
4. describe what is a mental status examination and what are the areas covered by such an examination
5. name the most commonly used psychological tests for assessing intelligence, personality, psychopathology, depression and anxiety, and discuss their strengths and weaknesses
6. describe the purpose, structure and main components of a psychological report.

KEY TERMS

clinical interview
Depression Anxiety and Stress Scale
Diagnostic and Statistical Manual of Mental Disorders
mental status examination
Minnesota Multiphasic Personality Inventory
Personality Assessment Inventory
psychological report
referral question
Wechsler Adult Intelligence Scale

Setting the scene

- Since her divorce three months ago, a 30-year-old woman has been feeling very sad and has lost interest in activities she normally enjoys. She was referred by her family doctor to a clinical psychologist for assessment of depression.
- A 60-year-old man who has been drinking heavily for the past 25 years was referred to a rehabilitative service for assessment and treatment of alcohol abuse.
- Since witnessing a bank robbery, a young bank teller has not been able to return to work. She has been anxious and agitated, and has nightmares. The bank referred her to a clinical psychologist for assessment and counselling.
- A young man in his early twenties has been acting strangely over the last two months. He reported that he was being unfairly treated by his boss and workmates, and he also reported hearing voices. He was admitted to a hospital for psychiatric assessment and treatment.
- Partners who have been married for 10 years are having difficulties maintaining their relationship. They referred themselves to a clinical psychologist to seek help.

Introduction

In Australia and other parts of the world, mental health services, public and private, are one of the largest employers of psychologists. This is not surprising, given that in 2007 a national survey found that 20 per cent of our population aged between 16 and 85 years had a mental disorder in the 12 months prior to the survey (Slade et al., 2009). These disorders were identified as one of the leading causes of healthy years of life lost due to disability and their annual cost in Australia has been estimated at $20 billion (Australian Bureau of Statistics, 2009–10). Clinical psychologists in this setting assess, diagnose and treat mental disorders (e.g. schizophrenia, depression, anxiety and personality disorders) as well as problems in everyday living (e.g. relationship problems, low self-esteem and stress). In all cases, the starting point for the psychologist is usually the **referral question**, which may be as broad as: Is the client suffering from a mental disorder? What is the likely cause of the client's problem? What is the client's current level of psychological functioning? What is the appropriate treatment for a client and how should the treatment be evaluated? In this chapter, we introduce the psychological assessment techniques most commonly used by clinical psychologists in the mental health setting. These techniques include history taking, clinical interview, mental status examination and psychological testing. For the psychological tests, we concentrate on some of the commonly used tests for intelligence, personality, psychopathology, depression, anxiety and stress. To conclude the chapter, we discuss the content and structure of a psychological assessment report and provide an example of such a report.

referral question
a request for psychological testing or assessment is usually raised by a client or other professionals who work with the client; it can be general or specific

Clarifying the referral question

In the mental health setting, the need for psychological testing and assessment for a client is usually triggered by a referral question. This question provides the justification or rationale for testing and assessment (Suhr, 2015). If the client is referred by another professional (e.g. a psychiatrist or general practitioner), the referral question will have been formulated by them. If the client is self-referred, there is a need to formulate the problem to be addressed. In either case, there may be a need to spend some time clarifying or refining the referral question so that it becomes realistic or answerable in terms of what current knowledge in psychology can provide (Groth-Marnat & Wright, 2016; Maloney & Ward, 1976). The question might be too broad (e.g. Why does my daughter have an eating disorder?) or generate expectations that cannot be met (e.g. Please assess and treat this client's depression in three sessions) and there might need to be a negotiation of the expected outcome with the referring agent. The formulation of a clear and specific referral question will facilitate the derivation of hypotheses about a case, selection of appropriate psychological assessment instruments, interpretation of results, and provision of recommendations. This process can be facilitated by the use of a standard referral form with explicit questions about the reason for referral, use of assessment results and the client's willingness to undertake the assessment (Bagby, Wild & Turner, 2003).

Case history data

After clarifying or agreeing on the referral question for a client, a clinical psychologist who works in the mental health setting usually begins a case by collecting demographical and biographical data about the client. These data are useful for providing the context in which to understand the referral question, interpreting results of other data collection procedures, making recommendations and preparing the psychological report. Although most of the data can be obtained during a clinical interview with the client, sometimes it is useful to collect them from a number of sources for verification purposes. For example, for clients who lack self-awareness or for those with memory or language problems, it might not be possible to find out details of their educational or vocational history by asking them direct questions. Family members or partners, in these cases, may be a better and more accurate source

for such information. In most mental health settings, standardised forms have been designed to summarise these demographic and biographical data. Having access to these forms facilitates the collection of information. An important consideration here is the need to be aware of and familiar with the privacy policies of various organisations (e.g. hospitals, private companies, non-governmental organisations and government departments) and the legal requirements (e.g. the *Freedom of Information Act 1982*) and ethical guidelines for obtaining and using information of this sort.

Clinical interview

The **clinical interview** is one of the oldest psychological assessment techniques used to collect information about a client or a patient, and the most widely used by clinical psychologists who work in a mental health setting (Hersen & Thomas, 2007). It can be unstructured, structured or semi-structured. Basically, during the interview the psychologist will ask the client a number of questions (both open- and closed-ended) that are related to the client and to the referral question. Sometimes questions are used to elicit information that is not readily available from the client's record or file. For example, although there may be some information on educational history, the interviewer may need to ask the client directly about the level of educational achievement or favourite subjects in school. Similarly, information on a client's file may indicate that a client is married, but to gauge marital satisfaction, the interviewer will need to ask about the duration and quality of the marital relationship. At other times, questions are used to test a hypothesis that the psychologist has formulated about the client's condition. For example, if the psychologist suspects that the client is suffering from a depressive disorder, questions about the person's recent level of activities, sleeping and eating habits, ability to concentrate, and prevailing mood become pertinent.

clinical interview
a technique for collecting information about a client; it may take many forms, for example, from a psychoanalytic perspective it includes detailed exploration of the personal and family history of the client, particularly with respect to psychosocial development, conflict, and defence, self and interpersonal processes

The clinical interview also provides the psychologist with a good opportunity to establish rapport with the client, to provide important information, and to establish whether the client has a reasonable understanding of what is happening to them and why. If the psychologist considers that the client does not feel comfortable during the initial stage of the clinical interview, she might want to

spend more time putting the client at ease before asking the more confronting questions. Information the psychologist can convey during the interview includes:

1. the purpose and nature of psychological testing and assessment
2. what the client or patient is expected to do
3. confidentiality of information collected during assessment
4. the need for informed consent (the client or patient consents to testing after being made aware, in language that can be understood, of the nature and purpose of testing)
5. who will have access to the information collected and how it will be used.

To conduct a successful clinical interview, the psychologist needs to establish good rapport with the client by being sincere and supportive (Giordano, 1997). To engage the client in the interview, a number of techniques can be used. These include: trying not to dominate the interview, reflecting what is said, paraphrasing, summarising, clarifying, confronting, using eye contact and a positive posture, and nodding (Groth-Marnat & Wright, 2016; Maloney & Ward, 1976).

Although most of the information collected by the psychologist during a clinical interview is verbal in nature (i.e. answers to questions), non-verbal information is provided by the client's demeanour during the interview, by how particular questions are answered, and at times by what is *not* said. For example, a matter-of-fact or flippant style of responding may be inconsistent with the seriousness of the content being revealed. This could be useful information for interpreting test results or answering the referral question later.

Clinical psychologists who work in the mental health setting typically obtain the following information during a clinical interview:

- demographic data
- medical history (self and family)
- family history
- educational and vocational history
- psychological history.

Although much of this information is of the sort that would be obtained by psychologists working in other settings (e.g. organisational or educational), an important additional source of information comes from the mental status examination (discussed in the next section), which is unique to the mental health setting (Bagby, Wild & Turner, 2003). Finally, psychologists who work in this

setting sometimes use structured clinical interview schedules such as the Structured Clinical Interview for DMS Disorders (SCID-5-CV; First et al., 2016) to ensure relevant information relating to various disorders are adequately covered and asked.

Mental status examination

Similar to the physical examination conducted by a medical doctor, the **mental status examination** is a comprehensive set of questions and observations used by a clinical psychologist or by other professionals in a mental health setting to systematically assess the mental state of a client. These questions include the following:

mental status examination
a comprehensive set of questions and observations used by psychologists to gauge the mental state of a client, which usually covers areas such as appearance, behaviour, orientation, memory, sensorium, affect, mood, thought content and thought process, intellectual resources, insight and judgment

- *Appearance:* How does the client look? What kind of clothing does the client wear? Is the clothing appropriate for the occasion or the weather? What is the personal hygiene of the client?
- *Behaviour:* How does the client behave during the examination? Does the client show unusual verbal and non-verbal behaviour?
- *Orientation:* Is the client aware of who or where he is? Does the client know what time (year, month, date, day and time) it is?
- *Memory:* Does the client show any problems in immediate, recent and remote memory?
- *Sensorium:* Is the client able to attend and concentrate during the examination? Does the client show problems in hearing, vision, touch or smell?
- *Affect:* Does the client display a range of emotions during the examination? What are these emotions and how appropriate are they?
- *Mood:* What is the general or prevailing emotion displayed by the client during the examination?
- *Thought content and thought process:* What does the client want to focus on during the interview? Does the client only want to talk about these things? Is the client able to clearly explain ideas during the interview? Does the client

show problems such as talking rapidly, jumping from one topic to another, being circumspect and tangential, or using illogical reasoning and arguments?

- *Intellectual resources:* Does the client have good verbal ability? Can the client answer questions that call for general information or arithmetical operations?
- *Insight:* Is the client aware that there is a problem? Does the client know what is causing the problem? Does the client know the reason for the referral to see a mental health professional?
- *Judgment:* Does the client have the ability to make their own decisions? Can the client make plans and solve problems?

Based on information gained during the clinical interview and mental status examination, the psychologist can begin to formulate or conceptualise the client's problem by referring to systematic classification systems such as the *Diagnostic and Statistical Manual of Mental Disorders* published by the American Psychiatric Association (see Box 9.1) or the *International Classification of Diseases* published by the World Health Organization (1992–94). To further clarify ideas and narrow down or test hypotheses, the psychologist may administer psychological tests to finalise the assessment.

Box 9.1

Diagnostic and Statistical Manual of Mental Disorders

The **Diagnostic and Statistical Manual of Mental Disorders** (DSM) is a standard classification system of mental disorders published by the American Psychiatric Association for use by mental health professionals. It is the most commonly used system adopted by professionals in the USA, Australasia and Asia. The main purpose of the DSM is to facilitate communication among mental health professionals. The diagnostic terms and codes included in the manual provide a shorthand for professionals to communicate information about clients and their conditions. Because the diagnostic system of the DSM is based on observed behavioural symptoms rather than on a particular theoretical perspective, it can be used by professionals with different theoretical orientations.

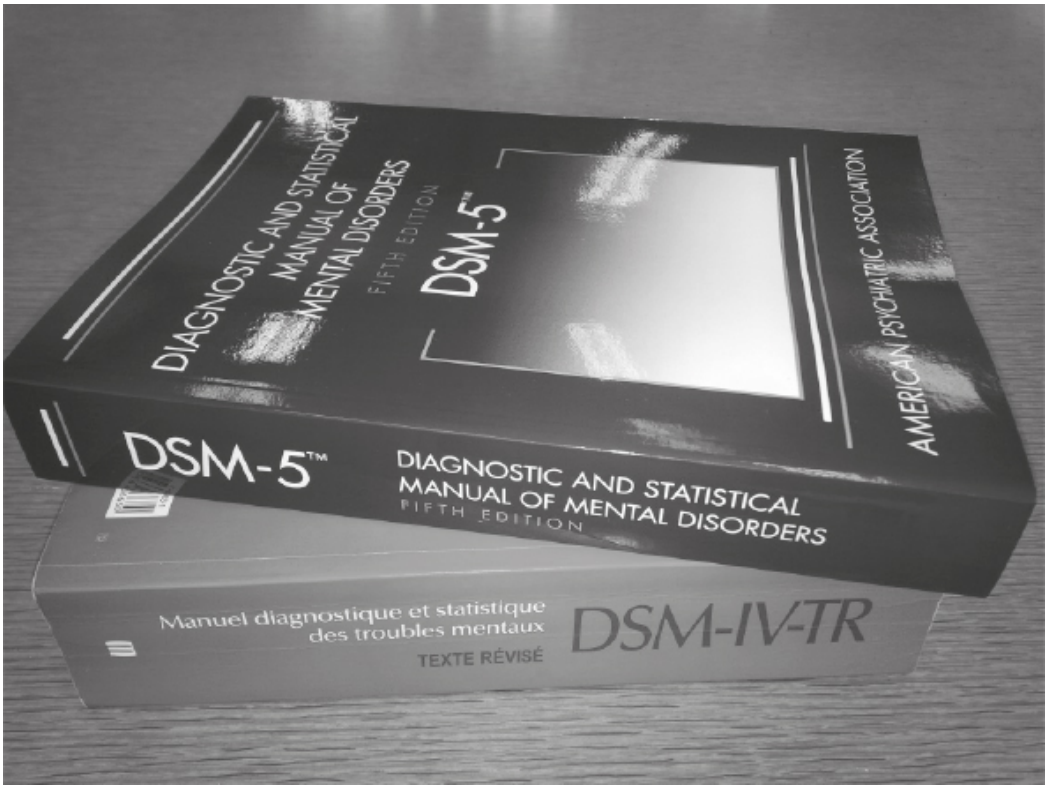
Diagnostic and Statistical Manual of Mental Disorders (DSM)
a standard classification system of mental disorders published by the

American Psychiatric Association for professionals to use to diagnose mental disorders

The first edition, DSM–I, was published in 1952 and the latest edition, DSM–5 (which replaced the DSM–IV–Text Revision (TR)), was published in 2013. Like the DSM–IV–TR, the DSM–5 contains a list of psychiatric disorders and their corresponding diagnostic codes. Each disorder is accompanied by a set of diagnostic criteria and text containing information about the particular disorder, including associated features, prevalence, familial patterns, age-, culture- and gender-specific features, and differential diagnoses. No information about treatment or ætiology is included. In addition, each client is not just given a single label.

One notable change from the DSM–IV–TR to the DSM–5 is the elimination of the multiaxial system of diagnosis (Axis I, II, etc.). The DSM–5 is now divided into three sections: Section I describes all the changes from the DSM–IV–TR to the DSM–5; Section II lists all the disorders with separate notations for important psychosocial and contextual factors (formerly Axis IV) and disability (formerly Axis V); and Section III is a new section that describes disorders which require further study but are not included in the main lists of disorders in Section II. The Global Assessment Functioning scale (GAF, Axis V) is now replaced with the WHO Disability Assessment Schedule (WHODAS) for measuring global functioning.

Figure 9.1 Diagnostic and Statistical Manual of Mental Disorders



The disorders in the DSM–5 are listed in such a way that it reflects a lifespan approach (e.g. developmental disorders being listed at the beginning of the section and disorders more applicable to later adulthood being listed at the end of the section). Lastly, the DSM–5 uses Arabic rather than Roman numerals for each edition. Updates for the DSM will now be classified using decimal numbers (e.g. DSM–5.1). Although the DSM–5 has attracted some criticisms (e.g. diagnostic inflation, inadequate empirical documentation and inadequate field trials; Frances & Widiger, 2012), the DSM manuals are still commonly used by professionals in preventing, diagnosing and treating mental health problems (Nathan & Langenbucher, 2003; Dziegielewski, 2015).

Psychological tests

Because of space limitation, we will confine our description of tests to a select number of instruments commonly used in the clinical and mental health area for the testing and assessment of intelligence, personality, psychopathology, depression, anxiety and stress. Our selection was based mainly on the tests listed by the Psychology Board of Australia in the national psychology examination curriculum. Interested readers can consult sources such as Goldstein and Hersen (2000), Groth-Marnat and Wright (2016) and Hersen (2004) for a more comprehensive treatment of instruments used in the mental health setting.

Intelligence

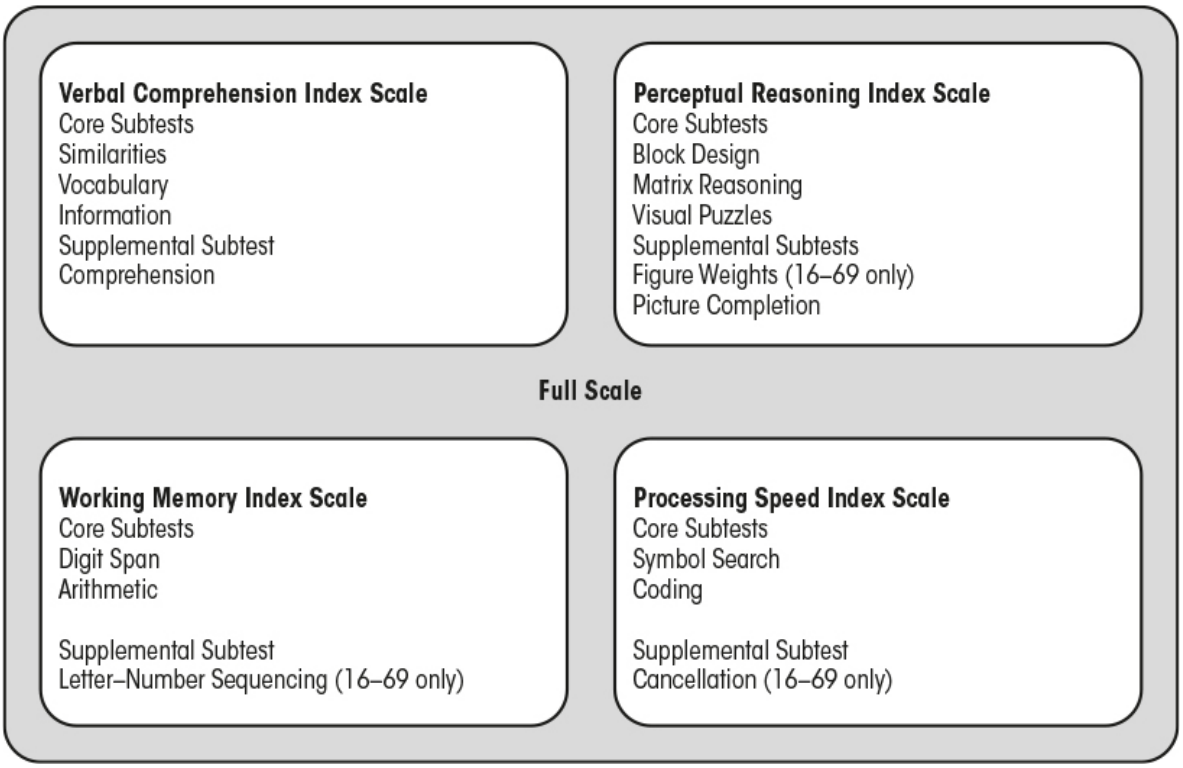
Since Binet's pioneering development of ways of assessing intelligence in children, psychologists in a number of settings have made use of measures of general intellectual ability. In Chapter 7, we covered the history, theories, issues and controversies of intelligence testing. Here, we describe one of the mostly commonly used individual tests of intelligence in clinical practice.

Wechsler Adult Intelligence Scale–Fourth Edition

The **Wechsler Adult Intelligence Scale** (WAIS) is a classic test and one of the most commonly used psychological tests throughout the world (Archer et al., 2006; Camara, Nathan & Puente, 2000; Rabin, Barr & Burton, 2005). The original version was published as the Wechsler-Bellevue Intelligence Scale (WB) in 1939. Other editions of this test include the WAIS (Wechsler, 1955), the WAIS–Revised (WAIS–R; Wechsler, 1981) and the WAIS–Third Edition (WAIS–III; Wechsler, 1997a). Developed for adults aged between 16 and 90 years old, the WAIS–IV was published in 2008 and, similar to its predecessors, its aim is to assess intellectual ability in adults. It is also used for assessing psychoeducational disability, neuropsychiatric and organic dysfunction, and giftedness. The main purposes of this revision are to update the norms, co-norm with the Wechsler Memory Scale–Fourth Edition (WMS–IV) and the Wechsler Individual Achievement Test–Second Edition (WIAT–II), reduce testing time, and improve psychometric properties.

Wechsler Adult Intelligence Scale (WAIS)
developed by David Wechsler, and one of the most widely used, individually administered, intellectual assessment batteries; the latest version, WAIS–IV, was published in 2008

Figure 9.2 Structure of the Wechsler Adult Intelligence Scale–Fourth Edition



Wechsler (2008)

The WAIS–IV is an individually administered test battery that comprises ten core subtests and five supplementary subtests (see Figure 9.2). In this version, two subtests in the WAIS–III (Picture Arrangement and Object Assembly) were dropped and three new subtests (Visual Puzzles, Figure Weights and Cancellation) added. In addition, in this version, new items were added to the subtests of the older versions and modifications made to the administration, recording and scoring of the subtests. According to the manual, the WAIS–IV (ten core subtests) takes on average about 67 minutes to administer. Table 9.1 lists all the subtests of the WAIS–IV and the abilities they measure. Core subtests are the ones needed to be administered to derive composite scores, and supplemental subtests are the ones that can be administered to assess other cognitive skills to provide additional clinical information. Where necessary, supplemental subtests can also be used to substitute the core subtests to derive composite scores. Altogether, five composite scores can be derived based on performances on the core subtests. They are Full Scale IQ, Verbal Comprehension, Perceptual Reasoning, Working Memory and Processing Speed (Figure 9.2 illustrates which core subtests are used to derive which composite scores). The traditional Verbal IQ and Performance IQ are replaced by the Verbal Comprehension Index and the Perceptual Reasoning Index. In addition, an optional composite score called General Ability Index can be derived from the three Verbal Comprehension and three Perceptual Reasoning core subtests.

Table 9.1: Subtests of the WAIS–IV

Subtest	Timed	Description	Abilities measured
Verbal - comprehension subtests			
Similarities	No	Test taker is provided with pairs of words that represent objects or concepts and has to describe why they are similar	Verbal concept formation and reasoning
Vocabulary	No	Test taker is required to name pictures and to provide meaning of words of increasing difficulty	Word knowledge and verbal concept formation
Information	No	Test taker is required to answer a number of general knowledge questions	Ability to acquire, retain and retrieve general factual knowledge
Comprehension (Supplemental)	No	Test taker is asked to answer questions based on understanding of general principles and social situations	Verbal reasoning and conceptualisation, verbal comprehension and expression, ability to evaluate and use past experience, and ability to demonstrate practical knowledge and judgment
Perceptual - reasoning subtests			
Block Design	Yes	Test taker is asked to arrange red and white coloured blocks to recreate designs, presented models or pictures	Ability to analyse and synthesise abstract visual stimuli

Subtest	Timed	Description	Abilities measured
Verbal - comprehension subtests			
Matrix Reasoning	No	Test taker is shown incomplete matrices or series and is asked to choose a response option that best completes the matrices or series	Classification and spatial ability, knowledge of part–whole relationships and perceptual organisation
Visual Puzzles	Yes	Test taker is required to view a completed puzzle and select three response options that when combined will reconstruct the puzzle	Non-verbal reasoning and ability to analyse and synthesise abstract visual stimuli
Figure Weights (Supplemental)	Yes	Test taker is required to view a scale with missing weight(s) and to select the response option that keeps the scale balanced	Quantitative and analogical reasoning
Picture Completion (Supplemental)	Yes	Test taker is shown pictures with important parts missing and asked to identify what is missing for each picture	Visual perception and organisation, concentration and visual recognition of essential details
Working memory subtests			

Subtest	Timed	Description	Abilities measured
Verbal - comprehension subtests			
Digit Span	No	Test taker is presented with series of randomised digits at one digit per second and asked to repeat them as given, in reverse order, and in ascending order	Auditory processing, attention, mental manipulation and working memory
Arithmetic	Yes	Test taker is presented with mathematical problems orally and asked to solve them mentally	Attention, mental manipulation, numerical reasoning and working memory
Letter-Number Sequencing (Supplemental)	No	Test taker is orally presented with a series of letters and numbers that are random in order and asked to repeat the numbers and letters separately but in order	Attention, mental manipulation, sequential processing and working memory
Processing speed subtest			
Symbol Search	Yes	Test taker is asked to scan and search for a target symbol among a group of symbols	Visual-motor processing speed, coordination and attention
Coding	Yes	Test taker is asked to use a key to copy symbols that are paired with numbers	Processing speed, learning ability, psychomotor speed, visual-motor coordination and visual scanning

Subtest	Timed	Description	Abilities measured
Verbal - comprehension subtests			
Cancellation (Supplemental)	Yes	Test taker is asked to scan a structured arrangement of shapes and to mark target shapes	Speed of processing, attention, perceptual speed and visual-motor ability

One of the strengths of the WAIS–IV is the size and representativeness of the standardisation sample used in test development. A total of 2200 individuals were included, ranging in age from 16 years 0 months to 90 years 11 months across thirteen age groups. Each of the nine younger age groups of the sample comprised 200 individuals and each of the four older age groups comprised 100 individuals. A stratified sampling plan was used to match the final sample as closely as possible to the population of the USA in 2005 in terms of demographic characteristics known to influence intelligence scores: namely, age, gender, race/ethnicity, educational attainment (self or parent) and geographical location. The 2005 US Census was used to provide the test developers with accurate information on the proportions of individuals in each of the demographic groupings, and these proportions were reproduced in selecting individuals for the standardisation sample.

Scoring the test involves two steps. The raw scores on the subtests are converted to scaled scores (with a mean of 10 and a standard deviation of 3) based on the appropriate age group of the standardisation sample. The subtest scores are then summed and transformed into a Full Scale IQ and four Composite Score Indices (Verbal Comprehension, Perceptual Reasoning, Working Memory and Processing Speed). The IQ and Score Indices all have a mean of 100 and standard deviation of 15. Results of the Composite Score Indices can be used for discrepancy comparisons; that is, deciding whether a client's abilities in these four areas are significantly different from each other. In addition, strengths and weaknesses analysis can also be conducted on scaled scores of the subtests.

Table 9.2 summarises the internal consistency and test-retest reliabilities of the subtests, Indices and FSIQ score of the WAIS–IV. The test-retest reliabilities were obtained based on a subgroup (298 adults) of the standardisation sample and the retest period ranged from eight to 82 days (average = 22 days). The manual also reports high inter-scorer agreement on some subtests that require the exercise of judgment in scoring (*r* for Similarities, Vocabulary, Information

and Comprehension = 0.93, 0.95, 0.97 and 0.91, respectively). This, together with the coefficients summarised in Table 9.2, indicates that the WAIS–IV has high reliability.

The test manual of the WAIS–IV also presents an impressive amount of evidence to support its validity. First, it has been shown that the WAIS–IV can discriminate between those with and without neurological, psychoeducational and developmental disorders. Second, results of exploratory and confirmatory factor analyses support the four-index model. Third, the WAIS–IV has been found to correlate significantly with other tests of intellectual ability (e.g. WAIS–III, WISC–III and WMS–III).

Table 9.2: Reliability of the WAIS–IV

Scores			Reliability
		Internal consistency	Test-retest
IQ	Full Scale	0.98	0.96
Indices			
	Verbal Comprehension	0.96	0.96
	Perceptual Reasoning	0.95	0.87
	Working Memory	0.94	0.88
	Processing Speed	0.90	0.87
Subtests			
	Similarities	0.87	0.87
	Vocabulary	0.94	0.89
	Information	0.93	0.90
	Comprehension	0.87	0.86
	Block Design	0.87	0.80
	Matrix Reasoning	0.90	0.74
	Visual Puzzles	0.89	0.74
	Figure Weights	0.90	0.77
	Picture Completion	0.84	0.77

Scores			Reliability
	Digit Span	0.86	0.82
	Arithmetic	0.93	0.83
	Letter-Number Sequencing	0.88	0.80
	Symbol Search	–	0.81
	Coding	–	0.86
	Cancellation	–	0.78

Note: split-half/coefficient alpha reliability for Symbol Search, Coding and Cancellation could not be calculated because of the nature of these processing speed subtests.

Despite its advantages and excellent psychometric properties, the WAIS–IV is not without limitations. These include, for example, the relatively long time needed to administer the core subtests required to obtain the necessary IQ or Index scores, failure to take into consideration recent advances in intelligence theories and failure to include new subtests to assess recently emerging concepts in the area of intelligence such as social intelligence and creativity. For users of WAIS–IV in Australia and New Zealand, although adaptations (e.g. using more familiar stimuli, relevant cultural content and appropriate language for test items) have been made to ensure examinees in these two countries were not disadvantaged by US content, it is still the case that the norms used for scoring and interpretation were collected in the USA. Thus, research evidence is needed to show that this is appropriate (see the box 'Using American norms with Australian populations' in Chapter 3).

Personality

The assessment of personality has been an area of some controversy in psychology in the past (e.g. Mischel, 1968) because of concerns about the validity of many of the tests developed for this purpose. There is now less concern on this point, for three main reasons. First, the accumulation of a very large number of individual studies using a technique termed meta-analysis has pointed to validity coefficients for personality tests that are modest in size but replicable and useful for assessment purposes. Second, factor analytic work with personality tests has helped to clarify the similarities and differences between them. Third, there are now much more realistic expectations about what information these tests can provide and about their limitations. Personality measures do not provide highly

specific predictions about what individuals will do; rather, they provide information about what people are generally like or what they usually do.

In Chapter 8, we reviewed a number of different systems or theories of personality and these lead to different approaches in practice to personality assessment. The choice of a system for assessment depends partly on the theoretical orientation of the assessor and partly on the referral question. Because there are a large number of personality theories currently discussed in the literature, space does not permit an extensive treatment of various approaches. Instead, we confine ourselves to one of the most widely used tests for this purpose.

Minnesota Multiphasic Personality Inventory–Second Edition

The **Minnesota Multiphasic Personality Inventory**–Second Edition (MMPI–2; Butcher et al., 1989) is a 567-item self-report inventory and one of the most commonly used psychological assessment instruments in the USA, Australia and New Zealand (Camara, Nathan & Puente, 2000; Watkins et al., 1995). In 2008, Tellegen and Ben-Porath published a shorter (338 items) alternative to the MMPI–2 called the MMPI–2 Restructured Form (MMPI–2 RF). The original MMPI was developed to measure major patterns of personality and emotional disorders in adults 18 years and older, using a technique called the criterion-keying approach (for details, see Chapter 8). In criterion keying, test items are selected from a pool of items if responses to them discriminate between a group presumed to show the characteristic of interest and a group who do not. In the development of the original MMPI, patient groups previously diagnosed by a panel of expert psychiatrists were compared with groups of visitors to a large hospital. Items that differentiated between, for example, a group of patients diagnosed with schizophrenia and a group of 'normals' (i.e. hospital visitors) were included in the schizophrenia scale. Content of the item is not important in criterion keying. The only consideration is the empirically demonstrated capacity of the item to discriminate.

Minnesota Multiphasic Personality Inventory (MMPI)
a test developed to assess major patterns of personality and emotional disorders using the empirical-keying approach; the latest version, MMPI–2 was published in 1989 and it requires a test taker to respond to 567 items and takes 60 to 90 minutes to complete

The MMPI–2 can be administered individually or to a group in 60 to 90 minutes. Test takers are asked to consider each of the items of the inventory and indicate whether the statements are 'true' or 'false' for them. The responses of the test takers can be hand- or computer-scored and T scores for ten validity

indicators (Cannot Say (?), Variable Response Inconsistency (VRIN), True Response Inconsistency (TRIN), Infrequency (F), Back F (FB), Infrequency-Psychopathology (FP), Symptom Validity Scale (FBS), Lie (L), Correction (K) and Superlative Self-Presentation (S) and ten clinical scales (1 Hypochondriasis, 2 Depression, 3 Hysteria, 4 Psychopathic Deviate, 5 Masculinity–Femininity, 6 Paranoia, 7 Psychasthenia, 8 Schizophrenia, 9 Hypomania and 10 Social Introversion) (see Figure 9.3). In addition, a number of supplementary scales (e.g. Anxiety, Repression, Ego Strength and Social Responsibility) can be obtained. If the T scores of the validity scales are elevated (e.g. >65 or 1.5 SD above the mean), corrections can be added to the clinical scales and care is needed in interpreting the results and profile. For the content scales, elevation of the T score of a particular scale (e.g. >65 or 1.5 SD above the mean) suggests problems or difficulties in that area. Clinicians are also provided with interpretation guidelines and suggestions about the combined effect of elevation of two scales, or what is termed the two-point code. More detailed interpretation procedures for the MMPI–2 can be found in the test manual and in Groth-Marnat & Wright (2016).

The standardisation sample of the MMPI–2 comprised 2600 non-clinical individuals and 423 individuals with psychiatric problems. The internal consistency of the MMPI–2 scales is typically in the 0.70s and 0.80s, but some coefficient alphas as low as 0.30 have been reported for some scales in some samples. In terms of test-retest reliability, correlation coefficients ranging from 0.50 to 0.90 have been reported for retesting after one week.

Figure 9.3 Minnesota Multiphasic Personality Inventory–2: Profile for Validity and Clinical Scales

The validity of the MMPI–2 has been supported by high correlations between scores on the second and the first editions. In the literature, a large amount of research has been conducted examining the validity of the MMP–I, and Graham

(1993) reported an average validity coefficient of 0.46. The MMPI–2 is a sensitive instrument, but it should be pointed out that the scales are highly correlated and the test is not based on a firm theoretical base. In addition, it has not been revised based on a recent classification of psychopathology (e.g. the DSM–5).

Psychopathology

Unlike the WAIS–IV or the MMPI–2, several instruments have been developed to provide a comprehensive assessment of mental health problems. One of the main advantages of this type of instrument is a systematic and comprehensive coverage of all major areas of potential problems. In this section, we review two examples of these instruments: one designed for adults and the other for adolescents.

Personality Assessment Inventory

Originally developed in 1991, the **Personality Assessment Inventory** (PAI; Morey, 2007) is a 344-item self-report scale designed to provide information relating to clinical diagnosis, treatment planning and screening for psychopathology in adults 18 years and older. It can be administered individually or in a group and usually takes 40 to 50 minutes to complete. The items of the test were written at a fourth grade reading level and test takers are asked to consider each of the 344 items and endorse each one of them according to a four-point scale: False (Not At All True), Slightly True, Mainly True and Very True. Some examples of the PAI items include: 'My health condition has restricted my activities'; 'Often I think and talk so quickly that other people cannot follow my train of thought'; and 'I have some ideas that others think are strange'.

Personality Assessment Inventory (PAI)
a 344-item self-report scale designed to collect information relating to clinical diagnosis, treatment planning and screening for psychopathology in adults

Hand or computer scoring can be used, and results are summarised in T-scores and plotted on a multi-sided profile form. There are altogether twenty-two non-overlapping scales: four validity scales (Inconsistency, Infrequency, Negative Impression and Positive Impression), eleven clinical scales (Somatic Complaints, Anxiety, Anxiety-Related Disorders, Depression, Mania, Paranoia, Schizophrenia, Borderline Features, Antisocial Features, Alcohol Problems and Drug Problems), five treatment consideration scales (Aggression, Suicide Ideation, Stress, Nonsupport and Treatment Rejection) and two interpersonal

Scales (Dominance and Warmth). Among the 344 PAI items, twenty-seven are regarded as critical items. These are items that have very low endorsement by individuals in the normal sample and are indicative of potential crisis situations. The standardisation sample of the PAI comprised a census-matched sample (n = 1000; stratified according to age, gender and race), a clinical sample (n = 1265 from a number of clinical sites) and a university student sample (n = 1051). Scores can be compared to means and standard deviations for the subsamples based on gender, race and age groups (i.e. 18–29, 30–49, 50–59 and 60+ years old).

The internal consistency of the PAI has been found to range from 0.70 to 0.80. The average test-retest reliability of the full scale of the inventory over 24 to 28 days was 0.76. In terms of validity, the four validity scales have been found to correlate significantly with the validity scales of the MMPI–2 and the Crowne-Marlowe Social Desirability Scale (Crowne & Marlowe, 1960). In addition, validity of the PAI-2 has been demonstrated with convergent and discriminant validity with other measures of psychopathology and comparisons between criterion and control groups.

Millon Adolescent Clinical Inventory

The Millon Adolescent Clinical Inventory (MACI; Millon, 1993) is a 160-item self-report inventory that was designed as a brief instrument that can be administered individually in 20 to 30 minutes to assess a range of personality patterns and clinical presenting problems in adolescents aged 13 to 19 years old. The MACI can be used across diverse treatment settings for adolescents (e.g. inpatient, school, correctional and residential placement settings), and has been developed specifically to assess the unique difficulties that young people may experience during adolescence. The MACI can be used to obtain information to inform diagnostic hypotheses, create individualised treatment plans, and assess progress before and after treatment. The developers of the MACI assert that the measure was constructed based on theoretical considerations of personality structure (i.e. the Grossman Personality Facets) and with reference to multi-axial diagnostic systems (i.e. DSM; McCann, 1999). There is a growing research base on the use of the MACI across varied adolescent populations and treatment and assessment contexts, including incarcerated adolescent offenders (Salekin, 2002), adolescent substance abuse (Grilo et al., 1996), adolescents with child abuse histories (Grilo et al., 1999) and hospitalised adolescents with depression (Hiatt & Cornell, 1999). A significant strength of the MACI is that it was specifically developed for adolescents to address the unique difficulties they face, rather than being adapted from adult instruments (McCann, 1999).

Each item in the inventory represents a statement that test takers respond to on a true/false scale. Items fall into twenty-seven scales that are organised into

three clinically relevant categories, including twelve Personality Patterns (Introversive, Inhibited, Doleful, Submissive, Dramatising, Egoistic, Unruly, Forceful, Conforming, Oppositional, Self-Demeaning and Borderline Tendency), eight Expressed Concerns (Identity Diffusion, Self-Devaluation, Body Disapproval, Sexual Discomfort, Peer Insecurity, Social Insensitivity, Family Discord and Childhood Abuse) and seven Clinical Syndromes (Eating Dysfunctions, Substance Abuse Proneness, Delinquent Predisposition, Impulsive Propensity, Anxious Feelings, Depressive Affect and Suicidal Tendency). Additionally, the measure contains three modifying/validity indices (Disclosure, Desirability and Debasement) to assist in identifying test-taking attitudes, as well as confused or random responding. The Personality Patterns were designed to parallel those described in the DSM–III, DSM–III–R and DSM–IV. The Expressed Concerns scales focus on feelings and attitudes about problems faced by troubled adolescences, and the Clinical Syndromes scales assess disorders typically observed in adolescent populations.

Responses on the MACI can be hand or computer scored and summarised as Base Rate (BR) scores, where each scale score fits in a scale of 1–115, with 60 being the median score. A BR score of 75 on a scale indicates the presence of a pattern, while a BR score of 85 indicates the prominence of a pattern. The standardisation sample consists of 1017 adolescents who were seen in clinical treatment settings for emotional and/or social problems in North America (Millon, 1993). The MACI manual reports internal consistency for scales ranging from 0.73 to 0.91 across the standardisation sample. Further, the test-retest reliability for the scales ranged from 0.57 to 0.92, with the median stability coefficient being 0.82. The validity of MACI scales scores has been examined using a variety of statistics, including cross-validation correlations between scale scores and clinician judgments, and correlations between scale scores and existing collateral instruments. Correlations between clinician judgments and scale scores have been found to be modest across standardisation subsamples, with the highest correlations being for the Clinical Syndromes category of scales (correlations ranging from 0.09 to 0.52). MACI scale scores were correlated with scores from collateral instruments purported to measure similar constructs (e.g. Beck Depression, Hopelessness and Anxiety Inventories, Eating Disorder Inventory-2 and the Problem Oriented Screening Instrument for Teenagers). Correlations between MACI scales and concurrent collateral scales were found to be quite high.

Depression and anxiety

As mentioned in the beginning of this chapter, mental health disorders were identified as one of the leading causes of healthy years of life lost due to disability

and their annual costs in Australia has been estimated at $20 billion (Australian Bureau of Statistics, 2009–10). Among these disorders, depression and anxiety are the two that contribute most to this burden. It is therefore not surprising that these are the two most commonly referred problems in mental health settings in Australia. It is estimated that about 20 per cent of adults will experience a major depressive episode in their lives (Hassed, 2000). Those suffering from depression experience feelings of intense sadness for a considerable time; those suffering from anxiety are affected frequently by a state of severe and distressing nervousness. Although the symptoms of depression and anxiety are different, both of these conditions, if left untreated, can lead to debilitating and life-threatening consequences. In this section, we discuss a number of commonly used tests of depression and anxiety.

Beck Depression Inventory–Second Edition

The Beck Depression Inventory (BDI) is a commonly used scale for assessing depression, and the second edition (BDI–II; Beck, Steer & Brown, 1996) is a major revision of the BDI. The test was developed for individuals aged 13 to 80 years to assess symptoms corresponding to criteria for diagnosing depressive disorders based on the DSM–IV. The BDI–II can be self-administered or administered verbally by a trained administrator and it takes about 5 minutes to complete. Test takers are asked to use a four-point scale (0–3) to indicate whether they are experiencing depressive symptoms and their intensity. A total score can be obtained by hand or by using computer software. The standardisation sample of the BDI–II included a group of 500 outpatients and a group of 120 university students.

The internal consistency of the BDI–II as reported to date is high (0.92 for the clinical sample and 0.93 for the non-clinical sample) and its test-retest reliability is 0.93 for a one-week retesting period. In terms of validity, scores on the BDI–II have been found to correlate significantly and substantially with other measures of depression (e.g. the Hamilton Psychiatric Rating Scale for Depression–Revised, and the Symptom Check List-90–Revised Depression subscale). In addition, it has been found to discriminate between individuals who suffer from clinical depression and those who do not. Results of factor analyses also provide support for the validity of this inventory.

Beck Anxiety Inventory

The Beck Anxiety Inventory (BAI; Beck & Steer, 1987) is a twenty-one-item self-report inventory designed to measure the presence and extent of anxiety in adults and adolescents. It takes only 5 to 10 minutes to complete and can be self-administered or administered verbally by a trained administrator. Test takers are

asked to indicate how much they have been bothered by the symptoms listed during the past week using a four-point (0–3) scale that ranges from 'Not at all' to 'Severely; I could barely stand it'. The total score for the BAI is simply obtained by summing the points endorsed by the test takers on each of the twenty-one items, which can be done by hand or using computer software. The norms for the inventory are based on 810 outpatients with a variety of diagnoses.

The reported internal consistency of the BAI is high, ranging between 0.85 and 0.94. The test-retest reliability of the inventory was 0.75 over one week. In terms of validity, Beck and Steer have provided evidence to support its content, concurrent, construct, discriminant and factorial validity.

State Trait Anxiety Inventory

The State Trait Anxiety Inventory (STAI; Spielberger, 1983) is a self-report questionnaire of current symptoms of anxiety (viz., state anxiety [S-Anxiety]) and propensity to anxiety (viz., trait anxiety [T-Anxiety]). This test has been used in thousands of studies, in many different languages and many different countries (Spielberger & Reheiser, 2009). There is both an adult's version and a children's version (STAIC) of the test. The STAI consists of 40 items–20 items for each S-Anxiety and T-Anxiety subscales. The questionnaire is generally administered via pencil and paper and takes around 10 minutes to complete. Higher scores on each subtest indicates greater anxiety (scores are reversed for anxiety-absent items). Test-retest reliability has been found to be quite high on the T-Anxiety scale, with coefficients ranging from 0.73 to 0.86 (Spielberger, 1983). The S-Anxiety scale, however, yielded a lower test-retest reliability than the T-Anxiety scale (0.33). But this is expected due to the transitory nature of the S-Anxiety subscale (Spielberger & Reheiser, 2009). Concurrent validity of the T-Anxiety scale with other measures of anxiety (e.g., Taylor Manifest Anxiety Scale Cattell and Scheier's Anxiety Scale Questionnaire) have been found to be high, with coefficients ranging from 0.73 to 0.85 (Spielberger, 1983).

Kessler Psychological Distress Scale

The Kessler Psychological Distress Scale (K–10) is a ten-item questionnaire that was developed for screening populations on non-specific psychological distress (Kessler et al., 2002). The K–10 scale requires respondents to answer a number of questions relating to anxiety, depressive or physical symptoms experienced in the past 30 days (e.g. 'In the past 30 days, how often did you feel nervous?'). Responses are recorded using a five-point Likert scale (None of the time, A little of the time, Some of the time, Most of the time, or All of the time). The responses are then added up to yield a total score for the K–10 scale, with a maximum score

(i.e. 50) indicating severe distress and a minimum score (i.e. 10) indicating minimal distress.

The K–10 scale has been included in a number of population health surveys in Australia, such as the National Mental Health Survey (conducted by the Australian Bureau of Statistics in 1997; Andrews & Slade, 2001), and a population-wide survey exploring the relationship between social capital and mental health morbidity (Phongsavan et al., 2006). Normative data was produced from the National Mental Health Survey, with the K–10 being validated against clinical diagnoses of anxiety and affective disorders in the Australian population (Andrews & Slade, 2001).

The Hospital Anxiety and Depression Scale

The Hospital Anxiety and Depression Scale (HADS) was developed by Zigmond and Snaith (1983). This fourteen-item test was designed to provide clinicians with a quick and reliable method for identifying anxiety and depression disorders among hospital patients. The HADS is divided into two subscales: an Anxiety subscale (HADS-A; seven items) and a Depression subscale (HADS-B; seven items). The anxiety and depression questions are interspersed within the questionnaire. Patients have to indicate how they have felt in the last week and respond to each item on a four-point response scale (not at all–all the time [0–3]). For both scales, scores within 8–10 indicate a mild case, 11–14 a moderate case, and 15–21 a severe case of a mood disorder (Stern, 2014). Both the HADS subscales have been found to have high sensitivity and specificity (0.80) and found to correlate highly with other measures of anxiety and depression (Bjelland et al., 2002). The HADS has also been validated in many different countries and languages (e.g. Herrmann, 1997).

Depression Anxiety and Stress Scales

The **Depression Anxiety and Stress Scale** (DASS; Lovibond & Lovibond, 1995a) is a forty-two-item self-report scale designed to measure the states of depression, anxiety and stress (fourteen items for each state) for individuals over 17 years of age. There is also a short version that consists of twenty-one items. It was developed in Australia and is popular here and overseas, with the scale being translated into about thirty languages (Antony et al., 1998; Brown et al., 1997; Crawford & Henry, 2003; Wang et al., 2016). The DASS is available in the public domain and can be administered individually or in groups, and takes 10 to 15 minutes to complete. Sample items for the three scales are as shown in Table 9.3.

Depression Anxiety and Stress Scale (DASS)
a 42-item self-report scale that aims to measure the state of depression, anxiety

and stress in adults over the previous week

Test takers are asked to use a four-point severity–frequency scale to rate the extent to which they have experienced the state referred to in each of the forty-two items of the DASS over the past week. The total scores for the three scales can be easily obtained by using a template and they can be compared with the mean total scores of a standardisation sample of 2914 non-clinical individuals (note that 1607 of these were university students) or to suggested cut-offs derived from this sample. Based on this standardisation sample, the internal consistencies for the three scales of the DASS have been found to be high (Cronbach's alpha = 0.91, 0.84 and 0.90, respectively). Similar values of alpha have been obtained in a sample of clinically diagnosed individuals (Antony et al., 1998). Test-retest reliabilities (retest period = two weeks) for the three scales have also been found to be adequate (r = 0.71, 0.79 and 0.81, respectively; Brown et al., 1997). In terms of validity, the three-factor structure of the DASS has been supported by results of both exploratory and confirmatory factor analyses, and its convergent and discriminant validity have been demonstrated by correlations with the BDI and the BAI (Crawford & Henry, 2003; Lovibond & Lovibond, 1995b; Wang et al., 2016). The DASS has been found to be sensitive in discriminating individuals with clinical problems from those not so diagnosed (Lovibond & Lovibond, 1995a). In 2011, Crawford et al. published percentile norms and accompanying interval estimates for an Australian general adult population sample (n = 497) for the DASS.

Table 9.3: Sample items of the DASS

Scale	Items
Depression	I felt sad and depressed.
	I felt that I had lost interest in just about everything.
Anxiety	I experienced trembling (e.g. in the hand).
	I felt I was close to panic.
Stress	I found myself getting upset by quite trivial things.
	I found it hard to wind down.

Psychological report

Once all relevant information about the client has been gathered using the particular tests chosen for the purpose, results need to be brought together to answer the referral question. This is usually done in the form of a written report that has a commonly agreed format (Ownby, 1997; Zuckerman, 2005; see Case study 9.1 for a sample report). A **psychological report** is important because it allows the referral agent and others to understand why and how the psychologist came to the particular conclusions and why particular suggestions are being made. A written report, compared with a verbal report, provides an enduring record.

psychological report
a report to provide a client or a referral agent with the answer(s) to the referral questions based on results of testing and assessment; it is usually provided in a written format that has a commonly agreed structure

The following headings and content are typically included in a psychological assessment report in the clinical and mental health setting:

1. *Demographic data:* This includes name, gender, address, age, date and place of birth, marital status, ethnic background (if applicable), name of psychologist and date of psychological testing/assessment session.
2. *Relevant background:* A client's or patient's family, educational, vocational, psychological and medical history are usually included in this section. It is important to include only the information that is relevant to the current referral question, otherwise this section may become too long and cluttered with trivial or irrelevant information.
3. *Previous assessment:* If the client or patient has been seen by another psychologist previously for a similar or related problem, it is necessary to briefly summarise the results of the previous psychological assessment. This will provide the reader of the report with an idea of what the functioning of the client or patient was like previously and allow the psychologist to compare the results of the two assessments (if similar techniques were used).
4. *Assessment techniques and date and duration of assessment:* In this section, the names and order of the psychological assessment techniques used should be listed chronologically. This will give the reader some ideas about the length of the psychological assessment and the number and types of techniques used to answer the referral questions. For referral agents who might not be familiar with the names and purposes of the psychological assessment techniques, it is useful to provide a one-line

description of the purpose of the tests used. For example: 'The MMPI–2 was administered on 25 Jan 2005. This test measures major patterns of personality and emotional disorders in adults 18 years and older.'

5. *Results and interpretation:* The results obtained using the various psychological assessment techniques are summarised and explained in this section. For tests that have a large number of scores and scales, it is easier to use a table to summarise and present the results. Score ranges (rather than exact scores) are sometimes used to indicate the margin of error (e.g. plus or minus one or two SEM) associated with the estimate (see the reporting of the WAIS–IV results in Case study 9.1). Apart from describing the results obtained, the psychologist will also need to interpret what the results mean. Further, these results should be interpreted within the context of the background information described earlier in the report. For example, it is easier to interpret why someone is showing a high score on the Beck Depression Inventory if it has been reported that there is a history of depression in the family and that a number of events (e.g. losing a job or a relationship break-up) have recently happened in the client's or patient's life.
6. *Recommendations:* Based on the findings of assessment, recommendations for further action are usually offered. These may be in terms of what can be done to assist the person to deal with the problem, such as suggestions for a certain number of sessions of treatment or therapy. They could also be suggestions for psycho-education for both the client and his or her significant others. Sometimes the recommendations could be for further assessment or for reassessment after a given time.
7. *Summary:* This is the final section of a psychological assessment report and it is a precis of all the previous sections. Although this is the last section of the report, it is often the first section a referral agent reads. Therefore, it needs to be factually accurate, clearly written, and consistent with the information included and discussed in the other sections.

From the observations of Shellenberger (1982) and Brenner (2003), a good report:

- is individualised rather than general
- answers the referral question directly
- focuses on and describes behaviour
- is written in a clear, precise and straightforward manner without jargon
- is written and delivered on time
- emphasises strengths of clients

- provides explicit, specific and implementable suggestions and recommendations.

It is good practice to seek an opportunity to explain and clarify the report rather than simply send it to the client or the referral source. This can be accomplished in a face-to-face session or by a telephone call. Some follow-up may be needed to ensure that recommendations are implemented and that they are working well (Brenner, 2003; Wise, 1989). Sometimes the client initiates the follow-up because progress is not being made, but systematic follow-up helps ensure a positive outcome.

Case study 9.1

Example of a psychological assessment report

This is a fictitious case developed to illustrate the content of a psychological report. As such, it is not meant to be comprehensive or in-depth.

Client's name:	John Smith	File number:	135782
Date of birth:	17/07/1968	Age:	39
Date of Initial Session:	03/01/2008	Date of Final Session:	10/01/2008
Number of Sessions:	2		

Referral information and presenting problem

John presented to the outpatient mental health clinic for psychological assessment. John works as an accountant for a large car company. He reported that he has always been very organised and efficient both at work and at home. However, he stated that over the last six months he has found it increasingly difficult to cope with work demands. He described having difficulty getting to work on time and meeting work deadlines, and making more mistakes with routine tasks. John reported being concerned by the level of difficulty he was having sustaining his attention and concentrating at work. He stated that he found it difficult to remember names, addresses and other information unless he wrote them down, and found himself easily distracted at work. John also reported experiencing periods of insomnia, lack of appetite, and low energy and mood. He reported first experiencing these difficulties about six months ago; at

about the same time his wife Sally separated from him, and he reported these symptoms had become worse since this time. John stated that he felt there was something wrong with his brain and wanted to have a cognitive assessment performed to ensure that he was not, as he put it, 'losing his mind'.

Sources of information

Clinical Interview: 03/01/08
Beck Depression Inventory–Second Edition (BDI–II): 03/01/08
Beck Anxiety Inventory (BAI): 03/01/08
Depression Anxiety Stress Scales (DASS): 03/01/08
Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV): 10/01/08

Mental status examination

John is of medium height and slightly underweight. He was dressed in a dishevelled manner in a T-shirt and jeans. His nutritional condition appeared to be poor; his skin was pale and his hair was matted. He appeared to be very tired; he sat slumped back in his seat and gazed at the floor. However, he was cooperative throughout the interview. John's speech was mumbled and slow. His responses were non-spontaneous and minimal; he appeared to struggle to find the words to express himself. He described his mood as 'hopeless', which was consistent with his depressed affect. John reported that he had experienced thoughts about suicide, but he did not have a specific plan to carry out this behaviour, nor did he believe it would be likely for him to do so. John was oriented in time, place and person. His concentration and memory recall appeared to be impaired.

History of presenting problem

When John was 9 years of age his mother died. John lived with his father until he completed his university degree in accounting in his mid-twenties. It was at this time he married Sally. John and Sally have two children, a 15-year-old boy and a 12-year-old girl. John reported that he and Sally have had a difficult relationship, but in the last five years it had deteriorated considerably. Approximately six months ago, John's wife Sally separated from him, taking the children. John is not aware of having any major medical or psychological problems during his childhood or currently, and has not previously undertaken any psychological assessment or treatment. He reported that, prior to the last month, he has always been very organised at work and has been surprised by the level of difficulty he is currently experiencing with remembering things and with

maintaining his concentration and attention. John also described the onset of his insomnia and low mood and energy levels as sudden and 'out of the blue'.

Assessment and results

The WAIS–IV is a test used to assess general thinking and reasoning skills. The following scores show how well John performed compared to other people in his age group.

Indexes	Score range	Percentile	Classification
Full Scale IQ	111–19	84th	High Average
Verbal Comprehension	112–23	88th	High Average
Working Memory	109–23	87th	High Average
Perceptual Reasoning	110–22	87th	High Average
Processing Speed	84–101	30th	Average

John's Full Scale IQ places him in the High Average range of intellectual functioning, achieving a score above that of approximately 84 per cent of his peers. The Verbal Comprehension Index (VCI) provides a measure of acquired verbal knowledge and verbal reasoning. John's VCI score exceeds that of 88 per cent of his peers and falls within the High Average range. The Working Memory Index (WMI) assesses an individual's ability to attend to verbally presented information, to process information in memory, and then to formulate a response. John's performance on the subtests requiring working memory is in the High Average range; he performed better than 87 per cent of his age-mates.

The Perceptual Reasoning Index (PRI) is a measure of non-verbal reasoning and concept formation. The PRI measures fluid reasoning, spatial processing, attention to detail and visual-motor integration. John's PRI score fell within the High Average range; he performed better than approximately 87 per cent of his same-aged peers.

The Processing Speed Index (PSI) provides a measure of an individual's ability to process simple or routine visual information quickly and efficiently, and to quickly perform tasks based on that information. John's PSI score fell within the Average range, although towards the low end, with his score better than approximately 30 per cent of his same-aged peers. John's relatively low performance on this index compared to his scores on the other indices could be interpreted as being the result of psychomotor slowing due to his depressive

symptoms. Depression has been found to be associated with slowed mental processing and attentional deficits.

Self-report measures of depression and anxiety

John was administered the Beck Depression Inventory (BDI), the Beck Anxiety Inventory (BAI) and the Depression Anxiety Stress Scales (DASS) to assess his level of depression, anxiety and stress. John's results from these self-report measures are provided in the table below. As can be seen in the table, John exhibited levels of depression, anxiety and stress in the severe range.

Tests	Date completed	Score	Norms range
BDI	03/01/08 (Session 1)	32	Severe Depression
BAI	03/01/08 (Session 1)	41	Severe Anxiety
DASS	03/01/08 (Session 1)	Depression 40 Anxiety 17 Stress 34	Severe Depression Severe Anxiety Severe Stress

Summary and recommendations

- John is a 39-year-old male who presented to the mental health outpatient clinic for the assessment of his cognitive and psychological functioning. During the intake session John described having difficulty sustaining his concentration and attention at work and reported dysphoric mood and insomnia.
- John's Full Scale IQ is in the High Average range (84th percentile). His index scores also fell in the High Average range except for his Processing Speed Index score, which is in the Average range (32nd percentile).
- John's BDI, BAI and DASS scores indicated that at intake he was experiencing a severe level of depression, anxiety symptoms and stress.
- My considered opinion is that John's presenting concerns—his difficulties with maintaining concentration and attention, insomnia and dysphoric mood—are the result of depression and there is currently no evidence of cognitive impairment.
- I recommended that John receive psychotherapy for his depressive and anxiety symptoms. After receiving such treatment, John's difficulties with maintaining concentration and attention should abate. However, if

this does not occur I recommended that John be referred for further psychological assessment.

Susan Brown

Psychologist

Discussion questions

1. After reading this chapter, what other psychological tests or assessment instruments would you use for this client? Give reasons for your answer.
2. What are the strengths and weaknesses of this psychological report? Give reasons for your answer.
3. How would you improve this psychological report?

Practitioner profile

Professor Amanda Gordon

1. **How long have you been a psychologist?**
   Following four years as a research psychologist, I have worked as a clinician for 31 years.

2. **What is your specialisation and how did you get the training and experience to do this job?**
   I have specialist endorsement in two areas—clinical and health psychology. My early years in research were in the health psychology area of chronic pain. I worked in a pain clinic in a psychiatric unit, then in a neurosurgical hospital, and finally had experience in an orthopaedic-based clinic. I was also involved in chronic disease research and psychological interventions with families where chronic disease or disability dominated.

   Subsequent to my research, I became involved in treatment of those with chronic pain conditions. I was then employed in a psychiatric unit and was trained more intensively in clinical psychology at that time. I was trained in the use of psychological and neuropsychological tests and the interpretation of the results by administering and reporting on them. There was an excellent and skilled team of clinical psychologists with whom I worked, who supported and mentored me throughout my hospital and early private practice. Throughout my training I have had mentors across disciplines—psychologists, psychiatrists and other medical professionals—who have all honed my skills.

3. **What kind of clients and referrals do you usually get?**
   My current client load is quite diverse, ranging from adolescents to those of later years. Many of my clients have diagnosable mental illnesses, particularly depression or the range of anxiety disorders. I also manage those with grief and bereavement—often around suicide or sudden death—and relationship crises are common presentations. I also deal with those with addictions, as well as other behavioural disorders.

4. **Do you use psychological tests in your practice?**
   I administer some simple screening tools to every new client before their first consultation, and on the third, sixth and tenth sessions. I also use a personality screening tool for a specific cohort to assess for potential risk factors.

I no longer administer neuropsychological tests, as that work has become very specialised and I do not feel equipped to adequately interpret my findings. I may from time to time do intelligence or basic skills testing, for educational purposes or recruitment, and use questionnaires that allow parents and teachers to give feedback about children.

5. Why do you use psychological tests and in what way do they help you in your practice?

The standard tests I use for therapy clients are useful in terms of noting change over time, as well as pointing out potential vulnerabilities and markers. In all other cases, I use tests to answer a specific question I may have that is best answered through standardised material, which will assist in clinical assessment.

Often clients are reassured that test results confirm clinical findings, or even assist me in asking different questions, and I am able to put them in the helpful category.

6. In your opinion, what is the future for psychological testing in your specialisation?

Clients of the twenty-first century are often quite psychologically sophisticated and may even have googled tests and self-administered them prior to coming in to the practice. So the value for me for the therapy cohort is as back-up for what I am doing clinically, to allow people to express themselves and their distress to me over time, and to confirm that a problem-focus isn't masking other issues. It is often useful to use a particular test as a population screen and note diversions from the mean, and then have a specialist do more formal testing on that small at-risk population.

I believe that psychological testing is more and more in the realm of specialists in particular areas, much as pathology and radiology are for medicine. The private practising psychologist working in health and mental illness will use selected tests as back-up for their clinical judgment, but will send their client for a full testing session with a specialist if there is a concern.

Chapter summary

The clinical and mental health setting is one of the main areas where psychologists conduct testing and assessment. In this chapter we discussed the main techniques for assessing mental health problems. After clarifying a referral question with a client or a referral agent, psychologists in this area usually use the clinical interview and the mental status examination to collect relevant information to assist them to develop hypotheses about the case. In addition, they have access to a large number of psychological tests to assess constructs such as intelligence, personality, psychopathology, depression and anxiety, and stress. Testing and assessment usually conclude with the completion of a written report that has the purpose of answering the referral question and a commonly agreed format.

Questions

1. What are the main functions of a clinical interview?

2. What is the purpose of a mental status examination? What are the main areas covered in this examination?
3. Briefly describe the purpose and content of the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV).
4. What evidence has been collected to support the validity of the WAIS–IV?
5. Compare and contrast the MMPI–2 and the DASS.
6. Briefly describe the purpose and content of the DASS.
7. What are the characteristics of a 'good' psychological assessment report?

Further reading

Goldstein, G & Hersen, M (2000). *Handbook of psychological assessment* (3rd ed.). New York, NY: Pergamon.

Groth-Marnat, G & Wright, A J (2016). *Handbook of psychological assessment* (6th ed.). New York, NY: Wiley.

Ownby, R L (1997). *Psychological reports: A guide to report writing in professional psychology* (3rd ed.). New York, NY: Wiley.

Sellbom, M, Marion, B E & Bagby, R M (2013). Psychological assessment in adult mental health settings. In J R Graham & J A Naglieri (Eds.), *Handbook of psychology: Vol 10, Assessment psychology* (pp. 241–60). Hoboken, NJ: John Wiley & Sons.

Slade, T, Johnston, A, Oakley Browne, M A, Andrews, G & Whiteford, H (2009). 2007 national survey of mental health and wellbeing: Methods and key findings. *Australian and New Zealand Journal of Psychiatry*, *43*, 594–605.

Useful websites

Depression Anxiety Stress Scales (DASS): www2.psy.unsw.edu.au/DASS

DSM–5 implementation and support: www.dsm5.org/Pages/Default.aspx

Structured clinical interview for DSM Disorders (SCID): www.scid5.org

10 Organisational Testing and Assessment

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. suggest different ways of measuring job performance
2. cite the advantages and disadvantages of different job performance measures
3. suggest possible predictors of job performance for particular occupations
4. discuss the assessment of work attitudes in employees
5. explain vocational personalities
6. understand how interest inventories work.

KEY TERMS

behaviourally anchored rating scales
graphic rating scale
integrity test
job analysis
KSAOs
performance appraisal
person–organisation fit
RIASEC
selection on the criterion
validity generalisation

Setting the scene

- A manager with a job vacancy to fill has identified the applicant characteristics desirable for the position. How do managers identify these desirable characteristics, how do they decide which candidate is the best one from those who apply, and, once an appointment has been made, how do managers tell if the decision to appoint was a good one?
- Part of a manager's responsibility is to identify which employees are performing well and who might need further training. How do managers evaluate their employees?
- University graduates might attend many job interviews when they complete their studies. What sort of questions can they be expected to answer at these job interviews?
- Experienced workers with a long history with a company can be asked to sit psychological tests of various types, and sometimes the results of these tests are used to determine who stays and who is asked to leave the organisation. Would it be fair and reasonable to retrench employees based on this use of psychological tests?
- Young people often struggle to decide on a career path and thus find it difficult to choose the right subjects at high school or the best training for them after school. How useful would it be for a young person to complete a vocational interests test?

Introduction

Industrial and organisational (I/O) psychology is one of the oldest fields of applied psychology. The importance of psychological issues in the workplace was recognised well over a hundred years ago. Among the earliest published works were Walter Scott's (1908) analysis of how to use psychological principles to improve advertising success, and Hugo Munsterberg's (1913) general text that focused on enhancing industrial efficiency. The contribution of psychology to the military during the First World War provided a great impetus to applied psychological testing. In Britain, the Industrial Fatigue Research Board, whose aim was the 'scientific study of the laws governing the healthy employment of the human mind and body in industry' (MRCDSC, 1920, p. 4), was created in 1918, and Charles Myers established the National Institute of Industrial Psychology in London at about the same time to apply the scientific method to increase industrial efficiency and improve working conditions (Welch, Welch & Myers, 1932). By 1919, the British Psychological Society had established an industrial psychology section. Across the Atlantic, the Association of Consulting Psychologists was formed in the USA in 1930 and included a number of I/O psychologists among its members. Later, in 1937, the American Association of Applied Psychology (AAAP) was formed. This association included a section dedicated to industrial and business psychology. The AAAP eventually merged with a number of other groups to form the American Psychological Association

(APA) in 1945. Division 14 of APA, the Society for Industrial and Organisational Psychology, is now one of the largest groupings of I/O psychologists in the world. In Australia, the organisational psychology division of the Australian Psychological Society was established in 1971, finally becoming the College of Organisational Psychologists in 1993.

industrial and organisational (I-O) psychology
the study of job performance and worker health issues to assist individuals, groups and organisations

Although psychological testing and assessment are core components of I/O psychology, the field is concerned with all aspects of human behaviour in the workplace. I/O psychologists attempt to improve organisational productivity and worker performance, as well as enhance the quality of working life in general. Major areas of application include work motivation, designing and redesigning jobs, recruitment and selection of new personnel, training and development of workers, managing individual and group performance, and facilitating organisational change processes. Psychological assessment and the evaluation of outcomes through tests and questionnaires figure strongly in all of these activities. Unlike other areas of applied psychology that deal with relatively tiny clinical populations, I/O psychologists are concerned with the vast majority of normal people who go to work.

While tests might be used to evaluate the effectiveness of any organisational intervention, there are two areas of application where principles of psychological assessment play a central role: the assessment of workers' performance on the job, known as **performance appraisal**; and the prediction of that performance, usually prior to appointment, for staff selection purposes. Reasons for the first are fairly apparent: the productivity of any organisation rests on the performance of its employees; hence, managers are always interested in the effectiveness of individual workers. Explanations for the second are also clear: all attempts to improve the performance of workers once they join an organisation—through further training, incentive schemes, redesigning jobs, introducing new technology, etc.—are dependent, often to a large extent, on starting out by selecting appropriate recruits. A highly skilled, high-quality workforce, identified by valid personnel selection programs, is likely to have positive repercussions throughout the organisation for many years. Indeed, a high-quality workforce is now widely recognised as one of the key ways of competing in the post-industrial age (Handy, 1994). New technology, plant and equipment can always be purchased, but an organisation's workforce—its human capital—is unique and cannot be easily duplicated by the competition.

performance appraisal
the assessment of a worker's job performance, typically carried out on a regular basis, such as six-monthly or annually

Employees, of course, are not passive participants in this process of selection, evaluation and development. While organisations select desirable applicants and then take action to help them be productive and satisfied (i.e. by appraising performance and implementing developmental strategies when needed), employees themselves are also focused, first, on ensuring they enter workplaces that are a good fit for their abilities and personality, and, second, that they do well in those organisations. Working in a job that is a good fit for personal characteristics has important payoffs for the individual in respect to satisfaction and career success (Kristof-Brown, Zimmerman & Johnson, 2005).

Right from the earliest years, children understand that adults work, and they set about playing these roles, which helps them clarify their interests, abilities, and values in relation to employment (Gottfredson, 2005). As a young child you might have wanted to emulate some of Australia's great tennis players, and practised tennis in your backyard; doing this would have informed you about how well you played tennis and whether it was something you enjoyed. In later years, as young adults and adults, individuals seek to apply for, 'fit' into, and succeed in organisations that meet their personal interests and abilities. This '**person–organisation fit**', which occurs when there is compatibility between the person and the organisation, benefits both the organisation and the individual (Kristoff, 1996).

Psychological testing has been central to assisting individuals to clarify their interests, values and personal characteristics. This testing has allowed people to identify the occupations that might suit them, and has helped them choose the courses and training needed to seek these occupations. It is very likely that you have undertaken some sort of vocational interest or ability test to help you understand which occupations, and thus university courses, might suit your strengths. Two names synonymous with vocational interest testing related to person–organisation fit are John Holland (1919–2008) and Edward Strong (1884–1963), who we will learn more about towards the end of this chapter.

Performance appraisal

The most valid assessment of the job performance of employees has always rested on the application of psychological assessment principles. For many years it was assumed that good indicators of job performance were available, but while this belief was fairly true for jobs located at either end of the production and distribution process (see Table 10.1), it was not true for most jobs in between. At

the manufacturing end, reasonably good indicators of job performance can be found in the form of simple productivity counts like the number of items produced, the amount of scrap material left behind, or the number of defective parts or errors. At the distribution end, the number of products sold or the dollar value of sales can serve as useful indicators.

Table 10.1: Some performance indicators

Type of indicator	Example
Quantitative production counts	Number of items produced Number of defects Number of products sold Dollar value of sales
Qualitative production measures	Number of defects or errors Amount of scrap material or wastage Number of products returned Number of customer complaints Number of dissatisfied customers Quality of work produced
Personnel information	Absenteeism Turnover Length of service Length of downtime Number of accidents Number of grievances Rate of promotion
Training proficiency	Scores on training exams Scores on performance tests conducted during training Trainer ratings
Judgmental data	Supervisor ratings of performance on the job or of samples of work Peer ratings Subordinate ratings Customer ratings

Box 10.1

Team performance

Modern organisations increasingly arrange groups of employees into teams. There are a benefits to this. First, teamwork provides a social context for work that does much to reduce the alienation felt as a result of division of labour and job simplification, trends that have resulted from job redesign approaches. Being a member of a team encourages communication and the exchange of ideas. Employees have a greater appreciation of where their job fits into the scheme of things, which facilitates motivation and innovation. There is also a social facilitation effect that results from working alongside other people, which leads to not wanting to let the rest of the team down. Finally, there is the opportunity for employees to support and cover for one another. The ability to perform one another's job leads to job enrichment, which makes work more interesting and meaningful.

Teamwork presents a problem for performance appraisal. Job performance in a group context is a function of the whole team rather than the individual. In this context, at what level should performance be assessed: the individual employee or the group? Devising an adequate appraisal of team performance has proven more difficult than assessing individual work performance. Approaches have viewed team performance as an aggregate of individual performance, but should the focus be on the best workers as they will raise the standards of the team, or should the emphasis be given to the weakest workers as they reduce the team standard? We could simply focus on supervisor ratings or team production counts, but this still leaves the issue of managing individual team members (Scott & Einstein, 2001).

Even where productivity counts are useful, it can usually be shown that focusing on purely quantitative measures has undesirable consequences. For example, salespeople evaluated solely in terms of number of goods sold can become so motivated to make sales that they ignore their customer's needs and sell them things they do not want. In other words, quantitative measures of performance are notoriously deficient in terms of quality (see Box 10.2 and Chapter 5). Productivity counts indicate little about the quality of production. As such, it was realised that job performance is a multidimensional construct, having both qualitative and quantitative features that need to be taken into account. Measures of quality could include things like the number of products returned or the number of customer complaints (see Table 10.1 for other suggestions).

Ultimately, it was realised that the multiple aspects of job performance must be combined in some way to obtain a true picture of someone's work performance, and that this aggregation invariably involves human judgment. For this reason, the most common form of job performance measure is the supervisor rating. A judgment by an informed supervisor who appreciates most

aspects of the job (in both quantitative and qualitative terms), and who has an adequate opportunity to observe an employee's performance, is widely recognised as the most viable single measure of job performance. This realisation leads to the question of how to capture supervisor judgment.

Box 10.2

Limitations of 'objective' information

One problem with simple, objective productivity counts is that the rate of production is often outside an employee's control. The production rate of assembly lines, for example, is governed by the pace of the line. Another problem with productivity counts is their limited applicability to jobs outside of the production and distribution chain, which includes most jobs in mature service economies. For example, what should be counted when assessing the performance of a manager or supervisor? The number of meetings attended or memos produced are unlikely to be linked to management effectiveness. Most professionals are also reluctant to view their performance in purely quantitative terms. Is the best police officer one who issues the most speeding tickets? Should a good surgeon rush an operation? Is a good psychologist one who administers lots of tests? For complex jobs, objects and events that can be counted often represent only a small fraction of what the job really entails, and alternative methods of appraisal that combine multiple indicators of performance are required. Objective indices might not be available for clerical staff, training officers, cleaners and many other occupations making up a modern organisation.

Rating scales

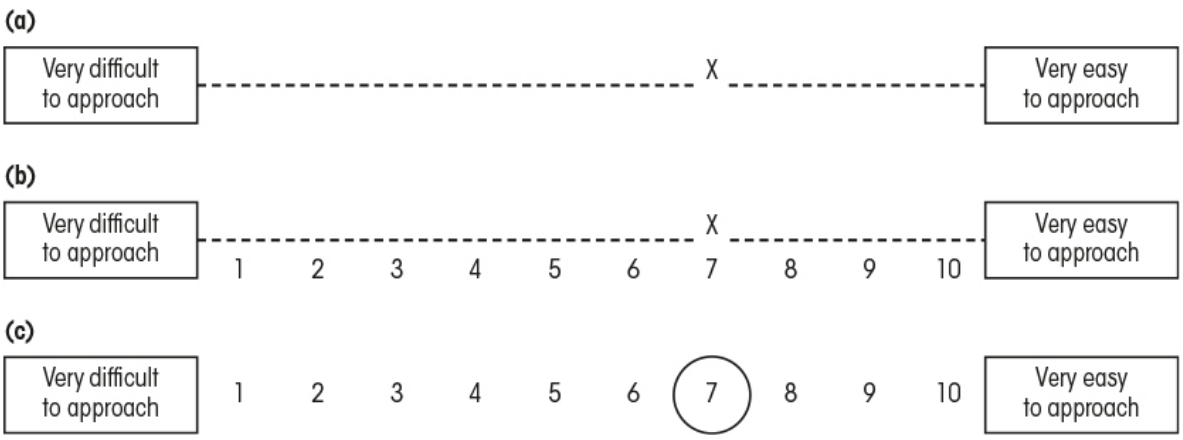
A simple method of capturing judgmental information is the **graphic rating scale**, which involves marking a line or circling a number to represent the level of performance (see Figure 10.1). Scale values can then be summed to produce an overall score on related items. Points along the scale can indicate variations in performance on a range of work-related dimensions. In spite of their apparent simplicity, users often find scales ambiguous in terms of the meaning of particular scale points. What, for example, might be meant by 'average performance' in the middle of the scale? This difficulty led to much research aimed at determining the best anchors for various points along rating scales.

Although this line of research never identified the single best scale format, it did identify a number of useful methods, one of which is the **behaviourally anchored rating scale**, or BARS.

graphic rating scale
a simple rating device used to elicit human judgment, typically completed by marking a point on a line or by circling a number (say from 1 to 10) to indicate the strength of agreement with the item

behaviourally anchored rating scale
(BARS) a rating scale that includes actual behaviours to indicate the response

Figure 10.1 Examples of graphic rating scales



BARS are rating scales with explicit behavioural statements to indicate the kind of behaviour expected at that point along the scale (see Table 10.2). The advantage of BARS is that actual behaviours are associated with values on the scale. Besides being more clear, behavioural anchors also provide an element of standardisation among raters: raters are not left to their own device when deciding what constitutes a score of, say, 4 out of 5. Further, the anchors also provide a basis for interpretation of the scale scores. Thus, inter-rater reliability can be improved and a more meaningful interpretation provided. Behaviourally anchored rating scales are created by identifying behaviours that are typical of good and poor performers, and having these ranked by experts (e.g. a group of supervisors) so that good examples can be found for the entire range of the scale.

Table 10.2: A behaviourally anchored rating scale for the role of customer service operator

Value	Behaviour
1	Does not attend to customers' needs; argues with customer
2	Attends to customers' needs, but does not take responsibility for finding a solution to problems
3	Attends closely to customers' needs, but does not defuse the situation
4	Defuses situation, but customer might not be completely satisfied
5	Takes responsibility and comes up with creative solutions to problems

The basic steps in developing a behaviourally anchored rating scale are as follows:

1. Obtain **critical incidents** indicative of especially good and bad performances from interviews with job incumbents and supervisors.
2. **Content analyse** these incidents and cluster them into coherent behavioural themes or dimensions. These themes will later form the basis of the individual questions.
3. Engage another group of incumbents and supervisors to rate the incidents within each theme; preferably on a 1 to 5, 1 to 7, or 1 to 9 point scale. The aim of this step is to identify incidents that yield a high level of agreement among raters. These will serve as the behavioural anchors on the scale.
4. Use the subset of anchors that survive Step 3 to represent scale points on the scale. The average rating across judges in Step 3 can provide a good indication of the appropriate scale points. Note that it is important that anchors be developed for the entire range of the scale.

critical incident
an example of extreme levels of behaviour or performance (both poor and exemplary behaviours), which are usually key determinants of subsequent outcomes

content analysis
the process of analysing textual information, either written or oral by, for example, searching for themes, examining frequencies of key words or constructs, and identifying repeating relationships; the procedure can be carried out manually or with computer-based software

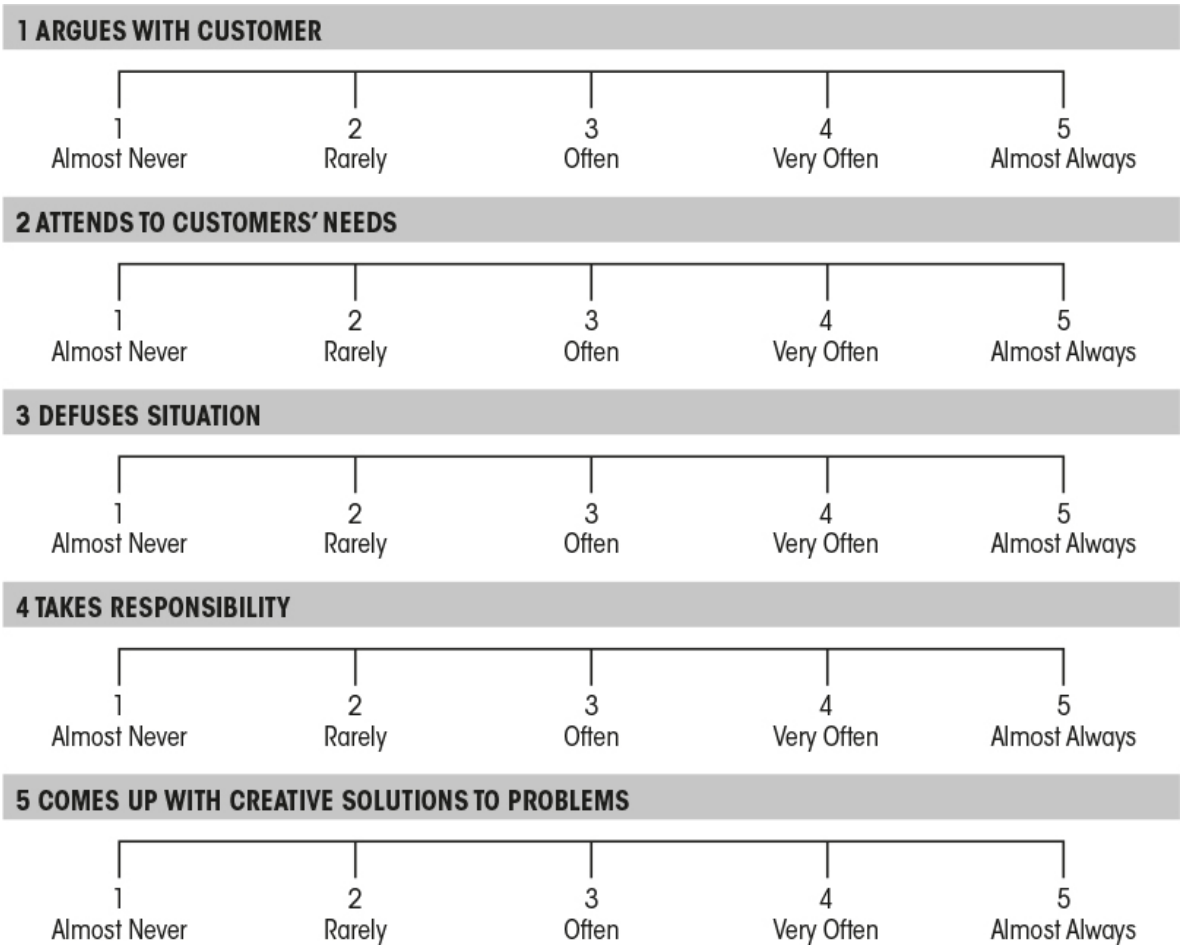
Ideally, a different group of incumbents and supervisors is used for each step in the process so that a wide range of viewpoints is canvassed and each set of judgments is independent.

In spite of their advantages, a drawback of BARS is that raters can sometimes become overly focused on the specific wording used. Rather than viewing the anchors as indicative of a general level of performance, some raters interpret them too literally. In extreme cases, the anchors can trigger specific memories of atypical events, thus biasing the result. Although the person being rated might generally perform well on the job, if they had been involved in one unfortunate incident that was actually used as an anchor, the rater might give them a low rating for that reason alone. Care needs to be given to training raters when using BARS, as rating employees is a skilled activity. Nonetheless, BARS have proven to be a highly popular performance appraisal method. A by-product of the development process of BARS is the involvement of many people from within the organisation, and this level of consultation often generates a high degree of 'buy in' and acceptance for the final system.

Behavioural observation scales

A drawback of the BARS is that some raters might not have actually observed the behaviours used as anchors. To counter this, Latham and Wexley (1977) proposed **behavioural observation scales** (BOS), where the items include the kinds of behaviours that form the anchors in BARS, and the rater is asked to indicate how often the behaviours are observed, where, for example, higher scores indicate higher performance. Identification of behavioural themes to include in the BOS format is determined by similar methods used to develop BARS. Figure 10.2 provides some examples of BOS.

Figure 10.2: Some examples of behavioural observation scales (BOS)



behavioural observation scale
(BOS) questions used in a rating scale that are based on actual behaviours; they are rated for their frequency of occurrence (e.g. from '1 = almost never displayed' to '5 = almost always displayed')

BARS and BOS are linked to typical workplace behaviours. No clear preference has emerged, and it seems that simple graphic rating scales are about as useful as the more sophisticated BARS and BOS (Landy & Farr, 1980).

Other methods

Although rating scales are the most common performance appraisal instrument, it is also possible to simply rank workers from best to worst. This is a fairly straightforward task for most supervisors, as long as the number of workers to rank is not too large. For large numbers, it might be easier to group workers into high, medium or low categories. Grouping might even be preferred when the

number of employees is small because it does not require fine discriminations to be made among individuals.

Another technique involves exhaustive comparisons of workers with one another. In this method of paired comparisons, each worker is paired with every other worker and the supervisor is asked to decide which member of each pair performs better. Such methods can lead to a strong ordering of performance, but becomes unwieldy as the number of pairs increases.

Box 10.3

The role of technology

Technology, especially computers and automation, greatly increases the productivity of employees. This is something all organisations want; however, the task of assessing job performance becomes more complicated as it is difficult to separate the contribution of individuals from the tools they use. In sport, we are interested in the unaided performance of individual athletes, and performance enhancers are seen as cheating (Hesketh & Neal, 1999). In contrast, organisations have no such qualms about increasing a worker's performance by any means possible, but 'performance enhancers' in organisations clearly muddy the water.

Theories of performance

Clearly, job performance is multidimensional and to a large extent contextually specific. This led researchers to focus on different aspects of performance instead of seeking one single, ideal measure (Campbell et al., 1993). Two broad components of job performance have thus far been identified. The first is '**task performance**', which comprises the core technical aspects of the job. The second is '**contextual performance**', which includes those behaviours directed at being a 'good citizen' at work, such as helping out fellow workers or volunteering for committees and incidental activities (Borman & Motowidlo, 1993). Neal and Griffin (1999) included non-observable factors (e.g. planning, problem solving and situational awareness) in their model of performance, and have also added the impact of technology on task performance. To these positive aspects of job performance can be added **'counter-productive behaviours'**, which work against organisations achieving their goals, and include activities such as elective downtime (i.e. 'go slows' or work avoidance), behaviours resulting from alcohol

and/or substance abuse, deliberately destructive or dangerous behaviours, and personal aggression and bullying (Rotundo & Sackett, 2002).

task performance
the core technical aspects and basic tasks that comprise a job

contextual performance
discretionary social behaviours directed at successful performance of the work group or organisation; sometimes referred to as 'citizenship behaviours'

counter-productive behaviours
behaviours that are largely under the control of the individual or reflect problematic employee characteristics, and which impede the progress and success of the organisation

Personnel selection

Personnel selection is the process of choosing which job applicants should receive an offer of employment (Sackett & Lievens, 2008). The aim is to make the offer to the applicants with the greatest probability of success. One way to do this would be to appoint everyone and monitor their performance for a period of time, say for six or 12 months, using the methods discussed above. At the end of the monitoring period, those with the best performance appraisal would be retained and the rest let go. This strategy is known as '**selecting on the criterion**' (the criterion of job performance) and has the great advantage of allowing selection to be made on the basis of actual job performance, which, after all, is what the organisation is really interested in.

personnel selection
the process of choosing which job applicants should receive an offer of employment

selection on the criterion
in personnel selection, the process of appointing all job applicants for a trial period and then retaining only those who have performed satisfactorily

Unfortunately, selecting on the criterion is extremely costly for everyone involved. No organisation can afford to appoint people it does not need, and all

applicants want to know the outcome of their application fairly quickly so they can get on with applying elsewhere if necessary.

Instead of selecting on the criterion, organisations attempt to make a prediction about future job performance based on information collected during the selection process. Hopefully, the prediction will correlate with success on the job, and this is where psychological assessment comes in. Thus, personnel selection is the process of predicting from among a group of job applicants those with the greatest probability of success, based on measurements of personal characteristics that make them more or less suited to the position. Once the prediction has been made, candidates are usually rank ordered and selected top-down from the list. Performance measures described above (i.e. task and contextual performance) serve as the dependent variable against which to validate these predictions. Organisations sometimes try to get some of the benefit of selecting on the criterion by instituting a probationary period for a short time immediately after appointment, usually a few weeks or months.

Predictive validities in personnel selection are usually not perfect because there are always extraneous and unpredictable factors that influence job success, with technology being one such ingredient (see Box 10.3). This means that selection errors inevitably occur (see Chapter 5 on validity). Some people will be appointed whose performance ultimately does not measure up. Such cases are known as false positives or mis-hires. They are positives in the sense that a positive decision was made in their favour, but false in the sense that the decision was ultimately found to be in error. Conversely, some people will miss out on a job offer even though they would have performed well enough if only given the chance. Such cases are known as false negatives. The only way to eliminate these errors would be to use a predictor with a perfect validity of 1.0, but such predictors do not exist. As we have seen in the previous section, even assessment of job performance itself poses many pitfalls. So organisations and applicants must live with an imperfect process, hoping that, in the long run, it works reasonably well most of the time.

Viewed as an assessment and prediction problem, personnel selection is based on the assumption that applicants differ in the knowledge, skills, abilities and other characteristics (**KSAOs**) needed for the job. The task is to identify those applicants whose KSAOs most closely match the requirements of the job. Personnel selection, therefore, is fundamentally the study of individual differences. For over a hundred years, psychologists have been studying the dimensions along which people differ, primarily through psychological tests and other assessment devices. The main outcomes of this research effort are our current theories of cognitive abilities, personality and interests (see the other chapters in this book for discussion of these areas).

KSAOs
the knowledge, skills, abilities and other characteristics of an employee or prospective employee needed to be able to undertake their job satisfactorily

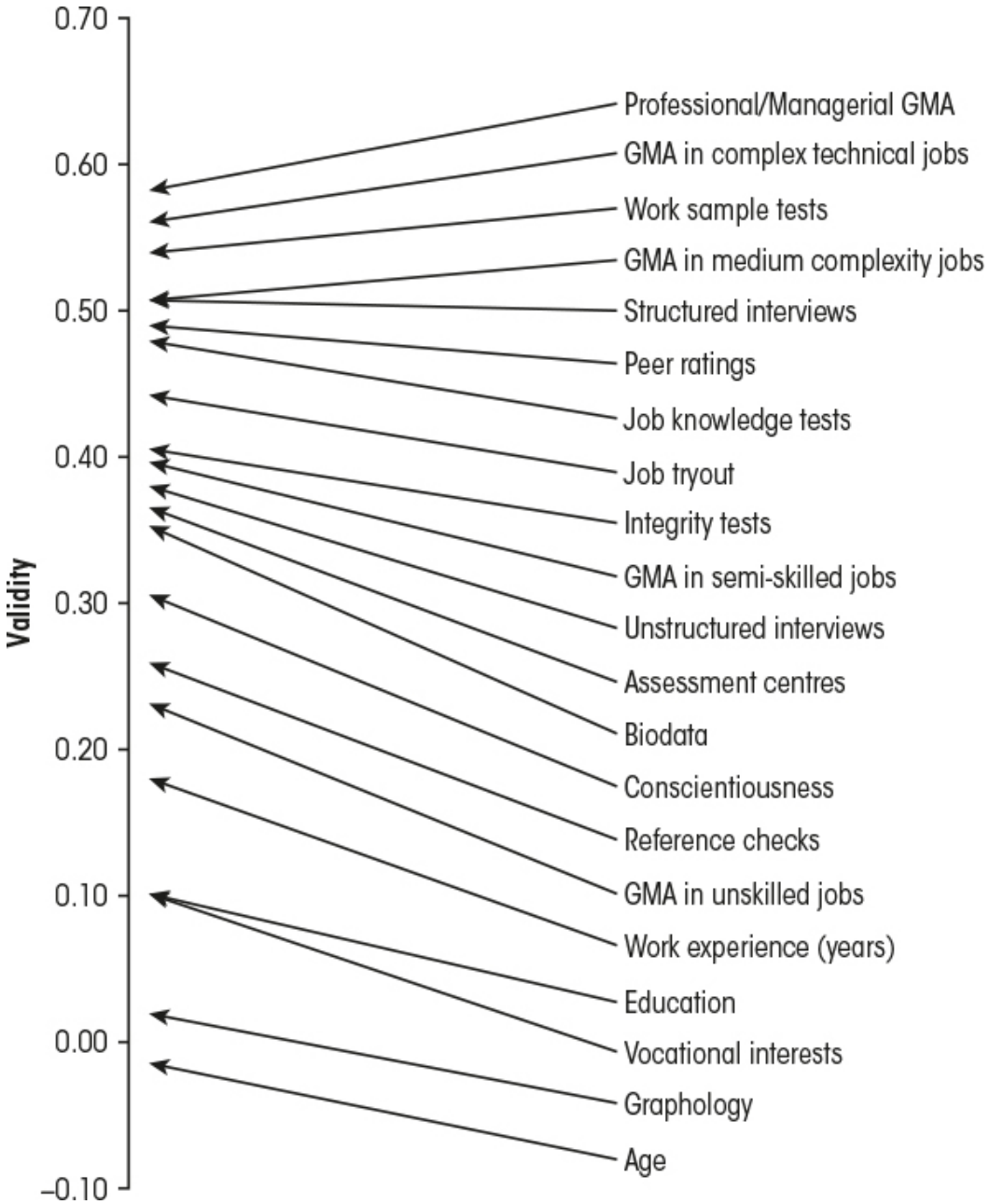
With such a huge research base of attempts to validate many different predictors of many different jobs, meta-analysis has been highly influential in the area of personnel selection. The process of meta-analysing validity coefficients is called **validity generalisation** (VG), because it has been shown that meta-analytically derived validities are highly generalisable across different jobs; that is, the type and level of job does not moderate the validities found.

validity generalisation
(VG) the demonstration that validity generalises across job selection exercises for different jobs by conducting meta-analyses of studies reporting validity coefficients

Psychologists have looked at virtually every individual difference for personnel selection purposes. This literature has been summarised in the Validity Generalisation League Table (Schmidt & Hunter, 1998; see Figure 10.3). The first thing evident is that **general mental ability** (GMA; see Chapter 7) sits as the top two items in the table as the best single predictor of performance in virtually all occupations. This is moderated, to some extent, by job complexity—the only moderator evident in the table—with the validity of GMA being higher for more complex jobs. Schmidt and Hunter (1998) hypothesised that GMA is so successful because it predicts learning, both prior learning and on-the-job learning, which translates into job knowledge and then into successful performance. Apart from its high validity, another advantage of GMA is the low cost of its assessment. Over the decades psychologists have developed a large number of very good intelligence tests that measure GMA. A number of these tests are reviewed later in this chapter, and also in Chapter 7.

general mental ability
(GMA) global intellectual ability

Figure 10.3 The Validity Generalisation League Table



Adapted from Schmidt and Hunter (1998, pp. 262–74)

Given the range of assessment devices evident in the table, an important question is: How does one choose among them? The answer to this question is 'by job analysis'. **Job analysis** is the process of gathering detailed information about a particular job, including the main tasks carried out, and the main requirements for performing the job. Methods of job analysis include questionnaires and tests, interviewing, and questioning workers and supervisors involved in the job. Job analysis is a specialised field within I/O psychology and is largely beyond the scope of this text. However, a selection system cannot be constructed without a detailed understanding of the job in question, and the

process of developing that understanding is job analysis. Indeed, all applications within industrial and organisational psychology begin with developing an understanding of the job through job analysis.

job analysis
the process of gathering detailed information about the main tasks and contextual responsibilities for a particular job

Closely following GMA in the League Table are **work sample tests**. These are specifically designed, hands-on simulations of the main work tasks to be performed. For example, an applicant for the position of a customer service representative might be asked to role-play dealing with a customer complaint, or a computer technician might be asked to diagnose a faulty circuit board. Not surprisingly, the ability to demonstrate good performance on actual job tasks is indicative of high potential for performing the job. The biggest disadvantage of work sample tests is their expense. Off-the-shelf tests are available for a few occupations, but new work sample tests usually need to be developed for each job.

work sample tests
based on the assumption that current, observed behaviour will predict future behaviour, they require job applicants to carry out tasks that mirror those that will be required on the job

Selection interviews are ubiquitous in personnel selection because few managers are prepared to appoint someone they haven't met in person. Unstructured interviews have been found to have some validity, but extensive research shows that providing structure is the best way to increase validity. Structure refers to the degree of discretion the interviewer has in deviating from a predetermined set of questions and format. A structured interview using questions based on job analysis information, asked in a fixed format to each applicant, is the best way to maximise the validity of an interview. Ideally, even the same interviewers and location would be used. In other words, interviews become more valid when they look like standardised tests, even to the extent of scoring them like open-response questions. This usually amounts to thinking of the most likely answers to each question and categorising them as good or poor prior to conducting the interview. Formally scoring an applicant's answers has the advantage of decreasing human judgment and further increasing the objectivity of the process. The same principles of standardisation that apply to psychological tests can be applied to interviews. In some ways an interview can be thought of as a test or questionnaire administered verbally.

selection interviews
usually included as part of any selection exercise, interviews generate ratings based on job applicant responses to questions, which are used to predict success on the job

Not surprisingly, **peer ratings**, which are evaluations of one's performance by co-workers and colleagues, often provide a good indicator of job performance; after all, your peers have usually had many opportunities to see you in action. The main problem with peer ratings is that they are virtually impossible to obtain for applicants from outside the organisation and they can be strongly influenced by interpersonal skills and perceived friendliness, reducing selection to more of a popularity contest.

peer rating
a rating of the KSAOs of an internal job applicant by the job applicant's co-worker/s

Job knowledge tests ask questions about specific aspects of the job. If job knowledge translates into 'know-how', which in turn translates into performance, as suggested above, it is not surprising that job knowledge is a good predictor of performance. The main disadvantage of job knowledge tests is that they are only relevant for experienced workers. New entrants into a field cannot be expected to have developed much job knowledge, unless formal qualifications are a prerequisite for entry, in which case some formal certification should constitute the job knowledge test.

job knowledge test
a test designed to assess knowledge, such as specific technical or professional knowledge, required for a job

Job tryout is a form of selecting on the criterion that involves hiring someone for a few months and seeing how well they fare. This can be expensive and requires great commitment on the part of the applicant and organisation. The effectiveness of job tryout is often undermined by the fact that supervisors can be reluctant to fire people once they are appointed. Probationary periods are a good way of implementing job tryout, but it is not uncommon to see performance decline after the probationary period is over.

job tryout
hiring someone for a short period of time to determine how well they fit in and

perform on the job; a probationary period has a similar purpose

A great deal has been written about **integrity tests** in recent years. Integrity tests attempt to gauge someone's honesty or good character and have also been found to assess dependability and conscientiousness. Integrity tests became very popular in the USA after the use of the polygraph method of lie detection was discredited in the early 1990s. They are relevant for jobs for which a high degree of trust is required, such as security personnel or cash handlers. Two broad classes of integrity test exist: those that are overt and make no attempt to disguise their intent and those that are less obvious or covert. Overt tests are made up of items like: 'How much money have you stolen from your employer during the past 12 months?' It does not take much of an inference to doubt the integrity of an applicant who freely admits to such actions, although it seems odd that questions of this type are not rendered useless by social desirability. After all, who would admit to such offences? Nevertheless, the validity associated with these types of integrity tests shows that they do work, and theories have been proposed as to why they don't lead to blatant distortion. Covert tests are more like personality or biodata tests (discussed below) and do not openly tap honesty behaviours. The inferences made about integrity from such tests are much less direct and it is thought that they are more likely tapping into broad tendencies towards delinquency or antisocial behaviour that might be precursors of specific bouts of dishonesty.

integrity test
either a specific type of personality test or a direct measure to assess a job applicant's honesty, trustworthiness and reliability

An **assessment centre** is a method of assessment, not a place to go to be assessed. Assessment centres usually involve a large battery of tests assessing many different behaviours, and applied to groups of around ten to twenty people at a time. They have been particularly popular for identifying managerial potential. Any of the assessments discussed in this book could be included as an assessment centre exercise, along with various group activities, such as management simulation games, group discussions and oral presentations. Trained observers follow the performance of each participant and compare scores and ratings at the end of the assessment centre to identify the most promising candidates. All exercises used in an assessment centre should have some relevance for the job in question.

assessment centre
a comprehensive testing procedure applied to groups that includes a diverse range of testing tools and techniques

Biodata is short for '**biographical data**' and comprises information about a person's past experience and life history. Some life experiences are highly predictive of job performance. For example, the most famous biodata item was used to select pilots during the Second World War. This single question was almost as predictive of pilot performance as an extensive selection process that included a whole battery of knowledge tests and simulation exercises. The question referred to applicants' childhood hobbies and was simply: 'Did you ever build a model aeroplane that flew?' Answering 'yes' to this question was indicative of a long interest in flying and a level of technical mastery of its basic principles. Not all jobs are amenable to such good biodata items. Traditionally, biodata questionnaires tended to be constructed on purely empirical grounds by trying out different items and seeing how well they correlated with performance. Information supplied in application forms provided a good starting point for such empirically keyed biodata questionnaires (see Chapter 8). More recently, empirical keying has been criticised as being atheoretical, and modern biodata items are chosen on more rational grounds. A good collection of biodata items can be found in Glennon, Albright and Owens (1965).

biographical data
(biodata) measures of past activities, effort and interests that reflect motivation, personality, values and interest, which assume that past behaviours will be consistent with future behaviours

Due to the persistent problem of adverse impact of GMA and other ability assessments in selection, much effort in recent years has been directed towards finding predictors that do not exhibit group differences (see Chapter 7), and hence do not result in adverse impact. In particular, personality constructs have been the focus of much research, as has integrity testing and, more recently, vocational interests. Conscientiousness is one of the Big Five personality factors. Its definition shows why conscientiousness is likely to be related to job performance. Conscientiousness is the only personality factor that appears in the League Table. In spite of its popularity and apparent relevance, personality does not have a great deal of bearing for many jobs. In fact, most jobs are open to a wide range of different personalities as long as the people have the requisite KSAOs for performance. In general, personality factors are more likely to have a greater impact on contextual performance than task performance.

Reference checks are recommendations from previous employers and knowledgeable others who can vouch for you. These are potentially a very useful source of information, although recent experience suggests that many referees are reluctant to pass on negative information. In light of this, a good approach is to conduct a telephone interview, thus avoiding written statements, and simply ask whether or not the referee would be prepared to rehire the candidate.

reference check
a means of verifying job applicant information provided in a resume and collected in an interview; typically done by contacting past employers and/or individuals who can vouch for the applicant

Work experience is the number of years a person has been employed in the line of work applied for. Such experience should represent practice of job-relevant tasks, learning and 'know how', which are all important for high performance. However, the benefit of experience and learning seems to be better captured by GMA than by time on the job per se.

Education refers to the amount of formal schooling, education and training completed. It is somewhat surprising to see education positioned so far down on the table, especially given that educational qualifications are often mandatory for entry into many careers. Does this result mean that unqualified people are likely to perform as well as someone with all of the necessary qualifications? Probably not. The lower validity associated with education is probably a reflection of the restriction of range in education among participants in validation studies. A strict entry requirement for any occupation will be held by all applicants and thus show little variation and, hence, low correlation with any performance measure, no matter how vital it is for performance in the job. In other words, for many jobs, everyone has already been selected for education. If they did not complete their training they would not even be considered. Clearly, minimal educational qualifications are not artefactual, so restriction of range due to education operates legitimately in the world of work and has the result that once educational standards are achieved, education provides little additional information about expected performance.

Research underlying the League Table suggests that vocational interests, discussed below, have little to do with one's ability to perform well in one's chosen occupation. Just because you are interested in a particular line of work doesn't mean that you will be good at it, although there is an interplay between interests and ability to the effect that most people are interested in things they are good at. This has led to a reappraisal of the interests–performance relationship and the role of vocational interests in personnel selection (Van Iddekinge, Putka & Campbell, 2011). In a meta-analysis of more than seventy

studies, Van Iddekinge, Roth, Putka and Lanivich (2011) found somewhat improved validities for interests from that shown in the League Table. The mean validity of interests for predicting job performance was 0.14. The authors also examined a number of other outcome variables and found validities of 0.26 for training performance, –0.19 for turnover intentions and –0.15 for actual turnover (negative validities indicate reduced turnover with greater interest in the line of work). For each criterion, there remained a degree of variability across the validities and the authors found that interest scales that specifically focused on a particular line of work were more effective than generic interest scales, such as scores for RIASEC (see below).

Selection as a social process

Up to now, we have been considering selection purely from the organisation's point of view. From this perspective, personnel selection is the prerogative of the organisation and assumes that all applicants will gratefully accept a job offer if it is made. In reality, applicants turn offers down. A more general framework sees selection as a social process: the outcome of mutual decision making that incorporates the views of both the job applicant and the organisation (Herriot, 2002). While the organisation is appraising the applicant, the applicant is also appraising the organisation and might eventually decide to apply elsewhere. From this perspective, selection is more like a process of negotiation rather than prediction, with each party weighing up the other and searching for grounds for a continuing relationship. The organisation or the applicant can end the process at any point. Even after employment begins, both parties continue to appraise each other.

The social process perspective explains why not all job offers are accepted and why some selection practices persist in spite of their apparent lower validity. In particular, social process theorists argue that unstructured interviews remain popular because this is the arena in which the negotiation takes place. Much current research in selection has attempted to gauge applicant reactions to the selection process, and the findings broadly support the social process view. This research primarily conceptualises applicant reactions in terms of justice perceptions, which are used to predict whether applicants are likely to be inclined to accept a job offer if one is made (Gilliland, 1994). Not all applicant reaction research is framed within the social process perspective, however, because many organisations now realise that all dealings with the community, including job applicants who might not join their organisation, play a role in determining the organisation's public image.

Box 10.4

Equal employment opportunity

Psychological assessment in organisations exists within an extensive legal framework. Laws that directly apply to personnel selection involve principles of equal employment opportunity (EEO); that is, the basic idea that all members of society should have equal access to employment and that employment decisions should be based on merit rather than characteristics irrelevant to the job. It might seem odd that anti-discrimination legislation exists in a context of selection. Isn't the whole point of selection to discriminate among a group of applicants? This might be so, but society seeks to eliminate discrimination on certain dimensions deemed irrelevant and/or likely to be the cause of significant social tension.

EEO principles were first introduced to Australia with the ratification of the International Labour Organisation Convention No 111, Discrimination (Employment and Occupation) in 1973. Since that time they have been enacted through numerous state and federal acts, such as the *Human Rights and Equal Opportunity Commission Act 1986* (Cth) and the *Workplace Relations and Other Legislation Amendment Act 1996* (Cth).

To summarise this legislation, it is unlawful to discriminate against someone in employment decisions on the basis of:

- race (including racial vilification)
- gender (including sexual harassment)
- sexuality (i.e. heterosexual or LGBTI)
- age
- marital status
- pregnancy
- parenthood
- breastfeeding
- status as a carer
- family responsibilities
- political beliefs/activities
- trade union or employer association activity
- medical record

- physical impairment
- intellectual impairment
- physical features
- religion
- criminal record.

Information about any of the above issues should not be collected unless its job relevance can be clearly demonstrated. The types of employment situation covered by EEO legislation include: job advertisements, contents of application forms, interviews, job offers, conditions of employment, opportunities for promotion, access to training, retrenchment and retirement.

Indirect discrimination

When adhering to EEO principles, it is important to realise that discrimination can sometimes be inadvertent. For example, height requirements for police officers or fire fighters, which might at first seem relevant to physically demanding jobs, can discriminate against women because women are, on average, shorter than men. Another example is requesting a photograph as part of a job application. This might at first seem useful for administrative purposes, such as remembering discussions with a particular applicant, but a photograph clearly provides information that falls under the Acts (e.g. information about race, gender and age). For this reason, photographs should, as a rule, not be requested.

Some tests used in selection
Wonderlic Personnel Test

A popular test of general mental ability (GMA) for personnel selection purposes is the Wonderlic Contemporary Cognitive Ability Test (Wonderlic Inc., 2012; formerly the Wonderlic Personnel Test). The test comprises fifty items of varying item types, including vocabulary items that ask for the definition of words and tap verbal abilities; knowledge items that ask about everyday events and tap crystallised abilities; arithmetical items that require some degree of calculation and tap numerical abilities; and figural items that require different shapes to be imagined or compared and tap spatial abilities. The heterogeneous nature of the items allows the total score on the test to reflect GMA. Recall that most definitions of general intelligence emphasise its presence in a broad and diverse

battery of tasks. It is as if the Wonderlic was made up of a few items taken from many different second stratum tests in the CHC model (see Chapter 7). One reason for the popularity of the Wonderlic is that it can be administered and scored in only 20 minutes; including 12 minutes for completion of the test itself. Further, the Wonderlic is available in multiple languages, can be administered to groups and can be taken online. These factors make it a very efficient, cost-effective and popular test with organisational recruiters. The Wonderlic comes with extensive norms for almost 150 occupations and educational groups. Australian norms, however, are not available. The manual reports test-retest reliabilities ranging between 0.82 and 0.94 and validities with the WAIS and WAIS–R in the 0.80s and 0.90s.

ACER General Select and Professional Select Tests

A set of ability tests widely used in Australia are the ACER General Select and Professional Select Tests (Australian Council for Educational Research, 2003; formerly the ACER Higher and Advanced Tests). The ACER General Select Tests are suitable for applicants who have completed Grade 10 and are applying for mid-range technical or administrative positions. The ACER Professional Select Tests, which are more challenging, are suited to applicants who have completed high school and are applying for positions requiring high-level problem solving. Both tests include a verbal or language test and a numerical or quantitative test. There are thirty-four items in the General Select language test, and twenty-nine in the Professional Select version. These tests tap language abilities, including verbal reasoning, similarities, vocabulary and verbal analogies. Verbal reasoning items involve short paragraphs where examinees are asked to evaluate an argument or derive a logical conclusion; similarities items ask examinees to identify a pair of words of similar meaning; vocabulary items ask about the explicit meaning of words; and verbal analogies items are of the form 'A is to B as C is to…'. The quantitative tests (also thirty-four and twenty-nine items, respectively) measure numerical abilities, including number series, matrices and numerical reasoning. Number series items are of the form: '1, 2, 3, 4, 5… Which number comes next?' Numerical reasoning problems are short paragraphs posing problems with an arithmetic solution, such as: 'I started out with X cents and spent Y cents. How much money do I have left?' There is a 15-minute time limit for the language test and a 20-minute time limit for the quantitative test. Both tests can be presented individually or in groups.

The tests also tap general mental ability, although they are deficient if this is the main reason for administering them, as, for example, there is no measure of non-verbal perceptual abilities (see Chapter 7 on intelligence). Australian norms

are available for both versions based on Grade 11 high school students (439 students for General Select and 409 for Professional Select). Cronbach alpha internal reliability coefficients range from 0.80 to 0.89 across the four tests based on these student samples, and the manual reports evidence for validity of the scales.

ACER Short Clerical Test

Tests have been devised specifically for the selection of clerical and administrative staff. In these tests, applicants are assessed for their speed and accuracy and for their numerical ability. A good example is the ACER Short Clerical Test (O'Connor, 2002). This test has two components: one assesses the speed and accuracy of the applicant in checking written and numerical data, and the other assesses basic arithmetic abilities (see Table 10.4). Applicants have five minutes to work through the speed and accuracy test and indicate whether members of each pair are the same or different, and five minutes to complete as many of the 60 items as they can on the arithmetic test. The score on each reflects the number of items correctly answered minus the number of mistakes. Normative data are provided for adult administrative trainees, sales employees, typists and graduate applicants.

Table 10.3: Typical speed and accuracy and numerical ability test items

Speed and Accuracy			
95068	95088	Same	Different
TG Smith Pty Ltd	T.B. Smith Pty. Ltd.	Same	Different
58903	58923	Same	Different
Alright Computer Rentals	Alright Computer Rentals	Same	Different

Arithmetic			
Add	36 + 44 = _____	Multiply	27 x 6 = _____
	$3.28 + $12.73 = _____		$8.25 x 5 = _____
Subtract	62 – 18 = _____	Divide	77 ÷ 11 = _____
	$26.42 – $18.77 = _____		$429.66 ÷ 3 = _____

The tests surveyed here represent only a small fraction of ability tests used in personnel selection. Other tests discussed elsewhere in this book are also used. Although individually administered tests are usually deemed too expensive for use in personnel selection, group administered tests such as Raven's Progressive Matrices (discussed in Chapter 2) are commonly used.

Work attitudes

Psychological assessment figures extensively in surveys of worker attitudes after appointment. Many issues have been investigated using organisational surveys, from classic concerns about job satisfaction to more recent issues of organisational justice and ethics. Other job characteristics that have been examined included degree of task control and complexity, organisational stress and well-being, organisational commitment and work–life balance.

Job satisfaction

Probably the oldest and most naturally interesting characteristic of workers once they have started their job is how satisfied they are. Intuitive notions that satisfied workers are more productive have not been totally borne out by research (George & Jones, 1997), but job satisfaction remains a key antecedent in many theories of turnover (Mowday, Porter & Steers, 1982). The extent to which employee expectations are met has figured highly in understanding the development of job satisfaction. Aspects of job satisfaction have included satisfaction with pay, working conditions, prospects for promotion, level of autonomy and opportunities for training, as well as the social aspects of work including level and type of supervision, and interactions with co-workers, customers and clients. In keeping with theoretical discussions, measures of job satisfaction have tended to emphasise global (Judge, Boudreau & Bretz, 1994; Warr, Cook & Wall, 1979) or facet satisfaction (Spector, 1997; Weis et al., 1967).

Organisational commitment

Organisational commitment refers to how much a worker identifies with or is attached to their organisation, especially in terms of shared values and goals. Perhaps even more than job satisfaction, organisational commitment reflects one's willingness to remain with the organisation and work towards its mission. The most influential model of organisational commitment is Meyer and Allen's (1997) three component model, which defines three aspects of commitment: affective commitment (one's overall degree of liking or attachment to the

organisation); continuance commitment (the sense of one's need to stay with the organisation, especially in terms of how difficult it might be to find alternative employment); and normative commitment (one's sense of obligation to one's employer, built up through past interactions). An alternative conceptualisation is provided by O'Reilly and Chatman (1986), who emphasise compliance, identification and internalisation. Compliance refers to an instrumental agreement between the employee and the organisation to remain together for mutual pragmatic benefit; identification refers to an agreement by the employee to adopt the organisation's values for the sake of ongoing stable employment, without necessarily implying an internalisation of common values; whereas internalisation occurs when the individual's and organisation's values are closely aligned.

Organisational justice

Justice perceptions have been linked to many organisational outcomes (Greenberg, 1987). A basic sense of fair treatment and natural justice can strongly influence an employee's job satisfaction, commitment, organisational citizenship behaviours, absenteeism and even intention to quit. Most people are sensitive to their rights and fair treatment relative to co-workers (known as distributive justice), and sensitive to the basic processes of decision making within an organisation (referred to as procedural justice). Other components of justice often examined include interpersonal justice and informational justice. These aspects of justice refer to fairness in treatment, access to information, and input into decision-making processes within the organisation. Models of organisational justice in the literature tend to differ depending on the emphasis given to the basic forms of perceived justice and fairness.

Vocational interests

The assessment of **vocational interests** is another important area of assessment. Vocational interest tests help to determine what line of work, career or course of study someone might be interested in pursuing. They employ a range of techniques to identify an individual's preferences in this regard. The oldest and perhaps most face-valid approach is to present examinees—often young people who are contemplating a career direction—with a list of occupations, such as police officer, fire fighter, farmer and accountant, and ask them to rate their level of interest in each pursuit. As examinees typically are not experienced in most of the occupations listed, this approach relies on widely held stereotypical beliefs about what constitutes particular lines of work. The lists need to be carefully

prepared to present easily recognisable, prototypical occupations, so that as much of the world of work as possible is portrayed using only a few dozen of the many thousands of occupations available. Once the broad direction of someone's interests is identified, the next stage involves vocational counselling where a more detailed list of possible jobs within a particular job family or theme can be considered. Following this, an even more accurate picture of preferred occupations, one not based on stereotypes, can be achieved through supplementary reading material, the internet, videos, work experience, observation, and discussion with actual job holders.

vocational interests
interests with specific relevance to the workplace, which tend to be stable over time, influence motivation and behaviour, and indicate the type of activities and environments the person prefers

It is clear that just presenting a list of occupations is not completely satisfactory. After all, if one truly knew what was involved in the occupations listed, what would be the point of taking the test? Other lines of questioning about likes and dislikes, hobbies and leisure activities are also found to be very useful in complementing the list of jobs.

The Self Directed Search

By far the most popular vocational interest test is John Holland's Self Directed Search (SDS; Tinsley, 1992), which was first developed in the early 1970s. It has now been taken by millions of people in dozens of countries throughout the world (Ciechalski, 2009). Holland (1919–2008), a US professor of sociology, began working in the vocational interest field in the 1950s and developed a list called the Vocational Preference Inventory (Holland, 1958). Using this instrument, he first identified the so-called 'hexagonal model' of interests (depicted in Figure 10.4), which developed into the theory on which the SDS is now based (Holland, 1992). The SDS consists of five sections:

1. Occupational Daydreams: where respondents are asked to generate a list of jobs that they might find interesting; this list provides insight into the person's ideal occupational aspirations.
2. Activities: where respondents are asked to specify whether or not they like or dislike the type of activity indicated (e.g. repair a car engine, babysit children or write poetry); responses here identify broad likes and dislikes.

3. Competencies: where respondents are asked to specify how competent they think they are at performing certain activities or in using a range of tools (e.g. carpentry tools or word processors); these responses tap into what the person considers they can do well and not do so well.
4. Occupations: where respondents are explicitly asked to indicate their liking for particular jobs (e.g. automobile mechanic, chef or journalist); these tap occupational likes and dislikes.
5. Self-Estimates: where respondents are asked to rate their abilities in various areas, such as mechanical ability, teaching ability or sales ability; these tap broad competency areas (Shears & Harvey-Beavis, 2001).

Clearly, there is a strong element of self-reported competencies and abilities in this test, in addition to the ratings of interest for certain activities and occupations. This is based on the idea that most people like doing things they are good at, even though self-rated ability does not necessarily indicate true ability. There is not always much correspondence between one's interests and one's actual abilities; that is, just because a person is interested in something doesn't mean they will be good at it. Nevertheless, the assessment of interests as well as self-efficacy is a hallmark of modern vocational measurement.

Holland believed that interests were more an expression of personality than ability, and his theory centres on six ideal personality types: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E) and Conventional (C). These types are so important to the theory that **RIASEC** has become a common acronym for it. An important distinguishing feature of the theory, and a reason for much of its popularity, is that it classifies work environments according to the same basic scheme as well. Thus, theory proposes that there are six distinct types of occupation: Realistic, Investigative, Artistic, Social, Enterprising and Conventional. As such, the activities, competencies, occupations and self-estimates in the SDS are organised around these themes.

RIASEC
John Holland's codes for the six types of individual and workplace 'personalities' that he identified (Realistic, Investigative, Artistic, Social, Enterprising and Conventional)

The theory can probably be best understood in terms of the ideal personality types. Thus, according to the theory, Realistic people like interacting with the physical world in a way that involves much practical knowledge but little need for abstract thought, social interaction or self-expression:

- Realistic people tend to be independent, no-nonsense and thrifty, and to prefer being outdoors, working with tools or machinery and to work solo (Holland, 1992). Occupations that permit expression of these personality characteristics include farming, mining, construction, transport and many small business operations such as electrician, motor mechanic, smash repairer, hairdresser or corner shop owner. Currently, over half the available jobs in the economy are Realistic in nature, although, with the rise of automation and new technology, this proportion will decline.
- Investigative people like analysing and solving problems, theorising and dealing with abstract concepts. In particular, they have little interest in business activities. Typical Investigative occupations include science, engineering and other occupations requiring high degrees of technical and theoretical knowledge such as computer programming and financial analysis.
- Artistic people tend to value creativity and have a need to express themselves in creative or artistic ways. The theory embraces the stereotype of Artistic people as being somewhat nonconformist and emotional with a dislike of routine. Besides the fine arts and music, some commercial occupations in the fashion and media industries are considered Artistic. On the whole, though, artistic occupations are relatively rare in the economy.
- Social people particularly enjoy interacting with others, especially in an educational or welfare role, and often have a heightened sense of ethics and social responsibility. They are also supposed to be somewhat impractical and uninterested in manual activities. Typical social occupations include teaching, counselling and the helping professions.
- Enterprising people have a strong business orientation, especially with regard to sales and management, and leadership positions in government and industry. A key source of satisfaction lies in their ability to organise and persuade others to certain courses of action, and they particularly value political and economic power. Consequently, they tend to dislike dealing with abstract concepts and intangibles, especially if they are difficult to explain to others or do not lead to a fairly immediate benefit.
- Conventional types are also reasonably business oriented, but more inclined towards administrative rather than leadership positions. They do not mind routine procedures or structured activities and especially dislike ambiguity and vague task requirements. As such, they tend to be fairly conservative. Typical Conventional occupations include accounting, secretarial, administrative and clerical occupations.

Although these descriptions are highly stereotypical, the theory places an emphasis on profiles rather than types per se when it comes to real people and real work environments. Thus a person's personality is not to be understood

simply in terms of one of the six types, but via their profile on all six types. The descriptions above are of *ideal types* that would rarely, if ever, be encountered in reality. Holland (1992), however, did take the familiarity of these stereotypes as an important source of evidence for their existence. Many other 'typologies' exist within personality theory (see Chapter 8), but Holland's theory is distinctive in its emphasis on occupational characteristics.

Although Holland conceded that all six scores could meaningfully be taken into account, he focused on the three highest scores to produce three letter codes for each profile. Thus a person whose complete set of scores might be, say, R = 6, I = 5, A = 4, S = 3, E = 2 and C = 1 would be coded as RIA, as these are the three strongest interests. An environment whose profile was R = 1, I = 2, A = 3, S = 4, E = 5 and C = 6 would be coded as CES, and so on. Clearly, then, Holland's personality theory comprises 6 × 5 × 4 = 120 types, rather than just six different types. The full profile of six scores would define 6 × 5 × 4 × 3 × 2 × 1 or 720 different types, which would be unmanageable. Environments also are given one of the 120 codes according to the profiles of the people living or working in them, or the types of activities and tasks they predominantly involve. This formulation has the advantage of allowing the same interest inventory to be used to assess both people and environments, and many studies have been carried out to classify occupations according to these terms (Gottfredson, Holland & Ogawa, 1982).

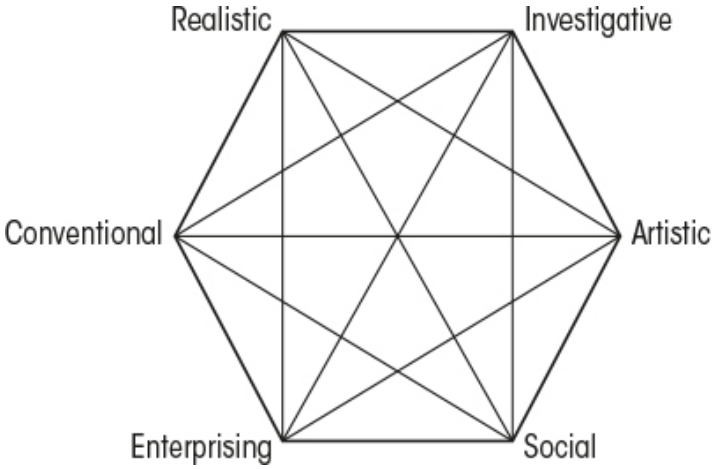
Finding a satisfying job involves essentially looking for a good person–environment fit (P–E fit). For example, a good match occurs for a Realistic person in a Realistic job or an Investigative person in an Investigative job. Evidence broadly in favour of the theory continues to accumulate and its basic principles have not changed appreciably in recent years (Prediger & Vansickle, 1992).

As might be evident from the descriptions above, the types are not independent. The **Holland's hexagon** shown in Figure 10.4 illustrates the hypothesised relationships among the types. According to Holland, 'the distances between the types and environments [on the hexagon] are inversely proportional to the theoretical relationships between them' (Holland, 1992, p. 5). Thus, Realistic types are fairly similar to Investigative and Conventional types (and thus appear closer to one another on the hexagon), but very dissimilar to Social types (which is further away); Artistic types are fairly similar to Investigative and Social types and very dissimilar to Conventional types, and so on. Types opposite one another on the hexagon indicate least preferred activities, which is how stereotypes such as 'social types are impractical', 'enterprising types are impatient with analyses' and 'conventional types are not creative' are derived. Distance or similarity is determined from proximity matrices such as the correlation matrix between scores on scales measuring the six types. Understandably, the hexagonal model has been a major source of construct validity evidence for the theory, and

many studies have sought to verify the hypothesised hexagonal structure among scales measuring the Holland types.

Holland's hexagon
a model that indicates the relationships among Holland's personality types and environments, with similar types placed closer to one another and dissimilar types placed farther away

Figure 10.4 The hexagonal model



The Self Directed Search or SDS is scored by summing the number of votes in favour of each type. The process is very simple and self-scoring is encouraged. Holland believed reflections during self-scoring lead to a greater understanding of the types and yield additional personal insight (Shears & Harvey-Beavis, 2001). Internal consistency reliability for the SDS ranges from 0.83 to 0.91. Validating interest inventories can be difficult because there is not necessarily an external criterion on which to base a prediction. Interest theorists are quick to point out that interests do not necessarily indicate success or suitability for a particular career, so basic career success might not be a good indicator, and information-gathering needs to follow administration of any interest inventory. The SDS has been validated by comparing agreement between the letter code for the highest score generated by the SDS (i.e. R, I, A, S, E or C) and the first letter codes of the occupations listed in the Occupational Daydreams section. The agreement is usually around 50 per cent. This is not an overwhelming endorsement of validity, but it is of the order typically found for other interest inventories (Holland & Rayman, 1986). Perhaps a more acceptable validation method would be to correlate interest scores with measures of job satisfaction among incumbents with several years of experience in a particular occupation.

The simplicity and elegance of the theory has led to its great popularity among career guidance practitioners, and its straightforward and unambiguous

predictions have encouraged much research. Because it can summarise occupational personalities and organise a vast number of occupations into a simple, plausible scheme, it has become one of the most influential theories in vocational psychology. Indeed, perhaps its greatest asset is sheer user-friendliness (Nauta, 2010). The SDS has been tailored for use in many countries and the latest Australian edition, covering over 1000 occupations and specialisations available within Australia, was published in 2012 (Holland, Shears & Harvey-Beavis, 2012). Alternative measures of the RIASEC themes are also available, including a public domain version, which can be accessed via the internet (Armstrong, Allison & Rounds, 2008).

Strong Interest Inventory

The grandparent of all interest inventories is the Strong Interest Inventory (SII), first developed by Edward K Strong in the 1920s (Strong, 1927). The inventory is composed of 325 items that ask about an examinee's interest in occupations, activities, hobbies, school subjects and types of people. Unlike the SDS, the SII has extensive norms, including normative data for Australia, and each examinee's pattern of scores is compared with patterns obtained by satisfied incumbents in over 200 occupations. This is the unique strength of the SII, but means that it must be computer scored in order to tap into its extensive occupational database. Output is divided into three levels of abstraction. At the most abstract level are scores on the six RIASEC themes, followed by twenty-five Basic Interests, followed by, at the lowest level, scores on 211 Occupational Scales. A testament to the influence of Holland's theory is that the SII was eventually reorganised in terms of the Holland codes (Campbell & Holland, 1972). Rounds, Davidson and Dawis (1979) have further suggested that the SII is the best available measure of the RIASEC types.

Figure 10.5 Edward K Strong (1884–1963)

Strong followed an empirical route when developing the SII. He did not start with a theory of vocational interests, but rather began by gathering statements of interest from many different people in many different occupations. These were refined into scales and, later, the hexagonal model was introduced as an organising principle.

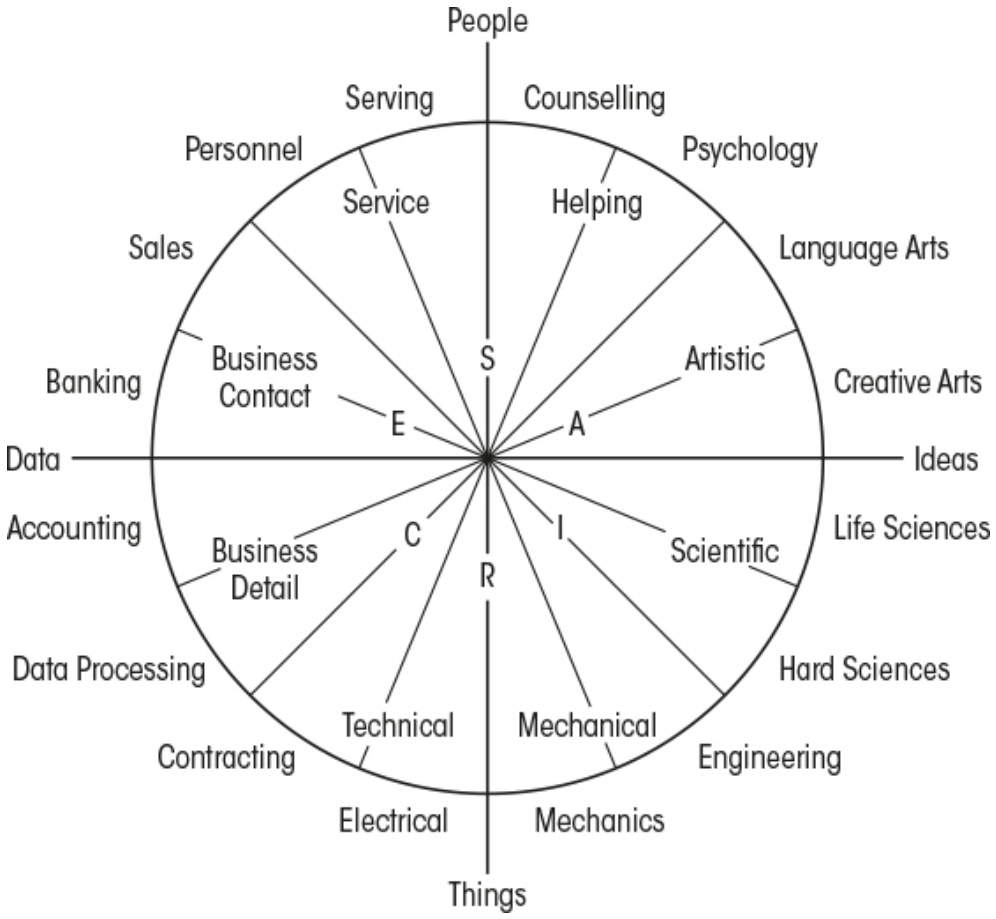
The instrument has been studied extensively during 80 years of use. Test-retest reliabilities are reported for three samples and are generally in the high 0.80s and low 0.90s. Even three-year stability is of the order of 0.80. In terms of validity, the scores have been shown to differentiate among people working in different occupations.

The Circumplex Model

While the SDS and the SII are structured according to Holland's hexagonal RIASEC model of vocational interests and occupations, some researchers have suggested that other arrangements provide better explanations of people and the world of work. Tracey and Rounds (1995), for example, argued that the space of vocational interests is better represented as circular rather than hexagonal (see Figure 10.6). If the vocational interest pie is cut into six slices, it yields the hexagonal model, but finer detail can be obtained by making eight slices or even sixteen, as shown in Figure 10.6. The two main dimensions of the circumplex

have been labelled as the contrast between an interest in 'people' versus an interest in 'things' (the vertical diameter), and an interest in 'data' versus an interest in 'ideas' (the horizontal diameter; Prediger, 1976). Tracey and Rounds (1996) added a third dimension of 'prestige', which resulted in a spherical theory of vocational interests. Other researchers have described these three dimensions in terms of persuasion versus problem solving, structured versus dynamic, and social service versus solitary work (Armstrong et al., 2004).

Figure 10.6 The circumplex structure of vocational interests



One of the biggest challenges for vocational psychology in the twenty-first century is the changed nature of the modern workplace, which is quite different from workplaces of the mid-twentieth century, when these vocational interest theories were developed. A critical question is: Are theories developed some 50 years ago still relevant? Holland's theory of static types, and the assessment devices based on it, are especially vulnerable to this criticism. Traditional notions of career as a choice made largely during adolescence seem particularly outmoded. The world of work is more a moving target than ever before. Unemployment is high in mature Western economies and many young people feel lucky to get any job at all, let alone be able to choose one they find

interesting. Further, with increasing rates of technological change, flattening of organisational structures and global competition, many middle-aged workers, who would have reached career maturity according to the traditional vocational theorists, are now having to revisit career exploration later in life.

Practitioner profile

Dr Elizabeth Allworth

1. **How long have you been a psychologist?**
   I have worked as a psychologist for over 20 years.

2. **What is your specialisation and how did you get the training and experience to do this job?**
   My area of specialisation is organisational psychology. As one of the principals of Allworth Juniper Organisational Psychologists, a small consulting practice based in Sydney, I manage and deliver psychological assessment services for employee selection, and career and leadership development. I have a Bachelor of Arts with Honours from the Australian National University, a Master of Applied Psychology from the University of New South Wales and a PhD from Macquarie University. In my early years as a psychologist I offered vocational counselling to the long-term unemployed and people with work-related injuries. My first experience in psychometric assessment in employee selection came from working in a large recruitment firm, also very early in my career as a psychologist. It was here that I learned about the kinds of assessments that are typically used in the selection context, the practical application of models of person–job fit, and the generation of competency-based reports that integrate results across a range of measures. This experience was very important to the development of my understanding of the management of client relationships in a commercial context.

3. **What kind of clients and referrals do you usually get?**
   Our organisational clients range from small businesses to global corporations, and represent a range of industries. Referrals for assessment are typically made by the human resources or line manager (up to and including the CEO), or might come through the organisation's recruitment agency. In the context of employee selection, the client might refer short-listed job applicants to undertake psychometric and behavioural assessments to help determine their suitability and potential for a particular role or for employment in the organisation. Alternatively, organisations might refer employees for career or leadership assessment to determine their potential career paths within the organisation and their development needs for the future. We also offer support to organisations in the selection, evaluation and validation of assessment measures and methods.

4. **Do you use psychological tests in your practice?**
   Psychological testing is the core of our practice. We use a range of tests, including measures of cognitive ability, personality, motivational needs, career interests and values. For specific purposes, we use measures of sales style, emotional intelligence, leadership or team orientation. We draw upon a range of test publishers and choose those measures that are most likely to assess the attributes required in the job and that best meet the client's needs on a particular assignment. Test takers are advised of the tests they will complete, the nature of the testing process, how they can access feedback on their results, and who will

receive a written report. The assessments are analysed by our consulting psychologists and integrated into a comprehensive, competency-based report. Both the organisation and the individual who takes the test also receive a verbal debrief on the results.

5. **Why do you use psychological tests and in what way do they help you in your practice?**
   We use psychological tests for two reasons. First, there is strong evidence to support the validity of measures of cognitive ability in predicting overall job performance. Although there is less agreement in the literature on the validity of personality measurement as a predictor, there is growing evidence that the validity of personality measurement is enhanced when the attributes measured are linked conceptually to those that are required on the job. The psychometric assessment can therefore help to improve the accuracy of selection decisions and minimise the risk of mis-hiring. The second reason for using tests is to gain a better understanding of the relative strength of the candidate or employee, the potential areas for development, their motivational needs and their career interests. Information gathered from the assessment can be used to guide reference checking, job interviews and development planning.

6. **In your opinion, what is the future for psychological testing in your specialisation?**
   There have been some significant changes over the past 20 years in the way in which psychological tests are used for occupational purposes. While psychological testing was once the domain of psychologists, occupational tests are now used widely by non-psychologists. This has led to a massive growth in the market demand for psychological tests and in the number of tests that are available to test users and organisations. Another major change in the delivery of psychological testing is the capacity for online and unsupervised administration. While the internet and other technological advances open up opportunities for innovation in testing, psychologists are mindful of the threats to the standardisation of tests and the assessment process, and to the authentication of test takers when assessments are completed remotely. These and other changes in the way in which psychological tests are used in occupational settings have highlighted the need to ensure that tests are used ethically and appropriately, not only by psychologists but by all test users. Organisational psychologists have an important role to play in the current effort by professional associations and governments round the world to raise testing standards and to minimise harm to test takers. They also have the opportunity to contribute to the development of a new generation of psychological assessments that draw upon the delivery and analytical capabilities offered by advanced technology, and that demonstrate improved efficiency in administration and enhanced validity in prediction.

Case study 10.1

What do organisational psychologists do when testing and assessing?

Organisational psychologists, who operate as both self-employed practitioners and as practitioners and consultants within business organisations, offer a range of services, many of which rely on formal and informal psychological testing. These services can be grouped under two broad headings of 'employee selection' and 'employee development'.

Employee selection

This includes the processes around recruiting and selecting suitable candidates for advertised jobs. Organisational psychologists contribute to recruitment (i.e. the process of locating and encouraging suitable applicants to apply) by conducting a job analysis to identify the KSAOs for a position. This information allows for the preparation of a job specification (i.e. the detailed requirements of the job) and for identifying strategies for advertising the vacant position. Job analysis is a structured way of answering the 'What do you do in this job?' question. A wide range of strategies are used in a job analysis, including videoing the person doing the job, observing job activities, having employees keep a diary, interviewing the person, and using structured questionnaires.

One structured questionnaire or 'test' used in job analysis is the Position Analysis Questionnaire (McCormick, Mecham & Jeanneret, 1977), which has the person currently doing the job rate such things as the level of decision making, problem solving, oral communication, and social demands required of the job. Another is the Fleishman Job Analysis Survey (Fleishman & Reilly, 1992). This tool draws on multiple 'experts' (e.g. worker and supervisor) who have knowledge of the job to rate it on dimensions of cognitive, psychomotor and physical demands. As all job analysis procedures are taking measures (i.e. measuring the demands of the job), the organisational psychologist will want to use the most reliable, valid and fair procedure or combination of procedures, in the same way they require these qualities in more traditional psychological tests.

After the job analysis, the organisational psychologist will work with HR staff to decide how best to assess the most important tasks of the job. Assessing educational level might be straightforward, requiring the job applicant to produce the required certificate or diploma; whereas assessing cognitive ability and personality might involve the use of tests, such as the Wonderlic Contemporary Cognitive Ability Test (discussed in this chapter) and the NEO Five Factor Inventory (discussed in Chapter 8). Assessing counselling skills might require applicants to pre-prepare a video of themselves working with a client, and assessing managerial skills might involve attending an assessment centre where applicants can role play managerial roles. In addition, a wide variety of standard assessment tools is available to measure almost anything that might appear on a job specification, including tests of motivation, sales aptitude, customer service orientation, safety awareness, leadership approach, emotional

intelligence, and so on. Importantly, assessment tools would only be chosen if there was evidence that they could contribute to predicting success on the job. Activities that are not valid are of no use and should not be used in selection exercises. Examples of these are horoscopes and hand-writing style.

Employee development

This includes ongoing performance appraisal processes that evaluate previous performance (e.g. past 12 months) and set goals for the coming period, as well as processes that identify specific employees for focused development (e.g. in relation to leadership succession and promotion). Performance is commonly appraised by the immediate supervisor, but can include self-ratings, peer ratings and ratings from customers. Standardised measures for performance appraisal are also used. The goal of these tests is to provide an accurate assessment of the employee's performance. One example is the Management Excellence Inventory (Flanders & Utterback, 1985), which is used to assess competencies, such as planning, coordinating, supervising and communicating, held by employees in leadership and management positions. Another tool is 360-degree or multi-source assessment (Fleenor & Prince, 1997), where feedback is sought from a range of people, such as supervisors, subordinates and peers, as well as the person being rated. These tools are available off-the-shelf or are designed for specific organisational situations. Of particular interest to organisational psychologists with these tools is inter-rater reliability and whether those giving feedback have the skills and ability to provide unbiased feedback (i.e. validity and fairness).

Performance appraisal processes, as well as being linked to performance management and promotion decisions, also identify training needs, which might be met within an organisation (e.g. job rotation) or outside of it (e.g. by attending courses). Self- and supervisor evaluation play important roles in identifying training needs, but more systematic tools are also available. The tests in this domain assess strengths and development needs across competencies relevant to the person's current or future job, and are conducted in the context of a training needs analysis (Barbazette, 2006). A free, public domain framework is the General Employee Training Needs Analysis (GETNA) provided by the Westinghouse Electric Corporation Technology Transfer Program (1997).

An important aspect of employee development is assisting employees with their career development. Employees are largely responsible for setting their own broad career direction. However, both employers and employees have needs to be met. The employer wants to develop skills and competencies that will progress their business; employees want to develop KSAOs that will lead them to satisfying and rewarding positions. Employees are attracted to organisations that have effective career development policies, and these policies

help retain personnel in the business. Career development plans need to consider the skills required for the employee's current job, the aspirations of the employee, and the strategies that need to be implemented to achieve the goals of both parties. Career counselling and the use of psychological tools can form an aspect of a career development program. Here the organisational psychologist might use tests to help clarify personal values, career interests or perceived career barriers. The Strong Interest Inventory (discussed in this chapter) is one such test widely used in Australia for this.

Other areas where organisational psychologists use assessment tools and testing are in team building, identifying safety awareness, identifying burn-out/stress in an individual or organisation, after training to assess transfer-of-training, when assessing organisational culture/climate, to undertake organisational reviews and change management processes, and for general staff surveys (e.g. in relation to job satisfaction). In other words, in almost all activities undertaken by the organisational psychologist!

Discussion questions

1. If you were to conduct a job analysis of the 'job' of tertiary student, what KSAOs would you want to include?
2. After identifying the major tasks of the job of 'tertiary student', how would you go about assessing these?

Chapter summary

Performance appraisal involves the assessment of workers' performance on the job. Quantitative indicators exist for some jobs, but need to be supplemented by assessments of quality. However, most jobs require the subjective judgment of performance by a relevant observer such as a manager or supervisor, and simple graphic rating scales have been found adequate for this. Personnel selection involves the assessment of individual differences among job applicants with a view to identifying the KSAOs that predict future job performance. A great many predictors have been identified and summarised in the Validity Generalisation League Table, but predictors should be chosen on the basis of job analysis.

Surveys of existing employees also make extensive use of psychological assessment to measure work attitudes such as job satisfaction, organisational commitment and perceptions of fairness and justice. Vocational interest testing is pervasive in most countries, as most individuals strive to identify and enter a career that meets their personal needs and values. Vocational interest inventories, augmented by measures of values, competencies, and motivation, play a key role in this process.

Questions

1. What are the relative advantages and disadvantages of quantitative and qualitative performance measures?
2. What is the difference between BARS and BOS?
3. What are the steps involved in constructing a BARS?
4. Why is general mental ability the best single predictor of job performance?
5. Why is job analysis so crucial to the process of employee selection?
6. Think of a particular job. What are some potential methods of selecting someone for that position?
7. How might someone's attitudes to their workplace or organisation be assessed?
8. What model did John Holland use to segment individual and organisational occupational 'personalities'?

Further reading

Cascio, W F & Aguinis, H (2011). *Applied psychology in human resource management* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Gatewood, R D, Feild, H S & Barrick M (2011). *Human resource selection* (7th ed.). Ohio: Cengage Learning.

Herriot, P (2001). *The employment relationship: A psychological perspective.* Philadelphia, PA: Taylor & Francis.

Holland, J L (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Murphy, K R & Cleveland, J N (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage Publications.

Patton, W & McMahon, M (2014) Theories of career development. In W Patton & M McMahon, *Career development and systems theory* (pp. 135–81). Rotterdam: Sense Publishers.

Useful websites

APS College of Organisational Psychologists: www.groups.psychology.org.au/cop

Australian blueprint for career development: www.education.gov.au/australian-blueprint-career-development

British Psychological Society—Division of Occupational Psychology: http://dop.bps.org.uk

Society for Industrial and Organizational Psychology: www.siop.org

11 Neuropsychological Testing and Assessment

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. describe the basic structures and functions of the human brain
2. understand different ways the brain can be damaged and the associated effects
3. define clinical neuropsychology and discuss what clinical neuropsychologists do
4. explain the purposes and procedures of neuropsychological assessment
5. list the functions commonly included in a neuropsychological assessment and give examples of the psychological tests/instruments used to assess these functions.

KEY TERMS

attention
clinical neuropsychology
executive functions
language
memory
motor functions
neuropsychological assessment
neuropsychology
sensory functions
visuo-spatial functions

Setting the scene

- The relatives of a 75-year-old war veteran noticed that he seemed to be more forgetful and less able to handle routine tasks than before. He was referred to a clinical neuropsychologist to determine if he was suffering from dementia.
- The disability officer of a university referred a first-year university student who had failed a number of courses to a neuropsychology clinic to find out if she had a learning disability.
- A 10-year-old boy was diagnosed with a brain tumour. After surgery and radiation therapy, he was referred to a clinical neuropsychologist to evaluate the effect of the tumour and treatment on his cognitive functions.
- A young woman was involved in a car accident six months ago. She was referred by a rehabilitation specialist for a neuropsychological assessment to determine whether her cognitive processes were affected by the injury; and, if so, which one and to what extent, and whether or not she could return to her job.

Introduction

In the twenty-first century, it is well known that the human brain is responsible for producing, controlling and mediating our behaviour (Box 11.1 provides a brief description of the structures and functions of the human brain). Damage to the brain caused by external or internal factors, as illustrated by the above examples, can lead to significant changes in functions such as sensation, attention, memory, problem solving, planning, language, visuo-spatial processing and movement—and, as a consequence, to problems in living. The branch of psychology that specialises in the assessment and treatment of brain injury is **clinical neuropsychology**. In this chapter, we provide an introduction to neuropsychological testing and assessment by addressing the following questions: What is clinical neuropsychology? When and how did clinical neuropsychology develop as a speciality area of psychology? What is neuropsychological assessment? What are the purposes and procedures of neuropsychological assessment? What are the commonly used neuropsychological tests?

clinical neuropsychology
a sub-branch of neuropsychology that is applied in nature and concerned with the assessment and treatment of cognitive impairments resulting from brain injury

Box 11.1

Structures and functions of the human brain

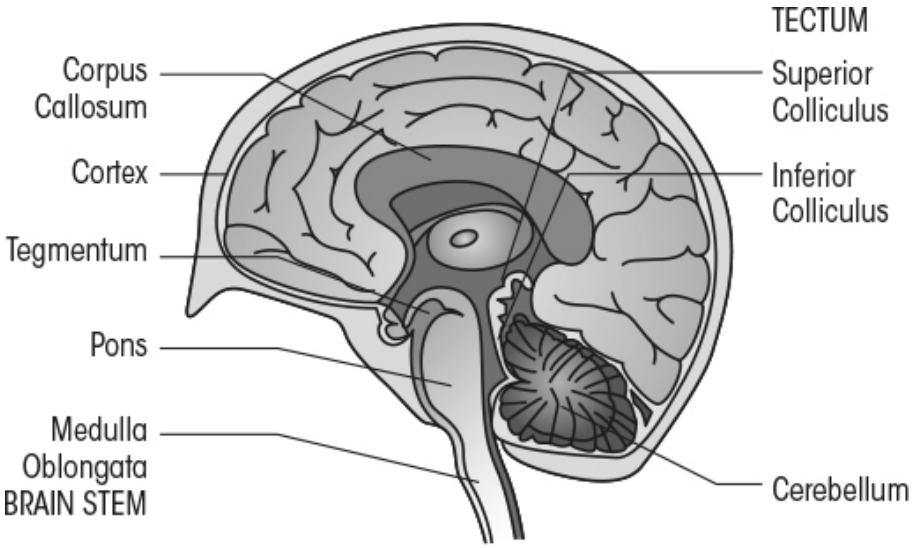
Although the human brain is on average only 1500 grams in weight and 1.4 litres in volume, it is the most complex organ of our body. The brain is made up of 100 billion ***neurons*** or brain cells and ten times the number of ***glia*** (meaning 'glue') or ***glial cells***. The neurons are the basic functional units of the brain. The three main types are ***sensory neuron***, ***motor neuron*** and ***interneuron***. The complex networks formed among the neurons via ***synapses*** (connections), and the electrical and chemical communications of this network enable us to encode and process information and to produce behaviour. As their names suggest, the glial cells hold the neurons together and provide supporting functions.

Structurally, and to a certain extent functionally, the brain can be divided into four main areas: the ***hindbrain***, the ***midbrain***, the ***between brain*** and the ***forebrain***. Continuing from the spinal cord, the hindbrain includes the ***cerebellum*** (meaning 'little brain') and the ***brain stem***. The cerebellum consists of two highly wrinkled structures attached to the brain stem (see Figure 11.1) and their functions include motor learning, coordination of complex motor movements, and coordination of some mental processes. The brain stem is made up of three structures: the ***medulla oblongata*** (meaning 'oblong marrow'), the ***pons*** (meaning 'bridge') and the ***reticular formation*** (meaning 'net-like formation') (see Figure 11.1). The medulla oblongata is situated just above the spinal cord and it has several nuclei that control vital life functions such as the regulation of breathing, swallowing and heartbeat. The pons is a key connection between the cerebellum and the rest of the brain, and it is involved in functions such as eye movements and balance. The reticular formation is located inside the brain stem. It consists of both nerve cell bodies (grey in colour) and nerve fibres (white in colour) and has a net-like appearance; hence its name. The reticular formation is involved in the regulation of sleep-wake cycles and in maintaining arousal.

The two structures in the midbrain include the ***tectum*** (meaning 'roof') and the ***tegmentum*** (meaning 'floor'). The tectum comprises the superior and inferior ***colliculi*** (meaning 'little hills') (see Figure 11.1). While the superior colliculus receives information from the visual pathways, the inferior colliculus receives information from the auditory pathways. These two structures are involved in the production of movements relating to sensory inputs; for example, orienting behaviour to sound or light. The tegmentum is not a single structure but is composed of a number of nuclei. The better-known and more

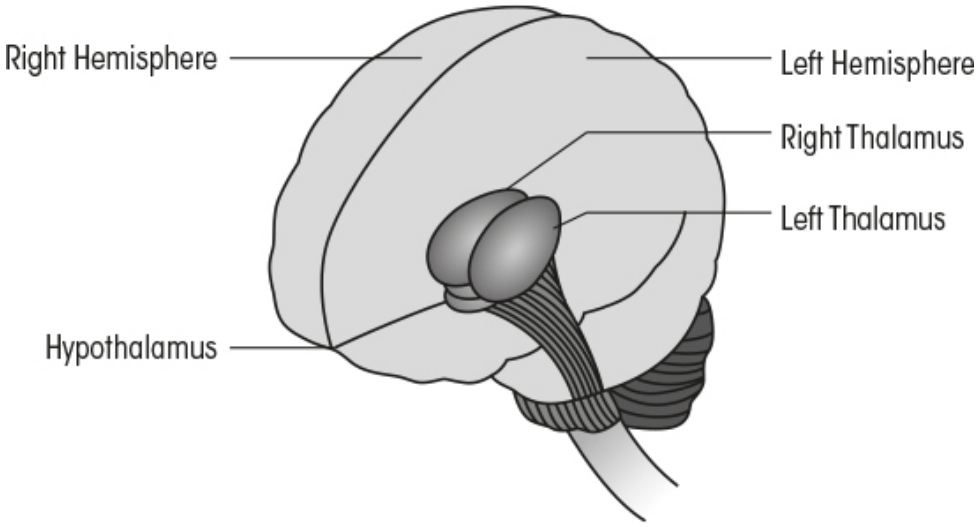
important nuclei include the ***substantia nigra*** (meaning 'black substance'), which is involved in movement initiation; and the ***red nucleus***, which is involved in limb movement.

Figure 11.1 Midsagittal section of the human brain showing structures and locations of the brain stem and midbrain



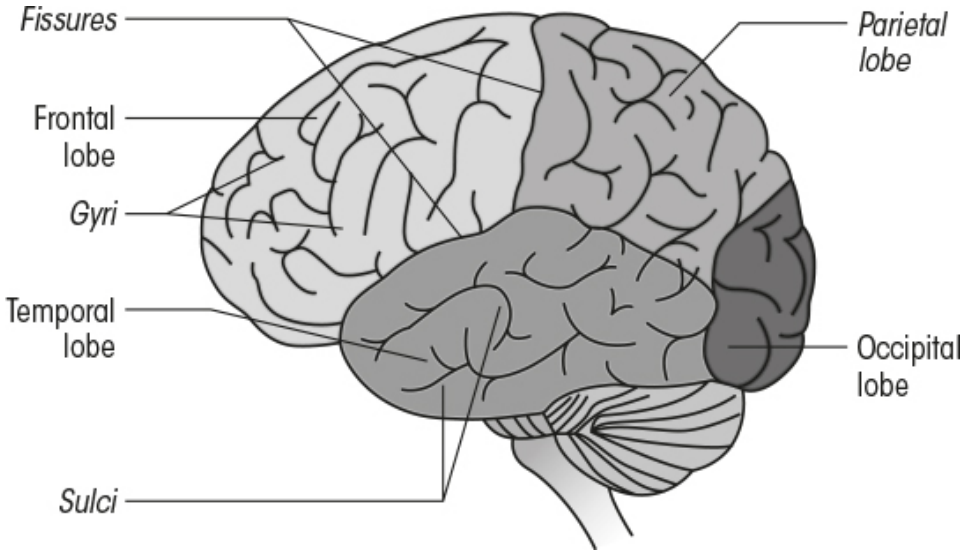
The two principal structures of the between brain are the ***thalamus*** (meaning 'inner chamber') and ***hypothalamus*** (meaning 'under thalamus') (see Figure 11.2). Despite its small size, the hypothalamus is made up of a large number of nuclei (twenty-two) and is involved in many important life functions such as eating, sexual behaviour, sleeping, temperature regulation, hormone function, emotional behaviour and movement. Similar to the hypothalamus, the thalamus is made up of a large number of nuclei, but these are much larger in size than those in the hypothalamus. The thalamus is located strategically between the forebrain and the brain stem. As such, it acts as a gateway or relay station between all the sensory information (with the exception of olfactory information) travelling to and from the brain.

Figure 11.2 Structures and location of the between brain (inside view)



The forebrain or the *cerebrum* is the largest part of the human brain. It is divided into two *cerebral hemispheres* (left and right) that are joined by a structure called the *corpus callosum* (meaning 'hard body'). The corpus callosum consists of 200 million nerve fibres that allow the left cerebral hemisphere to communicate with its right counterpart. The outer layer of the forebrain is the *cortex* (meaning 'bark') and it consists mainly of nerve cell bodies or *grey matter*. The inside of the forebrain comprises mainly nerve cell fibres or *white matter*. Like the cerebellum, the forebrain is wrinkled. This is because the large area of cortex in humans needs to be crinkled up and pushed together in order to fit within the confines of the skull. The bumps on the surface of the forebrain are called *gyri* and the grooves are known as the *sulci*. The deep, prominent sulci are called *fissures*. There are no clear anatomical demarcations for the cortex of the forebrain, but traditionally it is divided into four lobes (see Figure 11.3). The frontal, temporal, parietal and occipital lobes are named after the skull bones above the four areas. Such divisions are, therefore, arbitrary and should not be used as a strict functional guide.

Figure 11.3 The four lobes of the forebrain



The *occipital lobe* is situated at the back of the forebrain and its function is to register, process and interpret visual stimuli. As its name suggests, the *frontal lobe* is situated at the front of the forebrain and its function is to initiate, plan and produce motor behaviours. In addition, the frontal lobe is involved in a group of loosely related processes (i.e. planning, problem solving, working memory, inhibition and regulation) called the executive functions. In recent years, the prefrontal lobe has also been found to be involved in some memory functions. The *parietal lobe* is located immediately behind the frontal lobe and its function is to register, process and interpret somatosensory stimuli (stimuli from the skin and internal organs) and to control visual actions. In addition, because the parietal lobe shares boundaries with the other three lobes, it is involved in the integration of various sensory stimuli. The *temporal lobe* is located underneath the temple area of the human head and its function is to register, process and interpret auditory stimuli. Other functions mediated by the temporal lobe include memory and learning, regulation of emotional behaviour, and identification of visual objects.

Although the four lobes in the two cerebral hemispheres share similar functions for the left and right sides of the body, during the evolutionary process the two hemispheres developed to mediate different functions. Whereas the left hemisphere has become the specialised area for the comprehension and production of language, the right hemisphere has become the specialised area for processing visuo-spatial relationships.

The forebrain also contains two other important functional structures that are located beneath the cortex (see Figures 11.4 and 11.5). They are called the *basal ganglia* and the *limbic system*. The basal ganglia are a collection of nuclei that include the *caudate nucleus* (meaning 'tailed nucleus'), the *putamen* (meaning 'husk' or 'shell') and the *globus pallidus* (meaning 'pale

globe'). These nuclei are responsible for controlling and coordinating voluntary motor movement. The limbic system also comprises a large number of sub-cortical structures that include the ***amygdala*** (meaning 'almond'), the ***hippocampus*** (meaning 'seahorse') and the ***cingulate cortex***. The limbic system has been found to be involved in memory, motivation and regulation of human emotion.

In this section, we have provided a brief and general description of the structures and functions of the human brain. Interested readers who desire a more comprehensive and in-depth treatment of these topics can consult advanced texts and references such as Kolb and Whishaw (2015) and Vanderah & Gould (2016).

Figure 11.4 Structures and location of the basal ganglia (inside view)

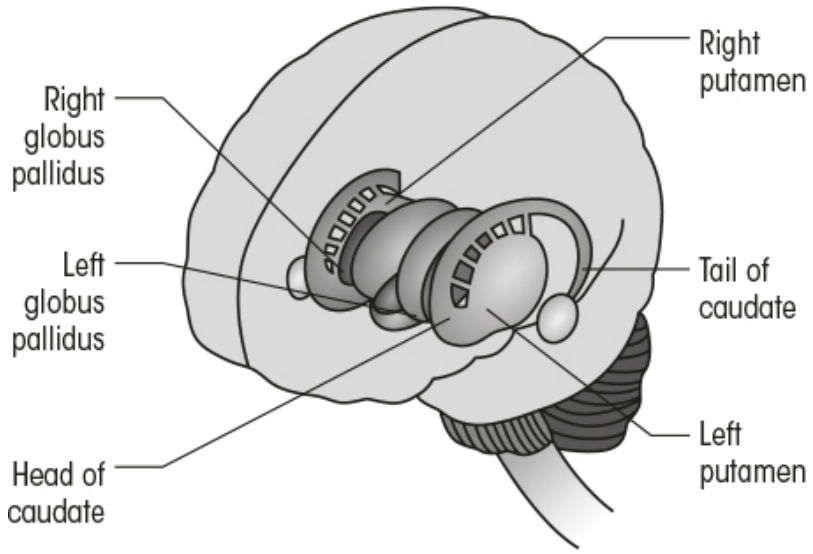
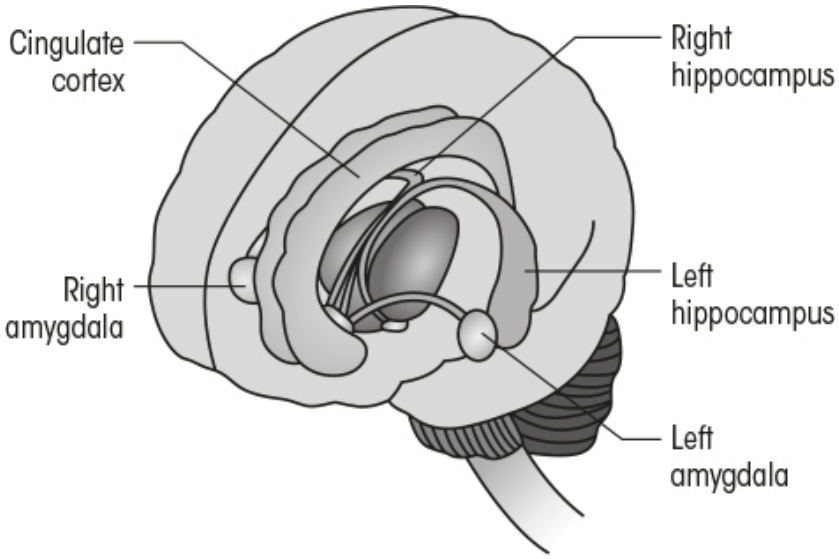


Figure 11.5 Structures and location of the limbic system (inside view)



What is clinical neuropsychology?

Donald Hebb was the first person to formally use the term '**neuropsychology**'—in his 1949 book *Organization of Behavior: A Neuropsychological Theory*—to describe the scientific study of the relationships between the brain and behaviour (Oliveira-Souza, Moll & Eslinger, 2004). In 1967, a group of psychologists formed the International Neuropsychological Society (INS) to promote this newly emerged discipline. Today, the INS has more than 4700 members throughout the world from various areas of practice. Within the discipline of neuropsychology, there are a number of sub-branches, including experimental neuropsychology, comparative neuropsychology, cognitive neuropsychology and clinical neuropsychology. Experimental neuropsychology aims to understand the behavioural organisation of the human brain by studying normal individuals in the laboratory. Comparative neuropsychology tries to achieve the same aim by studying animals such as primates and rats in the laboratory (Milner, 1998). Cognitive neuropsychologists and clinical neuropsychologists both have an interest in patients with brain injury. Whereas the cognitive neuropsychologist studies these patients to identify and clarify the underlying processes of human cognition, the clinical neuropsychologist specialises in their assessment and treatment (Coltheart & Caramazza, 2006; Darby & Walsh, 2005; Heilman & Valenstein, 2012).

neuropsychology
a branch of psychology that aims to study the relationships between the brain and behaviour

Clinical neuropsychology is one of the fastest-growing applied disciplines of psychology and is recognised as a speciality area of psychology in many countries (Hebben & Milberg, 2009; Parsons & Hammeke, 2014). In 1975, the National Academy of Neuropsychology was founded in the USA to represent and promote the interests of clinical neuropsychologists. The division of clinical neuropsychology (Division 40) was officially recognised by the American Psychological Association as a speciality area in 1996, the Special Group in Clinical Neuropsychology of the British Psychological Society was redesignated the Division of Neuropsychology in 1999, and the Board of Clinical Neuropsychology (later changed to the College of Clinical Neuropsychology in 1993) of the Australian Psychological Society was set up in 1983.

Typically, the job of a clinical neuropsychologist includes:

- conducting neuropsychological assessment on individuals with or suspected to have a brain injury
- providing psycho-education, counselling or psychotherapy for individuals with brain injury (and in some cases, their immediate family members or partners)
- planning, conducting and evaluating neuropsychological rehabilitation for individuals with brain injury based on the results of neuropsychological assessment
- conducting clinical neuropsychology research.

While some clinical neuropsychologists perform these functions in a multidisciplinary team with other health professionals (e.g. neurologists, neurosurgeons, physiotherapists, occupational therapists, speech pathologists, social workers and rehabilitation specialists) in a hospital or a rehabilitation centre, some clinical neuropsychologists undertake these tasks independently for clients and/or lawyers in private practice. In most countries, the training of clinical neuropsychologists is reserved for postgraduate programs. For example, in the USA clinical neuropsychologists usually have PhD or DPsych training and are certified by the American Board of Clinical Neuropsychology or the American Board of Professional Neuropsychology (Meier, 1997). In Australia, Masters level training is the minimum academic training for membership of the College of Clinical Neuropsychology.

A brief history of neuropsychological assessment

The field of **neuropsychological assessment** began in the 1940s and 1950s when psychologists were approached by other health professionals to assist in deciding if the behaviour of their patients was due to brain injury or other causes. (Box 11.2 provides a brief description of the major types of brain injuries.) Neurologists and neurosurgeons were interested in whether their patients showed signs of behavioural deficits or excesses caused by damage to the brain; while psychiatrists were concerned about whether the behavioural dysfunction of their patients was due to 'functional' (i.e. non-organic) causes. Before the development of imaging techniques such as computer tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET), psychologists used what they called a 'test for brain damage' or 'test of organicity' to assist them to diagnose damage to the brain. Although some of these tests were rather simple, they were shown to be quite sensitive to the effects of injury to the brain (Lezak et al., 2012).

neuropsychological assessment
the application of neuropsychological tests and other data-collection techniques to answer referral questions or solve problems for individuals with a known or suspected brain injury

Box 11.2

Major neuropathological conditions

Injuries to the brain are usually acquired after birth and they can be caused by either internal (e.g. burst of a cerebral artery) or external (e.g. introduction of neurotoxins to the brain) factors. The causes and effects of some common neuropathological conditions are summarised below.

Alzheimer's disease

This insidious degenerative disease accounts for 50–70 per cent of all dementias. It was named after the German neurologist, Alois Alzheimer, who in 1906 observed abnormal changes (i.e. the accumulation of amyloid plaques and neurofibrillary tangles) in the brain of a 51-year-old female patient with symptoms of dementia. In Australia, it is estimated that people over the age of 65 have a one in fifteen chance of developing this disease. For people over the age of 85, the chance is one in four. The total number of individuals suffering from this disease in Australia, according to Alzheimer's Australia, is about 353,800 and this number is expected to increase dramatically as the population ages (to about 900,000 in 2050). In 2009–10, the total direct health and aged care system expenditure on people with dementia was estimated to be about $4.9 billion. The symptoms of Alzheimer's disease include memory and learning problems, problems with abstract thinking, word-finding difficulty, loss of judgment, disorientation (loss of sense of time, place and people) and personality change. To date, there is no direct clinical test for Alzheimer's disease and diagnosis is by exclusion (i.e. making a diagnosis by excluding as many other causes as possible). The effect of this disease is progressive and irreversible, and the course from diagnosis to death usually takes about seven to ten years.

Traumatic brain injury

There are two types of traumatic brain injury: open and closed head injury. The former is caused by fast-moving projectiles (such as a bullet) or sharp objects (such as a knife). Closed head injury, on the other hand, is caused by the impact

of blunt external forces (e.g. in an assault) or by the sudden acceleration/deceleration of the moving brain (e.g. in a fall or in a motor vehicle accident). In open head injury, the skull is usually perforated, the effect of the injury is confined to the area of the brain damaged by the external object, and loss of consciousness is uncommon. In contrast, in closed head injury, the skull may be fractured but not perforated, the effect of the injury is more widespread, and loss of consciousness is common. In Australia and other countries, closed head injuries are more common than open head injuries and the incidence of closed head injury is estimated to be about 200 per 100,000 head of population. The highest number of closed head injuries occurs in the 15–35 age group and the ratio of males to females is about three to one. The severity of closed head injury is usually assessed using the Glasgow Coma Scale (GCS)—an index of the depth of coma—or the duration of Post-Traumatic Amnesia (PTA); that is, the duration between injury and the regaining of day-to-day memory and orientation. Although the effect of closed head injury depends very much on the severity of injury, common symptoms include slowing in speed of information processing, attentional and memory problems, personality change, impulsivity, emotional problems and speech problems.

Stroke

The initial symptoms of a stroke usually occur suddenly and they can include numbness, weakness or paralysis of the face, arm or leg on one side of the body; loss of speech; blurred or decreased vision; dizziness or loss of balance; headache; and confusion. A stroke occurs when the blood supply to one part of the brain is interrupted or severely reduced. There are two main types of stroke: ischaemic and haemorrhagic. The former type occurs when blood clots or other particles block one of the arteries that supplies oxygen and nutrients to the brain and leads to death of brain cells in one or more parts of the brain. About 80 per cent of all strokes are ischaemic in nature. The latter type occurs when a blood vessel in the brain leaks or ruptures because of hypertension or weak spots in the blood vessel walls called aneurysms. In Australia, stroke is the largest single cause of disability of all neurological disorders. According to the Australian Bureau of Statistics (2009), an estimated 381,400 Australians (1.8 per cent of the total population) reported they had suffered a stroke in 2009. People aged 65 years or older (69 per cent of the total) and males (55 per cent of the total) were more likely to have suffered a stroke. Among the 381,400 people who have suffered a stroke, 35 per cent had at least one impairment that lasted for six months or longer. In 2008–09, health care expenditure for stroke in Australia was over $600 million.

Brain tumour

A brain tumour is an abnormal growth of cells in the brain. There are two main types: primary and secondary. Primary brain tumours originate in the cells in the brain and they can be either benign (non-cancerous) or malignant (cancerous). Secondary brain tumours are metastases (migrating cancerous cells) that originate from other parts of the body. The former is usually less common than the latter (a ratio of about one to three). Brain tumours are most common in people older than 65 years and in children under 8 years old, and they are the second leading cause of cancer death in people under the age of 20 years. A brain tumour can cause different symptoms and these may develop gradually or appear suddenly. The nature and number of symptoms depend on the size, location and rate of growth of a particular brain tumour. Some of the more commonly reported symptoms include headaches, nausea, vomiting, vision problems, loss of sensation or movement in limbs, difficulty with balance, speech problems, personality or behavioural changes, epileptic seizures, hearing problems and hormonal disorders. A brain tumour can cause temporary or permanent damage to the brain, depending on whether it is diagnosed or treated early.

Epilepsy

Epilepsy is a condition in which a person suffers from a seizure or temporary disruption of brain function due to periodic disturbance of the brain's electrical activity. Epileptic seizures can be classified as symptomatic or idiopathic. In the former, the cause of the seizure can be identified; and in the latter, the cause of the seizure is spontaneous and cannot be traced. Epileptic seizures can also be classified according to the origin of the abnormal electrical activity in the brain. Focal seizures are those that originate in a specific area of the brain and then spread to the other parts. Simple partial seizures and complex partial seizures are subtypes of focal seizures. Generalised seizures are those that involve the whole brain without focal onset. Absence (petit mal seizures) and generalised tonic-clonic (grand mal seizures) are examples of generalised seizures. According to Epilepsy Action Australia (www.epilepsy.org.au/about-us), nearly 800,000 people in Australia will be diagnosed with epilepsy at some stage in life and over 250,000 Australians are living with epilepsy. The symptoms of epilepsy depend on the type of epilepsy, but they usually include disruption of sensory function, loss of consciousness and motor problems.

Infection

Because the brain is one of the most important organs of the body, it is well protected by the skull, the meninges (covering of the brain) and the blood-brain barrier (a thin barrier that limits the types of substances that can pass from the blood into the brain). Occasionally, however, the brain can be invaded by bacteria, viruses, fungi, protozoa or parasites, and become infected. The

consequences of these infections can be very serious if they are not treated in time. Meningitis is a general term that describes the infection of the meninges, which can be caused by bacteria or viruses. Encephalitis is the inflammation of the brain usually caused by a virus. Primary encephalitis occurs when a virus directly invades the brain and secondary encephalitis occurs when a virus first infects another part of the body and subsequently enters the brain. Some common symptoms of infection of the brain include headache, drowsiness, seizure, stiff neck, confusion and disorientation, fever, nausea and vomiting. More long-term effects can be generalised and affect the whole brain, but can also be specific. For example, some viruses have an affinity for a certain area of the brain and the behavioural effects depend on the area of infection.

The early successes were encouraging to both the psychologists and the referring health professionals, and led to the rapid development of the field of neuropsychological assessment and the proliferation of tests designed to assess brain damage. In the late 1960s and early 1970s, however, the future of neuropsychological assessment seemed to be in doubt when more sensitive neuroimaging techniques were developed to detect the location and size of structural damage to the brain. Contrary to expectation, neuropsychological assessment continued to flourish in the 1980s and continues to the present day, because both the psychologists and the referring health professionals realised that, apart from diagnosis, psychological tests can be used to provide a comprehensive assessment of the strengths and weaknesses of a person who has suffered a brain injury (Lezak et al., 2012). Results of neuropsychological assessment can be used to provide feedback to the client, monitor recovery, plan treatment and evaluate its effect. For a more comprehensive review of the history of neuropsychological assessment, readers are advised to consult Goldstein (1992), Groth-Marnat (2000a) and Meier (1997).

What is neuropsychological assessment?

Neuropsychological assessment is defined as the application of neuropsychological tests and other data-collection techniques to answer referral questions or solve problems for individuals with known or suspected brain injury. Because neuropsychological tests are sensitive to brain function, they are sometimes considered to be different from the other psychological tests. Although the use to which they are put is different, these tests still retain all the basic characteristics of a psychological test and they still have to fulfil all the required psychometric properties before they can be considered useful.

Purposes and procedures of neuropsychological assessment

A neuropsychological assessment is usually conducted for a number of purposes. These include:

- diagnosis
- description of neuropsychological functions
- prognosis
- treatment planning
- monitoring the rate of recovery
- evaluating the effects of treatment.

As mentioned earlier in this chapter, clinical neuropsychologists are now less involved in the diagnosis of suspected brain injury because of advances in neuroimaging techniques such as CT, PET and MRI scans. These techniques, however, are not 100 per cent reliable and they are not suitable for detecting all types of changes in the brain (e.g. early dementia or Attention Deficit Hyperactivity Disorder). Neuropsychological assessment is still required for diagnostic purposes in ambiguous cases. According to Lezak et al. (2012), a comprehensive description of neuropsychological function has become the most important purpose of a neuropsychological assessment, enabling a clinical neuropsychologist to document the functions that are impaired and those that are spared in an individual after a brain injury. Using this information, the clinical neuropsychologist can explain problems experienced by the individual in everyday living, provide psycho-education and make predictions about the person's return to the community and to work.

The process of neuropsychological assessment generally comprises five steps:

1. interviewing
2. gathering other relevant information
3. neuropsychological testing
4. interpreting test results and integrating information
5. report writing and providing feedback.

Similar to psychological assessment in other areas, neuropsychological assessment typically starts with an interview. During the interview, the client is asked to provide information about the nature and duration of the referral problem, the effect of this problem on their everyday functioning, and their

medical, educational, vocational, social and psychological history. Because brain injury can affect a person's ability to provide accurate information during an interview, information collected is usually checked for accuracy with family members or partners and official records. In addition, reports from the hospital and from other professionals are collected to understand the nature and severity of the injury and to assist in the interpretation of tests results.

Neuropsychological testing is usually the most time-consuming step in neuropsychological assessment. During this step, a client is administered instruments designed to measure a number of important neuropsychological functions. After the tests are administered, a clinical neuropsychologist scores and interprets the results in the context of the test taker's background. For example, an average level of performance on a test can be a good or bad sign, depending on the person's previous educational and academic achievements. For example, an average IQ found at testing for someone who was a university medallist might indicate deterioration of cognitive functioning. Finally, a report is written and feedback is provided to the person and, where appropriate, to family members, partners and the referral agency.

Neuropsychological functions commonly assessed

The functions commonly included in a neuropsychological assessment are sensory functions, attention, memory and learning, language, visuo-spatial functions, executive functions, motor functions and premorbid functioning (Groth-Marnat, 2000a; Lezak et al., 2012). Both fixed and flexible batteries have been used to assess these functions. As the name suggests, the fixed battery uses the same subtests for all clients referred for neuropsychological assessment. The Halstead-Reitan Neuropsychological Battery is an example of a fixed battery (see Box 11.3). The flexible battery approach, on the other hand, uses a number of core subtests for all clients, but uses different subtests depending on the referral question or the results of the other tests. Although both approaches are used in the USA, the flexible battery is more commonly used by clinical neuropsychologists in Australia. In this section, we briefly consider some of the neuropsychological tests commonly used to assess these functions. Because of space limitations, what follows is not meant to be a comprehensive or definitive list of tests for neuropsychological assessment. Readers interested in finding out more about neuropsychological tests of different functions can consult excellent references in the area (e.g. Lezak et al., 2012; Mitrushina, Boone & D'Elia, 2005; Strauss, Sherman & Spreen, 2006).

Box 11.3

Halstead-Reitan Neuropsychological Battery (HRNB)

This battery was originally developed in the 1940s by Halstead to provide a comprehensive measurement of neuropsychological functions. It was last updated in 1993 (Reitan & Wolfson, 1993). The HRNB is an individually administered test and completion of the whole battery takes about six to eight hours. The subtests of the battery and the functions they measure are summarised below.

Category Test

In this subtest, test takers are required to determine the rules for categorising pictures of geometric figures by using feedback based on whether they got the last item correct or incorrect. It measures abstract reasoning and complex concept formation.

Tactual Performance Test

For this subtest, test takers are blindfolded and required to place large wooden blocks of different shapes on the correct cut-out positions of an upright board using their dominant hand, nondominant hand and both hands. After the form board and the blindfold are removed, test takers are also required to draw the outline of the form board from memory. The subtest measures sensorimotor and kinaesthetic abilities and incidental spatial memory.

Speech Sounds Perception Test

Sixty nonsense syllables are presented using a tape recorder in this subtest and test takers are required to pick out the presented sound from four written choices. The functions measured are perception of auditory verbal stimuli, auditory-visual synthesis and sustained attention.

Seashore Rhythm Test

This subtest is presented using a tape recorder. Test takers are required to indicate if thirty pairs of rhythmic sounds are the same or different. The functions measured by the subtest include auditory perception and sustained attention.

Finger Tapping Test

In this subtest, test takers are required to tap as rapidly as possible on a telegraph-type key fitted with a mechanical counter. It measures gross motor speed.

Trail Making Test

This subtest has two parts. In the first part, test takers are asked to use a pencil to connect twenty-five numbered circles on a piece of paper as quickly as possible. In the second part, the task is to connect numbered and lettered circles alternately (1-A-2-B etc.). The functions measured by this test are simple and complex information-processing speed and cognitive flexibility.

Aphasia Screening Test

Test takers are required to undertake tasks such as repeating short phrases, naming pictures, following instructions and copying pictures. It is used to screen receptive and expressive language problems.

Sensory-Perceptual Examination

In this subtest, test takers are required to respond to a series of simple auditory, tactile and visual stimuli, both unilaterally and bilaterally. It measures a person's sensory-perceptual abilities.

The HRNB is one of the tests most commonly used by clinical neuropsychologists in the USA (Camara, Nathan & Puente, 2000; Rabin, Barr & Burton, 2005), but it is not commonly used in Australia and New Zealand (Knight & Godfrey, 1984; Sharpley & Pain, 1988; Sullivan & Bowden, 1997). The main strength of this test battery is the use of a standard set of measures on which patients' performances can be compared. However, it has been criticised because of its inflexibility and the amount of time it takes to complete the battery (Hebben & Milberg, 2009).

Sensory functions

Sensory functions comprise the ability to encode and perceive sensory stimuli in the visual, auditory and somatosensory domains reliably and accurately. Impairments in these functions are important because they limit the amount of stimulus information that can be taken in by the individual. According to Lezak et al. (2012), special care should be taken in assessing individuals with basic

sensory impairments and in interpreting their results. This is because failure to do so may lead to misinterpretations and incorrect conclusions. The Sensory-Perceptual Examination from the HRNB (see Box 11.3) can be used to assess sensory function. Other tests of visual, auditory and tactile perception can be found in Lezak et al. (2012). Sometimes information about these functions can be obtained from sources other than neuropsychological testing (e.g. neurological, audiological and ophthalmological examinations and assessment by occupational therapists and physiotherapists).

sensory functions
the abilities to encode and perceive visual, auditory and somatosensory stimuli reliably and accurately

Attention

Difficulties with **attention** are commonly reported by individuals with brain injury. It is now widely acknowledged that attention is not a unitary construct. Models of attention (e.g. Mirsky et al., 1991; Posner & Petersen, 1990; Shum, McFarland & Bain, 1990) suggest that there are at least three components of attention—attention span, focused attention and selective attention—each with a different neuroanatomical basis. The attention span component refers to the ability to encode and reproduce, in correct order, the stimuli presented, and it is mediated by the inferior parietal lobule (McCarthy & Warrington, 1990). Focused attention is the ability to scan stimuli for a specific target and respond to it. The superior temporal and inferior parietal cortices and structures of the corpus striatum have been found to be associated with this component (Mirsky, Fantie & Tatman, 1995). Selective attention refers to the ability to maintain cognitive or response sets in the presence of distracting stimuli and the cingulate cortex has been found to mediate this ability (Pardo et al., 1990).

attention
the ability to focus on or select one stimulus or process while ignoring another; it has at least three components (i.e. attention span, focused attention and selective attention)

The Digit Span subtest of the Wechsler Intelligence Scales is commonly used to assess attention span. In this subtest, number sequences of varying length are presented aurally to the test taker. The Forward condition of the subtest requires the test taker to repeat the number sequence in the order presented and the Backward condition requires repetition of the sequence in the reverse order. The

reliability of the subtest has been found to be excellent (split-half reliability = 0.90, test-retest reliability = 0.83; Kaufman & Lichtenberger, 1999). Performance on the Digits Forward and Digits Backward subtests has been found to be sensitive to damage to the left temporal area of the brain. Performance on the Digits Backward subtest has been found to be sensitive to right frontal-lobe injuries (Golden, Espe-Pfeifer & Wachsler-Felder, 2000). Visual attention span can be assessed using the Spatial Span subtest from the Wechsler Memory Scale Third Edition (WMS–III; Wechsler, 1997b).

The Trail Making Test (see Box 11.3) and the Digit Symbol subtest from the WAIS are commonly used tests of focused attention. On the Digit Symbol, a test taker is given a key that pairs a different geometric shape with the numbers 1 to 9, and asked to draw the shape appropriate to each number in a random sequence of the numbers 1 to 9. The subtest has a time limit of 120 seconds. It has been found to have a test-retest reliability of 0.86 and has been found to be very sensitive to the effect of brain injury (Kaufman & Lichtenberger, 1999).

Stroop (1935) found in a series of experiments that the names of colours were difficult to read if the colour in which the name was printed did not correspond to the name of the colour. Thus 'green' printed in red took more time to read than when it was printed in green. Golden and his colleagues (1978, 2002) developed a commercially available test based on these findings, the Stroop Color-Word Interference Test, which is generally considered a test of selective attention. There are three trials in the test. Each trial uses a different card on which five columns of twenty items are printed. In the first trial the test taker is asked to read, as quickly as possible, rows of colour names (i.e. red, green and blue) printed in black ink. In the second trial the test taker is asked to name as quickly as possible the colour of four Xs printed on a card. Finally, in the third trial, the test taker is required to name the colour of the ink in which words are printed. All words are printed in a colour conflicting with the name indicated (e.g. the word 'red' printed in green or the word 'blue' printed in red). Each trial has a time limit of 45 seconds and the key measure obtained for this test is an interference score derived from the three trials. The psychometric properties of this test are good. For instance, the Stroop has good test-retest reliability (0.75 to 0.90; Uttl & Graf, 1997); is moderately related to the Perceptual Organisation and Freedom from Distractibility factors of the WAIS; and loads in factor analyses on a component of attention called sustained mental processing (Shum, McFarland & Bain, 1990). Further, the Stroop interference score has been found to be sensitive in distinguishing those with brain injuries from their non-injured peers (Hanes et al., 1996).

To address the issue of ecological validity in neuropsychological assessment, Robertson et al. (1994) developed the Test of Everyday Attention (TEA), which uses everyday, familiar materials to assess various components of attention. This test takes 45 to 60 minutes to administer, is suitable for individuals aged 18 to 80

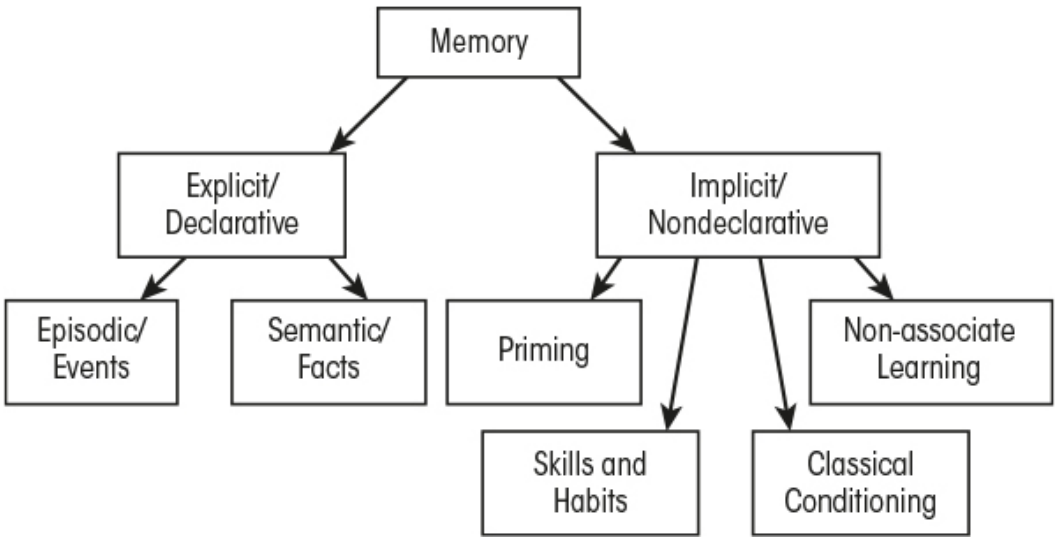
years and has three parallel versions. The subtests of the TEA include Map Search, Elevator Counting (with and without distraction), Visual Elevator, Elevator Counting with Reversal, Telephone Search, Telephone Search While Counting, and Lottery. The norms of the TEA comprise 154 volunteers (18 to 80 years old) and are divided into four age bands (18–34, 35–49, 50–64 and 65–80) and two levels of education. The test-retest reliability of the TEA subtests was found to range from 0.59 to 0.86 for normals and 0.41 to 0.90 for stroke patients. Moreover, the test has been found to be sensitive to the effect of closed head injury and stroke.

Memory and learning

According to Squire (1987), 'Learning is the *process* of acquiring new information, while memory refers to the persistence of learning in a state that can be revealed at a later time' (p. 3). Similar to attention, **memory** is not a unitary construct (Markowitsch & Piefke, 2010). Figure 11.6 is a model of memory adapted from Squire (1992). It can be seen from the model that there are two types of memory: declarative and nondeclarative. Most clinical tests focus on declarative rather than nondeclarative memory. This may be because deficits in nondeclarative memory are not commonly found after brain injury (Shum, Sweeper & Murray, 1996). Declarative memory can be further divided into episodic and semantic memory (Tulving, 1972). Semantic memory represents a person's knowledge of the world (e.g. date of major events or details of historical events). Episodic memory, on the other hand, is the memory for personal events (e.g. the name of one's primary school or what one did last Christmas). Most tests of memory and learning are involved in the assessment of episodic memory. Episodic memory can be subdivided into short- and long-term memory. Because of the lateralisation of brain function, it is also necessary to assess memory for both visual and verbal materials. One type of memory that is not included in the above model is prospective memory, or the ability to remember to do things in the future (Kliegel, McDaniel & Einstein, 2008). This construct has gained a lot of attention in recent years because prospective memory has applied implications and it has been found to be impaired in many clinical populations (Kliegel, Jager, Altgassen & Shum, 2008).

memory
the ability to encode, store and retrieve past information

Figure 11.6 A neuropsychological model of human memory



Adapted from Squire (1992)

One of the most commonly used batteries for memory and learning is the Wechsler Memory Scale–Fourth Edition (WMS–IV; Wechsler, 2009b). The earlier editions of this battery include the WMS (Wechsler, 1945), the WMS–Revised (WMS–R, 1987) and the WMS–Third Edition (WMS–III, 1997b). The WMS–IV was developed to provide a comprehensive assessment of memory functioning that is clinically relevant. It is an individually administered test designed for individuals aged 16 to 90 years.

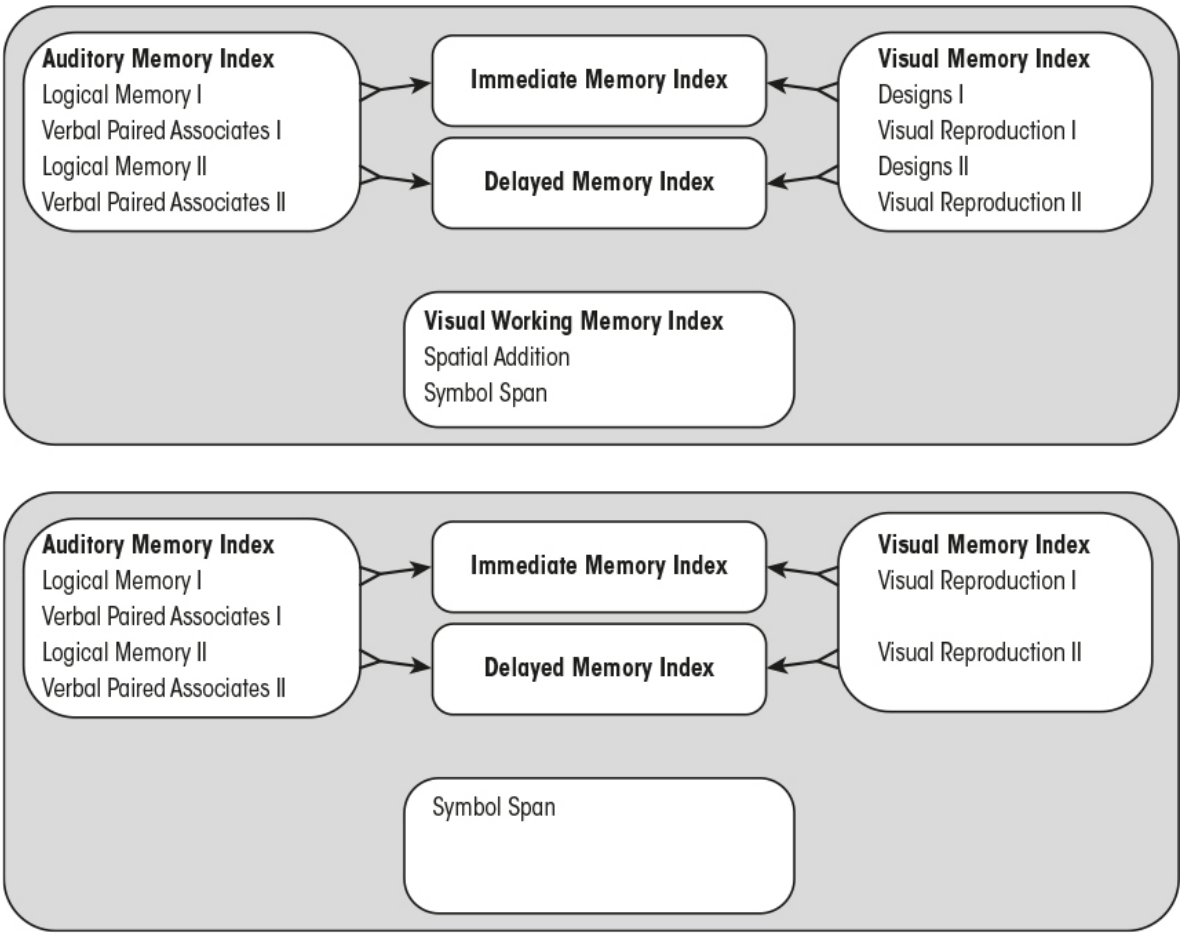
The WMS–IV comprises six subtests and an optional Brief Cognitive Screen (see Table 11.1). Of the six subtests, four of them (Logical Memory, Verbal Paired Associates, Designs and Visual Reproduction) have an immediate and a 20- to 30-minute delay condition. One of the major differences between the WMS–IV and the previous versions is the use of one battery for adults aged between 16 to 69 years old and another slightly modified battery for older adults between 65 to 90 years old. The adult battery takes longer to complete because all six subtests are administered. Five indices—Auditory Memory, Immediate Memory, Delayed Memory, Visual Memory and Visual Working Memory—can be derived based on results of these six subtests. The older adult battery takes less time to complete because it includes only four of the subtests: Logical Memory, Verbal Paired Associates, Visual Reproduction and Symbol Span. Four indices—Auditory Memory, Immediate Memory, Delayed Memory and Visual Memory—can be derived from results of these four subtests. Figure 11.7 illustrates the subtests and indices for the two batteries for the two age groups. To facilitate the use of the WMS–IV in Australia and New Zealand, adaptations (language and cultural) have been carried out. A training/demonstration CD is available and software packages can also be purchased for computer scoring and interpretation.

Table 11.1: Subtests of the WMS–IV

Subtests	Description
Logical Memory I & II (age range = 16–90 years)	Recall of details of two short stories read to the test taker
Verbal Paired Associates I & II (age range = 16–90 years)	Learn, over a number of trials, a list of eight word pairs (e.g. table–flower).
Designs I & II (age range = 16–69 years)	Learn, recall and recognise visual and spatial information of visual images presented within a grid (four items of increasing difficulty)
Visual Reproduction I & II (age range = 16–90 years)	Learn, recall and recognise a number of geometric figures
Spatial Addition (age range = 16–69 years)	Test taker is shown two grids with blue and red circles and asked to add or subtract location of circles according to rules
Symbol Span (age range = 16–69 years)	Test taker is shown a page with a series of abstract symbols and then a different array of symbol; they have to identify the correct order the symbols were presented on the first page
Brief Cognitive Status (optional)	Includes tasks such as orientation, mental control, draw a clock, recall objects named previously, inhibition of responses and verbal production

The norms of the WMS–IV comprise a sample of 1400 individuals from thirteen age groups (16–17, 18–19, 20–24, 25–29, 30–34, 35–44, 45–54, 55–64, 65–69 [adult battery], 65–69, 70–74, 75–79, 80–84 and 85–90 years of age [older adult battery], with 100 individuals for each of the age groups). Individuals in each age group were recruited according to a stratified sampling procedure according to education level, race/ethnicity and geographic region based on the 2005 US Census data. The WMS–IV was co-normed with the WAIS–IV (Wechsler, 2008). Similar to the WAIS–IV, raw scores of subtests are converted into scaled scores and then summed to derive various Index Scores (mean of 100 and standard deviation of 15). In terms of interpretation, a number of procedures such as process analysis, differences within indices and contrast scores can be obtained. Details of these steps can be found in the WMS–IV manual and in Groth-Marnat (2009).

Figure 11.7 Subtests and indices of the Weschler Memory Scale-Fourth Edition



Source: Wechsler (2009b).

Similar to the WAIS–IV, the WMS–IV has been found to have very good psychometric properties. The average internal consistency of the WMS–IV Index and Subtest Scores for the normative sample were found to range from 0.92 to 0.97 and 0.74 to 0.97, respectively. Test-retest reliability was evaluated by administering the battery to 244 individuals (adult battery for 173 individuals and older adult battery for 71 individuals) on two occasions. The average interval for test-retest was 23 days (range = 14 to 84 days). For the adult battery, the average stability coefficients for the Index and Subtest Scores were in the ranges of 0.81 to 0.83 and 0.59 to 0.77. The corresponding coefficients for the older adult battery were 0.80 to 0.87 and 0.69 to 0.81. The WMS–IV manual provides evidence to support the validity of the test. Basically the WMS–IV has been found to be sensitive to damage caused by brain injury. For example, individuals with intellectual disabilities, traumatic brain injury, schizophrenia and Alzheimer's disease have been found to perform significantly more poorly than normals on the WMS–IV. In addition, the WMS–IV has been found to correlate significantly with other memory and cognitive tests (e.g. WMS–III, California Verbal Learning Test, Children's Memory Scales, WAIS–IV and the Wechsler

Individual Achievement Test–Second Edition) and results of factor analyses support the memory indices used in the test battery.

The Rey Auditory Verbal Learning Test (RAVLT) and the Rey-Osterreith Complex Figure Test (Rey, 1964) are popular and are commonly used by clinical neuropsychologists in Australia and other parts of the world to assess verbal and visual memory (Rabin, Barr & Burton, 2005; Sullivan & Bowden, 1997). In the RAVLT, the test taker is read a list of fifteen words five times and asked to recall as many words as possible after each trial. After that, the test taker is read a second list of fifteen words and asked to recall as many words as possible from this list. This is followed by an immediate recall, a 20-minute delayed recall and a recognition trial of the first list of words. A number of indices can be obtained from test performance (e.g. number of words recalled for each of the eight trials, number of words recognised, total number of words recalled for the first five trials of the first word list, learning, and retroactive and proactive interference). Using a sample of fifty-one normal volunteers and a test-retest interval that ranged from six to 14 days, Geffen, Butterworth and Geffen (1994) found the test-retest reliability of the RAVLT to be modest (median r = 0.60). In terms of validity, the RAVLT has been found to be sensitive to verbal memory deficits in those with Alzheimer's disease or those with closed head injury (Bigler et al., 1989). Geffen et al. (1990) have collected normative data for this test in Australia. In the Rey-Osterreith Complex Figure Test, a test taker is first asked to copy a two-dimensional drawing made up of lines and shapes. The test taker's visual memory ability is assessed by an incidental recall of the figure and delayed recall 20 to 30 minutes after initial presentation. The Rey-Osterreith Complex Figure Test showed a high inter-rater reliability (i.e. >0.95) when strict scoring criteria are observed (Strauss, Sherman & Spreen, 2006). Further, this test has been shown to produce consistently different response patterns in those with posterior and frontal lobe lesions (Lezak et al., 2012).

The Cambridge Prospective Memory Test (CAMPROMPT) was developed by Wilson et al. (2005) as a psychometric test of prospective memory for individuals 16 years and older. In this test, test takers are required to perform three time-based tasks and three event-based tasks while performing some ongoing activities. Time- and event-based prospective memory are, respectively, the abilities to remember to carry out an intention at a certain time or after a specified duration of time and when a certain external cue appears. The test takes about 25 minutes to complete and spontaneous use of strategies such as note taking is allowed. Total CAMPROMPT scores are out of 36, with higher scores reflecting better PM performance. Norms have been collected for a group of 212 normal controls and a group of individuals with brain injury. The test has been found to be sensitive to brain injury and to correlate with retrospective memory and other cognitive processes.

Language

For most right-handers, the function of **language** is mediated by the left cerebral hemisphere. Assessment of the language function of an individual with known or suspected brain injury, therefore, enables a clinical neuropsychologist to draw some conclusions about the functioning of the left cerebral hemisphere of that individual. Because of the significance and utility of language in our society, language problems resulting from brain injury can have important implications for the recovery and rehabilitation of individuals with such injury. Clinical neuropsychologists, as well as speech pathologists/therapists, are interested in the assessment of language. A comprehensive assessment typically includes both spoken and written language (Mapou, 1995), with input (understanding written and spoken words) and output (speech production and writing) functions within each.

language
for most right-handers, the function of the left cerebral hemisphere; it includes the ability to understand and produce speech

Screening tests (e.g. the Aphasia Screening Test of the Halstead-Reitan Neuropsychological Battery) allow a brief assessment of a person's language functioning. However, other tests are needed to provide a more comprehensive assessment of the various areas of language functioning. The Western Aphasia Battery–Revised (WAB–R; Kertesz, 2007) and the Boston Diagnostic Aphasia Examination (BDAE; Goodglass, Kaplan & Barresi, 2000) are two examples of comprehensive language assessment batteries. According to Lezak et al. (2012), the WAB–R has satisfactory reliability and validity, and is sensitive in distinguishing the language abilities of those who have suffered stroke in the left versus the right hemisphere and those with mild Alzheimer's disease. The BDAE has inter-rater agreement that is typically above 0.75 (Lezak et al., 2012), and Davis (1993) found that BDAE scores predicted performance on other aphasia tests better than patient functioning in everyday circumstances.

Visuo-spatial functions

In contrast to language, **visuo-spatial functions** in humans are generally mediated by the right cerebral hemisphere of most right-handers. Damage to the right cerebral hemisphere has been found to affect a person's ability to perceive and understand visuo-spatial relationships and undertake three-dimensional constructional tasks.

visuo-spatial functions
usually considered functions of the right cerebral hemisphere include, the abilities to perceive and understand visuo-spatial relationships and undertake three-dimensional constructional tasks

Although a person's visuo-spatial abilities can be gauged from performance on other tests (e.g. the WAIS–IV), specific tests of visuo-spatial functions have been developed. The Hooper Visual Organisation Test (HVOT; Hooper, 1983) is an example. In the HVOT, a test taker is asked to identify thirty pictures of 'cut-up' objects (see Figure 11.8 for an example). Based on a sample of 166 college students and a sample of seventy-three psychiatric in-patients with mixed diagnosis, the split-half reliabilities of the HVOT were found to be 0.82 and 0.78, respectively (Hooper, 1948, 1958). Lezak et al. (2012) reported that the test-retest reliability for the HVOT varies from 0.68 to 0.86 across samples tested to date. Further, tests of its construct validity showed that a perceptual organisation factor on the WAIS accounted for 45 per cent of the HVOT variance, suggesting that the HVOT is a valid test of perceptual organisation.

Figure 11.8 A simulated example of a Hooper Visual Organisation Test item

To assess a person's spatial awareness ability, the Standardised Road-Map Test of Direction Sense (Money, 1976) can be used. In this test, the examiner traces a dotted pathway on a road map with a pencil and asks the test taker to tell the direction (right or left) taken at each turn. Lezak et al. (2012) reported that the road map test is able to distinguish those with parietal-lobe injuries from those with Huntington's or Alzheimer's disease.

Executive functions

Although there are disagreements among clinical neuropsychologists about the definition, nature and number of **executive functions**, it is widely accepted that these functions are primarily mediated by the prefrontal cortex (Chan et al., 2008). In addition, it is agreed that executive functions are responsible for goal-directed behaviours in humans, and that impairments in these functions are debilitating and difficult to rehabilitate. Working memory, concept formation, problem solving and planning are commonly considered executive functions. Because of space limitations, a comprehensive description of tests used to assess executive function is not included in this section. Instead, a discussion of a battery of executive functions—the Delis-Kaplan Executive Function System (D-KEFS; Delis, Kaplan & Kramer, 2001)—is used to illustrate the functions and tasks commonly included in the assessment.

executive functions
higher-level functions considered to be mediated by the prefrontal lobes; responsible for goal-directed behaviours, these functions usually include components such as working memory, concept formation, problem solving and planning

The D-KEFS consists of nine subtests: Trail Making Test, Verbal Fluency, Design Fluency, Color-Word Interference Test, Sorting Test, Twenty Questions Test, Tower Test, Proverb Test and the Word Context Test. Table 11.2 summarises the descriptions of these tests and the functions they measure. One of the strengths of this test is its comprehensiveness. The nine tests of the battery allow an examiner to assess all of the executive functions at the same time, using the same normative data. The size of the standardisation sample was 1750 and included age groups from 8 to 89 years. Although the reliability of the principal scores for the nine subtests is acceptable, some of the additional scores are not as high, and more research is needed to support the validity of the test (Ramsden, 2003).

Table 11.2: Descriptions of tasks and functions measured by the nine subtests of the D-KEFS

Subtest	Description	Function measured
Trail Making Test	This is a modified version of the test from the Halstead-Reitan Neuropsychological Battery. Test taker is required to join circles on a piece of paper.	Flexibility of thinking

Subtest	Description	Function measured
Verbal Fluency	Test taker is required to generate words in a phonemic format from over-learned concepts.	Fluent productivity (verbal)
Design Fluency	Test taker is required to generate as many figures as possible by connecting rows of dots.	Fluent productivity (spatial)
Color-Word Interference Test	This is a modification of the Stroop Color-Word Interference Test. Test taker is required to inhibit an automatic verbal response and generate a conflicting response.	Verbal inhibition
Sorting Test	Test taker is required to sort objects into sixteen different sorting concepts on two sets of cards.	Concept formation, cognitive flexibility and problem solving
Twenty Questions Test	Test taker is required to identify various categories and subcategories represented in thirty objects and formulate abstract, yes/no questions.	Hypothesis testing, abstract thinking and impulsivity
Tower Test	Test taker is required to move discs across three pegs to build a tower in the fewest number of moves.	Planning, reasoning and impulsivity
Proverb Test	Test taker is required to provide a correct abstract interpretation of a proverb.	Ability to generate and comprehend abstract thought and metaphorical thinking
Word Context Test	Test taker is required to discover the meaning of made-up or mystery words based on clues given.	Deductive reasoning and abstract thinking (verbal)

Motor functions

A comprehensive assessment of **motor functions** usually includes lateral dominance, strength, fine motor skills (speed and dexterity), sensorimotor integration and praxis (Mapou, 1995). A person might be able to encode, process, retrieve stimulus information and plan actions, but be prevented from achieving a behavioural goal because of problems with motor functions. As with sensory functions, information about a test taker's motor functions can be obtained from

other sources (e.g. neurological, occupational therapy and physiotherapy examinations). The following tests illustrate how motor functions are assessed.

motor functions
abilities such as lateral dominance, strength, fine motor skills (speed and dexterity), sensorimotor integration and praxis

The hand dynamometer is commonly used to assess motor strength. To obtain reliable scores, the test taker is asked to grasp the handle of the mechanical hand dynamometer as hard as possible three times, alternating between the dominant and non-dominant hand. Average motor strength (in kilograms) for each hand is obtained based on scores of these trials. To measure motor speed, the Finger Tapping Test of the HRNB (see Box 11.3) is commonly used. In this test, the test taker is asked to tap as quickly as possible for 10 seconds on a mechanical device similar to a telegraph key. A counter is fitted to the device for recording the number of taps. To obtain reliable measures, five trials are administered, alternating between the dominant and non-dominant hands, and the average number of taps per trial for each hand is obtained. Reliability data for the Finger Tapping Test are variable, with Lezak et al. (2012) reporting test-retest correlations of between 0.64 and 0.94 for those with brain disorders.

The Purdue Pegboard (Purdue Research Foundation, 1948) is usually used to assess motor dexterity. This test was originally developed to select assembly line workers. In this test, the test taker is required to place metal pins in two rows of holes with the dominant hand, non-dominant hand and both hands, within a 30-second time limit. An assembly trial (time limit = 1 minute) that requires more complex visual-motor coordination can also be administered. This trial requires the test taker to build an 'assembly'; that is, by placing a pin, a washer, a collar and a washer sequentially for each hole. The test taker alternates between the dominant and non-dominant hand. Lezak et al. (2012) again reported variability in test-retest reliability, with correlations from between 0.35 to 0.93 noted. In terms of its ability to predict a lateralised lesion, the test scores on the Purdue Pegboard represent a significant predictive gain over patients' base rate scores. By using appropriate normative data for these tests of motor function, an examiner can determine if the reductions of a test taker in motor strength, speed and dexterity are due to injury on the right, left or both sides of the brain.

Box 11.4

Screening neuropsychological status: the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS)

The RBANS, developed by Randolph in 1998, is a brief screening instrument (around 30 minutes in length) initially designed to assess cognitive decline in the older adult population (in particular mild cognitive impairments) (Randolph et al., 1998). In recent times, however, the RBANS has also been utilised as a brief screening measurement for cognitive functioning in younger patients (e.g. McKay et al., 2008). The RBANS consists of twelve subtests: two measures of attention (Digit Span and Coding), two measures of visuo-spatialconstructional abilities (Figure Copy and Line Orientation), two measures of language (Picture Naming and Sematic Fluency), two measures of immediate memory (List Learning and Story Memory), and four measures of delayed memory (List Recall, List Recognition, Story Recall and Figure Recall). A total of five index scores are computed from this instrument as well as a total global score (Randolph, 1998). The RBANS is expressed as a standard score (based on the normative data of a healthy US adult population aged 20–89 years) with a mean of 100 and a standard deviation of 15.

One of the main advantages of the RBANS is that it is easy to administer, and is able to assess a broad range of cognitive abilities in less than 30 minutes. By comparison, the HRNB can take six to eight hours, which can make the assessment process quite lengthy and fatiguing for older populations and clinical populations. Unlike the HRNB, the RBANS can be utilised by clinicians in a number of situations where a lengthy neuropsychological assessment is not possible or practical (e.g. hospital bedside evaluations, repeated evaluations, home visits etc.) (McKay et al., 2008). The RBANS has been used and validated in a number of clinical populations, including stroke (e.g. Larson et al., 2005), traumatic brain injury (e.g. McKay et al., 2008), dementia (e.g. Randolph et al., 1998), Parkinson's disease (e.g. Beatty et al., 2003), depression (e.g. Faust et al., 2016), and schizophrenia (e.g. Randolph, 1998).

The RBANS has been shown to distinguish between different diagnostic groups (e.g. patients with vascular dementia patients and patients with Alzheimer's; Randolph, 1998). Moreover, the RBANS has also been shown to have good psychometric properties. For instance, the RBANS has been shown to have good test-retest reliability, internal consistency, and construct and predictive validity in patients with schizophrenia (e.g. Gold, Iannone & Buchanan, 1999), in stroke populations (e.g. Larson et al., 2005) and in normal geriatric populations (e.g. Gontkovsky, Beatty & Mold, 2004). Supplementary normative data have also been collected and are available from several sources, including an Australian population (e.g. healthy Australian community dwelling adults; Green et al., 2008).

The main limitation of the RBANS is that, although the effects of age on performance are adjusted for, the effects of education are not (Beatty, Mold &

Gontkovsky, 2003). This is particularly problematic as education level has been found to predict performance in individuals (viz., higher education level predicts higher performance; Duff et al., 2003). However, some independent studies have provided age- and education- corrected normative data (e.g. in normal community-dwelling older adults; Duff et al., 2003).

Case study 11.1

Neuropsychological assessment and the case of Alan Bond

At the end of 1993, in one of Australia's biggest corporate fraud cases, Australia's most dynamic entrepreneur Alan Bond struggled as he faced his interrogators at Perth Magistrates Court. He faced two charges by Australian Securities and Investments Commission relating to failing to act honestly as an officer of Bond Corporation Holdings Ltd, and having the intent to deceive or defraud that Corporation. Both these charges pertained to the sale of a painting by the French impressionist Edouard Manet. Tim Watson-Munro, Mr Bond's psychologist, claimed that he was unfit to run a corner store let alone a large corporation. He also claimed that Mr Bond was incapable of instructing his lawyers. Mr Watson-Munro described Mr Bond in the court as being severely depressed, suffering from a high level of emotional stress, and having suicidal thoughts. Mr Watson-Munro also stated that Mr Bond had suffered brain damage as a result of heart surgery that he had earlier that year, and this (as well as breakdown of his marriage and the collapse of his business corporation) may have aggravated his physical and mental stress. Despite this testimony prompting high scepticism among onlookers and opponents, the magistrate was sympathetic and adjourned the hearing to a new date six months later (July 1994).

During this six-month period, however, it was alleged that Mr Bond was seen in public at expensive restaurants, and making business deals and phone calls at hotels. According to the prosecution, Mr Bond continued to conduct business as normal and they dismissed claims about his mental condition as being timely and convenient. During the nine-day application for a second adjournment in May 1994, scans of Mr Bond's brain were shown to the court by his medical team, which consisted of neurologists, forensic psychologists and nuclear physicians. They all agreed that Mr Bond's condition of depression, memory loss and stress was a result of symptoms that stemmed from Bond's heart surgery the previous year; namely, brain damage caused by the release of a tiny piece of tissue or gas into the bloodstream. Sitting in the public gallery

during the application hearing, Mr Bond was shaking and disorientated to the point where he had to be taken away to hospital by his lawyer and psychiatrist, as he was allegedly too ill to cope with proceedings. While the court found that Mr Bond did in fact have evidence of minor brain damage and 'reactive depression', it was not sufficient to adjourn the committal hearings for a further six months. Mr Bond was required to face the six-week proceedings at the Perth magistrate in July that year.

In March 1995, Mr Bond was released from bankruptcy and handed over a cheque for $1 million to his creditors (less than 1 cent in the dollar). He agreed to pay them a further $750,000 a year for the next three years. In 1996, Bond was convicted of four corporate fraud charges under the Western Australia Companies Code and was handed a three-year prison sentence, despite the strong pleas of his defence team that jail-time could be fatal for Mr Bond as he was a very sick man. In 1997, Mr Bond was further sentenced for four years for defrauding his own company, Bell Resources. In 2000, Mr Bond was released from jail.

Discussion questions

1. If you were Alan Bond's psychologist, what test(s) would you use to assess him? Provide justifications for your answer(s). In addition, what other information would you need to collect for this case?
2. Why do you think the court did not approve the second application for adjournment for Mr Bond? Do you agree with the court's decision? Why or why not?
3. What are some of the key issues discussed in this chapter that you think can apply in the case of Mr Bond?
4. Answer the above questions again after you have read Chapter 12.

Practitioner profile

Dr Jan Ewing

1. **How long have you been a psychologist?**
   I graduated from my Master's degree in 1975 and began practising as a psychologist that year, so I have now been a psychologist for 41 years.

2. **What is your specialisation and how did you get the training and experience to do this job?**
   I specialise in two areas and have sub-specialities in both fields. Firstly, I am a clinical psychologist who specialises in the treatment of post-traumatic syndromes, particularly combat-related trauma and adult survivors of childhood abuse and neglect. Secondly, I am a clinical neuropsychologist who specialises in medico-legal assessment of traumatic brain injury.

   I completed a Master's degree (course work, clinical placements and thesis) in clinical psychology from University of Melbourne. I then worked for four years in several settings

with a strong demand for the then newly emerging skills of neuropsychological assessment and treatment. Realising I needed more specialist training in that field, and there being no doctoral training in clinical neuropsychology in Australia at that time, I completed a PhD in clinical neuropsychology (course work, clinical placements and thesis) at the University of Victoria in British Columbia, Canada. I returned to Australia and have worked in both specialities since that time, with varying emphases depending on setting. I have continued to educate myself in my areas of speciality ever since with regular workshop training and seminar attendance. For example, I now have further certification in a range of therapeutic techniques (e.g. Eye Movement Desensitization and Reprocessing [EMDR], Hypnosis, Internal Family Systems [IFS]) and have gained experience as a neuropsychologist in a wide variety of settings including neurology/neurosurgery and psychiatric wards, rehabilitation units, a multiple sclerosis unit, a learning difficulties clinic and private practice. Forty-one years later, I am still learning.

3. What kind of clients and referrals do you usually get?

While I accept a wide range of referrals, my areas of specialisation are well known so that most client referrals reflect those areas. Therapy referrals tend to be for treatment of adult-onset or developmental trauma, usually many years after the incident/s. I have a particular interest in and specialised training in dissociative disorders and most referrals have reflected that speciality in recent years. My neuropsychological referrals tend to come from lawyers requesting a comprehensive medico-legal assessment of clients who have sustained personal injuries, including closed head injuries. While the whole range of severity is included in these referrals (from no brain injury through to severe brain injury), the majority of these involve mild to moderate head injuries and are often complicated by concurrent psychological symptoms.

4. Do you use psychological tests in your practice?

Yes, regularly. For therapy clients I use a range of diagnostic and progress monitoring tests. For neuropsychological clients, I follow a hypothesis-testing approach, the tests administered vary from case to case. However, most of my neuropsychological assessments in private practice relate to closed head injuries associated with diffuse effects and a comprehensive assessment of all areas of the client's functioning is requested, both to identify impairments attributable to the injury and also to allow for predictions of future occupational and personal functioning. Hence, I generally administer a relatively standard battery of tests that provide a broad picture of the client's abilities across various domains.

5. Why do you use psychological tests and in what way do they help you in your practice?

Psychological tests provide an objective tool with which to identify areas of abnormality (and avoid confirmatory bias or failure to detect subtle symptoms), to quantify the pattern and severity of those symptoms and to generate hypotheses for further investigation. With therapeutic clients, they generate questions for further enquiry and opportunities for further exploration that might otherwise be missed. With clients referred for formal evaluation, the identification of questionable symptom and/or performance validity is also essential, especially in the medico-legal context, and can only be reliably assessed with specifically designed tests.

6. In your opinion, what is the future for psychological testing in your specialisation?

In the medico-legal setting, the neuropsychologist's task is to assist the court to make a judgment regarding the extent of the client's impairments, the extent to which these impairments are attributable to the index injury, and their likely impact on his/her daily life

and employment potential. The courts rely on objective findings wherever possible and, hence, neuropsychological testing in this area is likely to remain important despite advances in imaging techniques, which provide little information regarding the impact of the injuries in the individual case. Future tests will hopefully provide more sensitive, specific and reliable tools for the identification of subtle impairments, tests that are able to better determine psychological versus organic patterns and, perhaps most importantly, more ecologically valid tests that allow greater confidence in our task of predicting the manner in which the client's impairments will compromise their daily functioning.

Chapter summary

Injury to the human brain can lead to long-term and significant disability for an individual. Neuropsychological assessment is an essential step for the management and treatment of individuals suspected or found to have brain injury. In this chapter we outlined the purposes for and the steps in neuropsychological assessment. We also discussed the functions commonly examined during a neuropsychological assessment and described some of the commonly used psychological tests that measure these functions. In so doing, we introduced you to one of the fastest growing sub-branches of psychology.

Questions

1. Psychological tests are different from neuropsychological tests. Do you agree?
2. Discuss the function(s) of the following brain structures:
    - **a.** cerebellum
    - **b.** thalamus
    - **c.** frontal lobes
    - **d.** basal ganglia.
3. What does a clinical neuropsychologist do?
4. What is neuropsychological assessment and what are the steps of a neuropsychological assessment?
5. What functions are measured by the following tests?
    - **a.** Halstead-Reitan Neuropsychological Battery
    - **b.** Stroop Color Word Interference Test
    - **c.** Rey-Osterreith Complex Figure Test
    - **d.** Purdue Pegboard

Further reading

Hebben, N & Milberg, W (2009). *Essentials of neuropsychological assessment* (2nd ed.). New York, NY: Wiley.
Kolb, B & Whishaw, I Q (2015). *Fundamentals of human neuropsychology* (7th ed.). New York, NY: Worth.
Lezak, M D, Howieson, D B, Bigler, E D & Tranel, D (2012). *Neuropsychological assessment* (5th ed.). New York, NY: Oxford University Press.
Otero, T M, Podell, K, DeFina, P & Goldberg, E (2013). Assessment of neuropsychological functioning. In J R Graham & J A Naglieri (Eds.), *Handbook of Psychology: Vol 10, Assessment psychology* (pp. 502–33). Hoboken, NJ: John Wiley & Sons.
Parsons, M W & Hammeke, T A (Eds.). (2014). *Clinical neuropsychology: A pocket handbook for assessment* (3rd ed.). Washington, DC: American Psychological Association.

Useful websites

APS College of Clinical Neuropsychology (Australian Psychological Society): www.groups.psychology.org.au/ccn
Clinical neuropsychology: Areas of specialisation (Australian Psychological Society): www.psychology.org.au/community/specialist/clinicalneuro/#s2
International Neuropsychological Society: www.the-ins.org
National Academy of Neuropsychology: www.nanonline.org
Society for Clinical Neuropsychology (Division 40 of the American Psychological Association): www.scn40.org

12 Forensic Psychological Testing and Assessment

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. define forensic psychology and describe what forensic psychologists do
2. list the settings where forensic psychological testing and assessment usually take place and give examples of assessment in these settings
3. explain the similarities and differences between forensic psychological testing and assessment and therapeutic psychological testing and assessment
4. give examples of the commonly used tests for forensic psychological testing and assessment and discuss their strengths and weaknesses
5. discuss some of the main issues relating to forensic psychological testing and assessment.

KEY TERMS

competency to stand trial
custody evaluation
expert witness
forensic psychological testing and assessment
forensic psychology
malingering
risk assessment
therapeutic assessment

Setting the scene

- While shopping at a supermarket, a young man slipped, fell and injured himself. His solicitor referred him to a psychologist for an assessment to determine the extent of damage caused by the injury for the purposes of compensation.
- A psychologist employed by corrective services was asked to assess a new inmate. Specifically, she was asked to determine if this person was suffering from a mental disorder.
- A judge at the Family Court ordered the parents of a young girl to be assessed by a psychologist. Results of the assessment were used to assist the resolution of the custody dispute.
- A young woman who was involved in a car accident complained that her ability to remember had dramatically decreased since the accident. Because the extent of her complaint was at variance with the severity of the accident and neurological findings, the possibility of her exaggerating her problem was raised by the insurance company.
- An inmate was referred for psychological assessment to gauge the risk of reoffending and to make treatment recommendations to reduce the risk of recidivism.

Introduction

The origin of the word 'forensic' can be traced to the Latin word *forensis*, which means 'of the forum' where the law court of ancient Rome was held. The *Australian Oxford Dictionary* defines the word 'forensic' as 'of or used in connection with courts of law'. **Forensic psychology** is a recently emerged branch of psychology that specialises in applying psychological knowledge and skills to the working of the legal and criminal justice systems. As illustrated by the above examples, forensic psychologists typically provide assessment services to clients of the legal and criminal justice systems to answer referral questions relating to diagnosis, decision making and prediction. In this chapter, we provide an introduction to forensic psychological testing and assessment by addressing the following questions: What is forensic psychology? What do forensic psychologists do? What are the main settings of forensic assessment? What are the similarities and differences between forensic and therapeutic psychological assessment? What are the common psychological tests and assessment techniques used by forensic psychologists? What are some of the issues and limitations relating to forensic psychological testing and assessment?

forensic psychology
a branch of psychology that specialises in the application of psychological knowledge and skills to the working of the legal and criminal justice systems

Forensic psychology and forensic psychological testing and assessment

As a branch of applied psychology, forensic psychology is relatively young. In the UK, the term 'forensic psychology' was introduced by Haward in 1953 to address the County Durham Psychology Group (Gudjonsson & Haward, 1998). In the USA, although psychologists have been asked to appear in courts as expert witnesses since the 1900s, forensic psychology was not formally recognised as a speciality area of psychology by the American Psychological Association until August 2001 (Heilbrun et al., 2000; Ogloff & Douglas, 2003). In Australia, the College of Forensic Psychologists was established by the Australian Psychological Society in 1993.

Broadly speaking, forensic psychology can be defined as the application of psychology to the legal system (Ogloff & Douglas, 2003). Heilbrun et al. (2000) put forward a more specific definition:

> Forensic psychology will be defined as the professional practice by psychologists within the areas of clinical psychology, counselling psychology, neuropsychology, and educational psychology, when they are engaged regularly as experts and represent themselves as such, in an activity primarily intended to provide professional psychological expertise to the legal system.

Testifying in court as **expert witnesses** (see Box 12.1), providing psychological treatment to offenders and victims of crime, and conducting research on the accuracy of testimony of witnesses are some of the domains of forensic psychology. However, one of the major contributions of forensic psychology is the provision of **forensic psychological testing and assessment** (Ackerman, 2010). (For ease of expression, the term 'forensic psychological testing and assessment' is shortened to 'forensic assessment' in this chapter.) The primary purpose of forensic assessment is the collection of relevant and useful data and information with psychological tests and other assessment techniques to assist decision makers in the legal and criminal justice systems to make decisions about offenders or those suspected of an offence (Ogloff & Douglas, 2003). Psychologists in other speciality areas are sometimes engaged in this work, but they can be considered to be conducting forensic assessment, and need to follow the guidelines and ethics for practice in this speciality. Furthermore, to practise in this area, they are required to have training and experience in the law that is relevant to the particular area of practice, be it clinical, organisational, counselling, neuropsychological or educational.

experts witness
someone who can or is required to provide factual information as well as an opinion, based on their background and training in a court of law

forensic psychological testing and assessment
the collection of relevant and useful data and information using psychological tests and other assessment techniques to assist professionals in the legal and criminal justice systems to make decisions about offenders or those suspected of an offence

Box 12.1

Forensic assessment and psychologists as expert witnesses

Before the results of forensic assessment are actually presented in a court of law as expert witness evidence, it has to be decided: (a) whether this evidence is really necessary; and (b) whether the evidence is admissible under the requirements of the court. The practical and legal criteria relating to these decisions can differ between countries and between states in the same country. According to Ogloff and Douglas (2003), the results of forensic assessment are needed if they are found by the court to be relevant and related to one or more legal standards raised by the case. In addition, the court needs to weigh up the relevance and utility of the evidence being presented (its probative value) against its potential to bias the jury (its prejudicial value). In deciding whether the evidence is admissible, three requirements must be satisfied: (a) the evidence is required by the judge or the jury to assist in decision making; (b) the person who provides the evidence must be suitably qualified; and (c) if the expert witness uses scientific facts or data, they must be widely accepted by other experts in the area (Ogloff & Douglas, 2003).

While other witnesses in a court case are required to provide factual information, expert witnesses may provide factual information as well as offer an opinion. In the USA, some specific criteria (known as the Daubert Criteria) have been developed based on the Daubert v Merrl Dow Pharmaceutical case (Ackerman & Kane, 1998). These require the psychologist to: (a) use psychological tests or assessment techniques that are theoretically and psychometrically sound; (b) draw conclusions based on scientifically validated theory; (c) weigh and qualify testimony based on theory and empirical research; and (d) know how to defend the scientific basis of the procedure used.

To assist the selection of psychological tests to fulfil these criteria, Heilbrun (1992) suggested the following specific guidelines:

1. Use commercially available tests that are adequately documented in at least two sources (e.g. a test manual and the *Mental Measurements Year Book*).
2. Unless there are justifiable reasons or explanations, use tests with reliability coefficients of at least 0.80.
3. Use tests that are directly relevant to the legal issue involved or at least use tests that assess psychological constructs that are relevant to the legal issue.
4. Make sure that tests are administered based on standardised instructions using materials or stimuli provided by the test publisher in an optimal testing environment (e.g. quiet, well lit and free of distraction).
5. Make sure that tests chosen are applicable or suitable (in terms of age, gender, ethnic and educational background) to the person being assessed.
6. If possible, select tests that provide formulae for making objective, actuarial conclusions or predictions.
7. If possible, assess the response style of the person and interpret the psychological test results of that person in light of this finding.

As applies to other fields of professional psychology mentioned in the earlier chapters of this book, training of forensic psychologists is usually reserved for the postgraduate level. In Australia, a minimum of six years of full-time university training, including two years of specialised postgraduate study and supervision, is the minimum requirement for membership of the College of Forensic Psychologists of the Australian Psychological Society. Most of the forensic psychology training programs in the USA are typically at the doctoral level (i.e. PsyD or PhD).

Settings of forensic assessment

In Australia and other Commonwealth countries, three jurisdictions are generally recognised: criminal, civil and family. Criminal law is concerned with crimes against the public or the Crown, civil law with the resolution of conflicts between

individuals or organisations, and family law with conflicts within families or between partners in married or de facto relationships. Apart from being employed to support courts in these jurisdictions, forensic psychologists are employed in other settings such as police departments, correction centres, corrective and forensic mental health services, private practices and research organisations.

Within these settings, forensic assessment is conducted for a number of purposes. For example, in the criminal law area, results of forensic assessment have been used by the defence, the prosecution or the court for pretrial, pre-sentenc cide if a defendant is competent to stand trial. In the civil law area, forensic assessment has been requested to establish the extent of personal injury (e.g. neurocognitive impairment as a result of a car accident or emotional harm as a result of a traumatic event such as a bank robbery or an assault), to determine the effect of an unfair dismissal, and to determine the capacity of individuals in making financial decisions or changing the content of a will. In the family law area, results of forensic assessment have been used to assist in deciding custody of and access to children, and removing children from the care of parents.

The results of forensic assessment can have significant and long-term impacts on the lives of the persons who are assessed and on the lives of those around them (Martin, Allan & Allan, 2001). This underscores the importance of advanced-level training and experience for psychologists working in the forensic area and highlights the responsibilities that come with conducting forensic assessment.

Differences between forensic and therapeutic assessment

Forensic assessment is considered by some (e.g. Melton et al., 1997) as specialised clinical psychological assessment because it requires the training and advanced knowledge and skills similar to that received by clinical psychologists. However, others (e.g. Greenberg & Shuman, 1997; Heilbrun, 2001) contend that, unlike other branches of psychology (e.g. clinical, counselling and neuropsychological) that are therapeutic in nature, forensic assessment is different from these disciplines in a number of aspects (see Table 12.1).

Table 12.1: Differences between forensic and therapeutic assessment

	Therapeutic	Forensic

	Therapeutic	Forensic
Purpose of assessment	Diagnosis and treatment of psychological problems	Assist decision makers in legal/criminal justice system
Psychologist–client relationship	Helper and client or patient	Objective or quasi-objective professional stance
Who is being served?	Individual client	Variable: may include the individual client, the lawyer and the court
Notification of purpose of assessment	Psychologist and client are assumed to share a similar purpose; formal, explicit notification is not necessary	Formal, explicit notification of purpose is necessary because psychologist and client do not necessarily share similar purpose
Nature of standard being considered	Medical, psychiatric and psychological	Medical, psychiatric and psychological but also legal
Source of data	Self-report, psychological tests, behavioural assessment and medical	Self-report, psychological tests, behavioural assessment and medical, but also other information (e.g. files and observations)
Response style of client	Assumed to be reliable	Not assumed to be reliable
Clarification of reasoning and limits of knowledge	Assumed and optional	Important
Written report	Comparatively brief and focused on conclusions	Lengthy and detailed; need to document findings, reasoning and conclusions
Court testimony	Not expected	Expected

The primary purpose of forensic assessment is to assist decision makers in the legal or criminal justice systems to address specific legal issues, such as whether a defendant is competent to stand trial or the risk of managing an inmate in a certain way. The primary purpose of **therapeutic assessment**, in

contrast, is to diagnose and treat clients with psychological or mental problems. The referral question for therapeutic assessment arises out of the needs of a client, but the referral question for forensic assessment is based on legal criteria for decision making. While the process of therapeutic assessment is usually (but not always) a means to an end (i.e. treatment), the process of forensic assessment is commonly an end to itself. Typically, the process of forensic assessment terminates after relevant information or data are collected to prepare a report that addresses a specific legal issue.

therapeutic assessment
an assessment conducted by psychologists with the purpose of assisting and treating a client

In terms of relationships between the psychologist and client, therapeutic assessment is similar to that between a doctor and a patient (i.e. between a helper and someone who seeks help). As such, it is assumed that in this relationship the psychologist will act with the best interests of the client in mind and that the client will voluntarily and truthfully provide the information required by the psychologist in order to be helped. In contrast, forensic assessment is different from its therapeutic counterpart in that the psychologist does not assume the role of a helper for a client. Instead, she adopts a more objective stance during the assessment and will neither accept nor reject the information or data provided by the person she assesses until they can be checked and validated.

For therapeutic assessment, the 'client' is the person who seeks help from the psychologist. The 'client' for forensic assessment, on the other hand, may be more than one person. Usually the 'client' is the decision maker or the person (e.g. a judge or a lawyer) in the legal or criminal justice system who referred the person to be assessed rather than the person who is assessed. Because of this difference, the purpose of forensic assessment is usually formally or explicitly explained to the person who is being evaluated before the assessment is done. Formal or explicit explanation of the purpose is not usually necessary for a therapeutic assessment because of an implicit understanding between the client and the psychologist (i.e. diagnosis and treatment of a problem).

The two types of assessment also differ in terms of the nature of the standards used. For therapeutic assessment, a psychologist is guided by scientific and professional standards. While these two standards are relevant and important for forensic assessment, legal standards need to be taken into consideration as well. The selection of assessment procedures can be used to illustrate this point. In therapeutic assessment, a psychologist can decide what psychological tests or assessment techniques to use for a client based on the referral question. In forensic assessment, the choice of an assessment technique needs to be

considered in the light of relevant legal standards to ensure that the constructs being measured and the test instruments bear on the legal standards.

Self-report during an interview, psychological test results and medical histories are usually used as data in both therapeutic and forensic assessment. Additional data commonly used in forensic assessment includes information recorded on legal files and observations made by personnel who work in legal or corrective settings. This additional information may be more directly related to the legal issue involved. Furthermore, this information is important for double-checking the accuracy of the other data collected. This is because, in forensic assessment, a psychologist does not automatically assume that the responses of the individual referred for assessment are accurate. The individual may want to exaggerate or minimise the extent of his problems or symptoms in order to gain a favourable outcome in his case.

The findings, reasoning and conclusions of a therapeutic assessment are not usually subjected to strict clarification or challenge because of an implicit acceptance of the professional expertise of the psychologist who conducted the assessment. Moreover, psychological reports written for therapeutic assessment are normally not expected to be brought to a court of law. In contrast, reasoning and findings of forensic assessment reports are expected to be scrutinised and challenged because of the adversarial nature of the legal system. As such, forensic assessment reports are comparatively longer and more detailed than is the case with therapeutic reports.

Case study 12.1

Psychologists in the Family Court

In the Australian legal system, psychologists play an important role in the Family Court. Following a break-down of marriage, where the couple cannot negotiate a mutually satisfactory arrangement with respect to child custody, they may approach the Family Court for a settlement of the matter. These are difficult situations in which emotions run high and recollection of events is fallible. The Court can ask a psychologist to provide an evaluation.

Wilmoth (2007) reviewed the role of the psychologist in the Family Court. The Court will provide terms of reference for the psychologist but there are often matters of motivation of both parties that need to be evaluated along with the parenting skills of each parent and the needs of the child. Interview is often the major method of assessment, but some forms of psychometric testing can assist in making normative comparisons. The task of the psychologist is not to

act as a judge but to provide an opinion to the court on what is in the best interest of the child.

Since the publication of Wilmoth's article, a set of guidelines for Family Court assessments have been published: *Australian Standards of Practice for Family Assessments and Reporting—February 2015* (Family Court of Australia, Federal Circuit Court of Australia & Family Court of Western Australia, 2015). Their purpose is to describe best practice by professionals in developing and reporting family assessments. The guidelines cover areas such as arranging the assessment, communicating with the parties, conducting assessment, formulating opinions, taking cultural issues into consideration, writing reports, notifying risk of harm, and recording and storing information.

Discussion questions

1. Why do think this area of practice is so difficult?
2. What matters are relevant to the opinion the psychologist must prepare in the case of a child custody settlement?
3. What forms of personal bias should the psychologist be alert to in making evaluations of this kind?

Psychological tests and assessment techniques commonly used in forensic assessment

According to Heilbrun, Roger and Otto (2002), three types of assessment techniques can be used in forensic assessment: forensic assessment instruments, forensically relevant instruments and clinical instruments. The first type is specifically designed for forensic assessment and these instruments are directly relevant to a specific legal standard. For example, the MacArthur Competence Assessment Tool–Criminal Adjudication (Poythress et al., 1999) was specifically developed to assess the US legal standards for competence to stand trial. The second type is not designed based on any specific legal standards, but the constructs measured by these instruments are related to a legal standard. Examples of this type of assessment technique include tests that measure constructs such as psychopathy, violence risk or malingering (the notion of malingering is discussed later in the chapter). The third type includes psychological tests or techniques that are not developed specifically for the purpose of forensic assessment but have been adopted by forensic psychologists to answer legal questions. Examples of these instruments include the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV), Minnesota Multiphasic Personality Inventory–Second Edition (MMPI–2) and the Beck Depression Inventory–Second Edition (BDI–2). In previous chapters of this book, we have

discussed quite a number of the third type of assessment techniques. Therefore, in this chapter we will focus on the first two types of techniques such as competency to stand trial, risk assessment, custody evaluation and malingering.

Competency to stand trial

In the area of criminal law, the issue of whether a defendant is competent to stand trial is an important but difficult one. This issue is based on the assumption that it is unfair to put someone on trial if he does not have the ability or capacity to understand the matters brought against him. This reduced ability or capacity could be due to intellectual handicap, mental illness, cognitive decline or psychological trauma. Compared with other requests for assessment of competency (e.g. competency to make financial decisions, change a will or consent to medical treatment), the number of requests for **competency to stand trial** is comparatively high. In the literature, a number of tests and techniques have been developed to assess a person's capacity to stand trial (Ackerman, 2010). In this section, we only review the Competency Screening Test (CST; Lipsitt, Lelos & McGarry, 1971) and the MacArthur Competence Assessment Tool–Criminal Adjudication (MacCAT–CA; Poythress et al., 1999) because of space limitations.

competency to stand trial
an assessment of whether a defendant is able to stand trial because his/her mental state was affected at the time of the offence or at the time of the trial

As the name suggests, the CST is a screening device used to decide if a more comprehensive assessment is necessary for defendants who may be unfit to stand trial. It comprises twenty-two unfinished sentences that the test taker has to complete and administration usually takes about 25 minutes. Basically, the sentences cover three areas of the legal/judicial processes: relationship between a client and a lawyer; a client's understanding of the processes of the court; and the ability of a client in dealing emotionally with the criminal processes. The responses of the test taker on the items are scored using a three-point scale (0, 1 and 2) depending on their appropriateness. The CST takes about 15 to 20 minutes to score and its score can range from 0 to 44. Randolph et al. (1982) reported a high level of inter-rater reliability (0.92) for the CST. The same authors found significant correlations between scores on the CST and opinions of court psychiatrists. Despite these positive findings, the CST has been criticised on the grounds that the sentence completion procedure and scoring method are not well justified, the construct(s) it assesses may not be directly related to the legal

standard of competency to stand trial, and it leads to relatively high false positive (indicating competent defendants as incompetent) and false negative rates (indicating incompetent defendants as competent) (Ackerman, 2010).

The MacCAT-CA was developed based on Bonnie's (1992, 1993) theory of legal competency and is an update of the MacSAC–CD (MacArthur Structured Assessment of Competencies of Criminal Defendants). It is an individually administered instrument intended for use with criminal defendants and takes about 25 to 55 minutes to administer. It comprises twenty-two items that are related to the formal functional abilities associated with the legal construct of competency to stand trial. These items include three discrete competence scales: understanding (eight items), reasoning (eight items) and appreciation (six items). A brief vignette describing a hypothetical crime is used for a test taker to respond to items for the first and second scales. The first scale covers the ability to understand general information related to the law and adjudicatory proceedings. The second scale covers the ability to discern the potential legal relevance of information and capacity to reason about specific choices that confront a defendant in the course of adjudication. The third scale covers the rational awareness of the meaning and consequences of the proceeding in the defendant's own case. Each item is rated on a three-point scale (0, 1 and 2) and a high score indicates a high level of capacity.

The MacCAT–CA was validated on a sample of 729 criminal defendants (90 per cent males) in the USA. Among them, 197 were competent, 249 were competent but receiving treatment for mental health disorders, and 283 were incompetent because of mental illness. Norms are not available for individuals with IQs under 60. In terms of reliability, the internal consistencies of the three scales for the standardisation sample were 0.81, 0.85 and 0.88, respectively. Inter-rater reliability for the three scales was reported as 0.75, 0.85 and 0.90. In terms of validity, the construct validity of the MacCAT–CA has been supported by expected patterns of correlations with measures of cognitive ability and psychopathology, and ratings of experienced clinicians. Despite these results, Rogers et al. (2002) cautioned that the MacCAT–CA items are vulnerable to faking by the test taker.

Risk assessment/prediction of aggression or dangerousness

Within the legal and criminal justice context, forensic assessments are increasingly used to assist in predicting the risk of future offending behaviour (i.e. recidivism) and identifying offender treatment needs to guide decision-making processes such as sentencing, parole, classification and treatment provision (Gottfredson & Moriarty, 2006). There are, in general, two methods of **risk**

assessment/prediction: one based on clinical-psychological judgment and one based on an actuarial formula. Clinical-psychological methods rely on the knowledge and experience of professionals to inform risk classifications, utilising techniques and instruments common with clinical psychological practice (e.g. clinical interviews and psychometric measures). Clinical-psychological methods are still commonly used, although research indicates that predictions of risk derived from these methods are frequently inaccurate and invalid (Lowenkamp, Latessa & Holsinger, 2006). Such assessments are subjective and involve decision-making processes that are difficult to observe and replicate (Bonta, 1996).

risk assessment
an assessment conducted to determine how risky or dangerous an inmate is for the purpose of sentencing, parole or classification

Actuarial methods are based on psychometric tools that are statistically developed by identifying those factors in the research literature that are most strongly correlated with the offending behaviour in question (e.g. violence and sexual offending). Important predictor variables for future criminal behaviour that have been identified and used in many actuarial tools include criminal history, education or employment, significant relationships in an offender's life, antisocial relationships, alcohol/drug use and abuse, mental health issues, attitudes, orientation and cognitive processes. Actuarial risk assessment scales generally consist of checklists of these predictor variables that are statistically scored for offenders, with higher scores generally representing a greater risk of recidivism.

Risk assessment predictions derived from actuarial methods have been found to significantly outperform human judgment in terms of accuracy and reliability (Gottfredson & Moriarty, 2006; Hanson, 2005; Hanson & Morton-Bourgon, 2009; Ogloff & Davis, 2004; Upperton & Thompson, 2007; Wormith et al., 2007). Actuarial methods provide a standardised process for making risk predictions that are less prone to biases in human judgment and decision making. However, clinical-psychological methods continue to provide some benefits to risk predictions, given that professionals can make use of information not readily available to actuarial instruments (e.g. professional experience, situational/environmental factors and demeanour during interview). It is increasingly common in forensic assessment to use a combination of clinical-psychological and actuarial methods to inform risk assessments. It must be noted that there is a range of methodological issues in the development and use of actuarial tools that may impact on professional practice in forensic settings (e.g. accuracy of prediction, base rates, and static and dynamic risk factors), although discussion of these issues is beyond the scope of this chapter. Interested readers

are directed to Gottfredson and Moriarty (2006) for an overview of methodological issues impacting on actuarial risk assessment.

There is a growing number of actuarial risk assessment tools that have been statistically developed to assist in making predictions about the probability of future offending behaviour. Table 12.2 lists some examples of risk assessment tools commonly used in forensic assessment.

Table 12.2: Examples of commonly used risk assessment tools

Risk assessment tool	Description
Violence Risk Appraisal Guide (VRAG, Quinsey et al., 1988)	• Clinician-scored instrument for the prediction of violent recidivism among adult convicted offenders • Considers a wide range of variables including age, marital status, criminal history, Psychopathy Checklist score, performance on conditional release, victim injury and gender, history of alcohol problems and psychiatric diagnoses, and developmental factors such as school problems and separation from parents
Static-99 (Hanson & Thornton, 1999)	• Brief ten-item clinician scored instrument designed for use with adult male sexual offenders at the time of release into the community (the instrument can be viewed at www.static99.org) • One of the most widely used instruments across English-speaking countries with high predictive accuracy
Youth Level of Service Inventory/Case Management Inventory (YLS/CMI; Hoge, 2005)	• Combined risk/needs assessment and case-management instrument designed for use with offending or at-risk youth aged 12 to 17 years • Both clinician scored and includes semi-structured interview • Considers variables including prior and current offences, education, substance abuse, family, personality/behaviour, peers, leisure/recreation and attitudes/orientation
Juvenile Sex Offender Assessment Protocol–Version II (J-SOAP-II; Prentky & Righthand, 2003)	• Clinician-scored instrument for use with juvenile sex offenders consisting of two static scales (factors not amenable to change)—Sexual Drive/Preoccupation Scale and Impulsive-Antisocial Behaviour Scale—as well as two dynamic scales (factors amenable to change): Intervention Scale and Community Stability/Adjustment Scale • Each scale has a number of items and higher scores are thought to be associated with increased risk of reoffending

In this section, we describe one instrument for risk assessment/prediction of aggression or dangerousness in more detail. One good predictor of violence recidivism is psychopathy. The Psychopathy Checklist–Revised Second Edition (PCL–R; Hare, 2003) was developed by Canadian forensic psychologist Robert Hare to assess psychopathic (antisocial) personality disorders in adult forensic populations. This widely used, individually administered rating scale comprises twenty items that cover a wide range of psychopathic traits and behaviours. To rate these items, a semi-structured interview (of about 90 to 120 minutes) and a

review of collateral information (of about an hour) needs to be conducted. Ratings on the twenty items using a three-point scale (0 = absent, 1 = possible or to some degree, and 2 = present) produce a total score that ranges from 0 to 40. This score provides an overall assessment of psychopathy or the degree of match to the prototypical psychopath (cut-off score of 30). Two factor scores can also be derived from the ratings: the callous, selfish, remorseless use of others and a chronically unstable and antisocial lifestyle. Norms are available (for male offenders, female offenders and male forensic patientse).

The PCL–R has gained popularity in the forensic area and much research has been conducted to support its utility. The PCL–R is considered the 'gold standard' in predicting violence and recidivism because it has been found to have very good psychometric properties (Acheson, 2005; Ackerman, 2010; Martin, Allan & Allan, 2001). Its internal consistency has been found to be high based on data obtained from male offenders, with 0.85 for the total score and 0.64 to 0.71 for the two factor scores. Inter-rater reliabilities have been found to range from 0.84 to 0.93 for data obtained from male offenders and from 0.93 to 0.97 for female offenders. In terms of validity, the PCL–R has been found to be a very good predictor of many problem behaviours. However, it is rather time-consuming to administer, score and interpret, and, according to its developer, competent use of the instrument requires a high level of training (Hare, 1998).

Custody evaluation

In the area of family law, one of the most difficult decisions for a judge in cases of divorce, abuse or neglect, or guardianship is to determine who should have custody of a child and what provisions, if any, should apply to the custody. Forensic assessment is frequently requested by decision makers in family courts in Australia and other parts of the world to assist in making decisions about child custody (Ackerman, 2010; Powell & Lancaster, 2003), a process known as **custody evaluation**. According to the *Australian Standards of Practice for Family Assessments and Reporting—February 2015* (Family Court of Australia, Federal Circuit Court of Australia & Family Court of Western Australia, 2015), forensic assessment in this area is 'an independent, professional forensic appraisal of the family, done from social science and non-partisan perspective'. It has 'the functional value of contributing to informed and child-centred decisions'. Importantly, the assessment should include assessment of any risk factors and family violence (where such concerns are expressed).

custody evaluation
an evaluation conducted to determine in cases of divorce, abuse or neglect or guardianship which parent should have custody of a child

One of the most obvious psychological assessment techniques for use in child custody cases arising out of parental divorce is the interview (Ackerman, 2010). Usually the child or the children involved are interviewed separately from the adults and different sets of questions are used for these two groups. During an interview with adults, the areas typically covered include demographics, place of residence, current marital situation and marital history, place of employment, current employment and employment history, educational history, names and ages of children, and whether they are living at home, history of medical and psychiatric problems, alcohol and drug use, problems with the law (including sexual abuse or sexual assault), problems with developmental milestones (both parents and children), current life circumstances (including stressors) and functioning. Questions asked during an interview with children depend on the age of the child. In general, questions used during the interview can cover areas such as the child's reaction to the divorce, the child's perception of his or her role in the divorce, the child's view of the parents during the divorce process, how the divorce has affected the relationships between the child and the parents—and the child and his or her siblings (if any)—the impact of the parents' new social life or relationships on the child, who disciplines the child at home and how it is done, and the level of involvement of each parent in the family and in family activities.

Apart from the interviews, a number of psychological tests and assessment instruments may be used to assess the general cognitive ability and personality of the parents and the children involved in a custody evaluation. Because most of these tests (e.g. WAIS–IV, WISC–V, MMPI–2, etc.) have been reviewed early in this text, in the remaining part of this section we focus on an assessment instrument specifically developed for child custody.

One of the commonly used instruments in the USA and Canada is the Ackerman-Schoendorf Scales of Parent Evaluation of Custody (ASPECT; Ackerman & Schoendorf, 1992). This instrument was designed to directly evaluate the suitability of a parent for custody based on characteristics that are related to fitness of custody as identified in the literature. Basically, the ASPECT requires the assessor to respond to fifty-six questions using a yes/no format. The assessor's responses are based on information collected from a parent questionnaire, interview and observation of each parent with and without the child, scores obtained from tests routinely used for child custody evaluation, and the results of an IQ assessment of the child. Responses are collated and an overall index called the Parental Custody Index (PCI; T score with a mean of 50 and standard deviation of 10) is obtained for each parent. Recommendations can be made by comparing the index for each parent. T-score difference of 10 points or more are considered significant and interpretable. The PCI of the ASPECT has been found to have adequate internal consistency (0.76 based on 200 participants) and high inter-rater reliability (0.96). There are also some data that

support its predictive validity (Ackerman, 2010). Nevertheless, concerns have been raised about some ASPECT items that do not seem to be related to custody outcomes and that some factors related to custody decisions were not included in the instrument (Ackerman, 2010).

Malingering

It is assumed in psychological testing and assessment that when a test taker fills in a self-report measure or completes an objective test, the result is a true reflection of their thoughts and feelings or their ability. In some cases this assumption may be false. Clients might want to present themselves in a negative or positive manner. For example, a person who is given a personality test might want to endorse items that are contrary to his behaviour or belief because he wants to present himself in a positive light to the psychologist or the referral agents. This problem has been found to be more common in the forensic assessment area. This is because many clients in the legal and criminal justice systems may not actually want to undertake the assessment, or they know that the results of the assessment may have serious implications for their lives. On some personality tests (e.g. the MMPI–2 or the 16 PF) items have been added in an attempt to detect the tendency to fake good or bad or to adjust scores. In this section, we discuss two psychological tests that have been developed specifically to detect **malingering**: the attempt to exaggerate symptoms or claim symptoms one does not have.

malingering
responding or behaving in such a way to present oneself in a negative or positive manner during a psychological test

The Structured Interview of Reported Symptoms (SIRS; Rogers, Bagby & Dickens, 1992) was designed to 'detect malingering and other forms of feigning of psychological symptoms' in adults 18 years and older. Specifically, it focuses on deliberate distortions in self-presentation. The SIRS is an individually administered instrument that comprises 172 items. These items cover a wide range of psychopathology, including symptoms that are unlikely to be true. Thirty-two of the items are repeated to detect inconsistency in responding (e.g. providing different answers to the same question). The SIRS uses a structured interview method and takes 45 to 60 minutes to complete. Ratings on the items are categorised into eight primary scales (Rare Symptoms, Symptom Combinations, Improbable or Absurd Symptoms, Subtle Symptoms, Blatant Symptoms, Severity of Symptoms, Selectivity of Symptoms and Reported Versus

Observed Symptoms) and five supplementary scales (Direct Appraisal of Honesty, Defensive Symptoms, Symptom Onset and Resolution, Overly Specified Symptoms and Inconsistency of Symptoms). Scores on each of these scales are classified as being 'honest', 'indeterminate', 'probably feigning' or 'definite feigning' based on research findings collected with psychiatric patients, normals, simulators and malingerers.

The internal consistency for the SIRS primary and supplementary scales has been found to range from 0.66 to 0.92. Its inter-rater reliability has been found to range from 0.89 to 1.00. In terms of validity, the SIRS has been found to be effective in discriminating between individuals instructed to feign mental illness (i.e. simulators), honest responders and suspected malingerers. Construct validity of the SIRS is supported by results of factor analyses. Finally, its construct validity has been supported by correlations with the validity scales of the MMPI.

One of the most common symptoms associated with malingering is memory impairment (Rogers, 1997; Shum, O'Gorman & Alpar, 2004). This is because problems with forgetting and remembering are frequently associated with the effect of compensable brain injury (motor vehicle accident, assaults, falls, sports injury etc.). This fact is also reinforced in the community by popular books, films and television programs. To 'assist neuropsychologists in discriminating between bona fide memory-impaired patients and malingerers', Tombaugh developed the Test of Memory Malingering (TOMM; Tombaugh, 1996). The TOMM is an individually administered test that is suitable for adults aged 16 to 84 years and takes only 15 minutes to complete. It aims to detect response bias, intentional faking and exaggeration of symptoms by showing a test taker fifty line-drawings of ordinary objects and then asking her, after a delay, to recognise the target among a choice of two drawings. The TOMM was developed based on the assumption that on a two-choice recognition test for fifty target items, a person's performance should not be lower than the chance level (i.e. less than 25 items). According to Vitelli (2001), the TOMM has been found to have high coefficients of internal consistency (0.94 to 0.95), but no information on test-retest or inter-rater reliability has been included in the test manual. Furthermore, validation studies found that simulators and suspected malingerers performed significantly more poorly on the TOMM than normals, and individuals with TBI and other organic problems. In one particular study, the sensitivity (the proportion of simulators correctly classified) was found to be 93 per cent and specificity (the proportion of non-simulators correctly classified) 100 per cent. (See the Technical Appendix for comments on sensitivity and specificity.) Finally, test performance on the TOMM was not found to be sensitive to age, education and cognitive impairment.

Limitations of forensic assessment

It has to be acknowledged that the practice of forensic assessment is not without its critics. Faust and Ziskin (1988), in the prestigious journal *Science*, questioned the contribution of psychologists and psychiatrists as expert witnesses in courts. They argued that the evidence provided by psychologists and psychiatrists in court is of low reliability and validity, and does not assist decision making by the court. Faust and Ziskin expanded their arguments in a number of books (Faust, Ziskin & Hiers, 1991; Ziskin & Faust, 1988) that provided lawyers with a resource for challenging psychological and neuropsychological evidence in court. In response to the original article, a number of psychologists (e.g. Fowler & Matarazzo, 1988; Heilbrun, 1992; Matarazzo, 1990) argued that Faust and Ziskin were selective in reviewing evidence relating to the utility of forensic assessment and had overstated the case.

The controversy has, however, alerted psychologists who work in the legal area to the limitations of forensic assessment. These limitations include: self-report instruments are prone to malingering; actuarial formulae have not been developed for many assessment instruments to interpret and predict behaviours; and small sample sizes were used in some of the validation studies for forensic assessment instruments. The controversy has also prompted professional psychological societies to develop clear guidelines to improve the practice of forensic assessment. For example, the Australian Psychological Society has developed resources to assist its members to manage legal requests for client files, subpoenas, third party requests for psychological report, and disclosure of test data and test materials.

Practitioner profile

Dr Danielle Schumack

1. **How long have you been a psychologist?**
   I have worked as a forensic psychologist for the past 12 years. I completed my postgraduate training in forensic psychology at the Griffith University School of Psychology. In my final year of study I was introduced to and later employed at the Griffith Youth Forensic Service (GYFS).

2. **What is your specialisation and how did you get the training and experience to do this job?**
   I currently specialise in the assessment and treatment of adolescents who engage in sexually abusive behaviour. GYFS was established in Queensland Australia in 2001 to provide specialist assessment and treatment services on a state-wide basis for young people aged between 10–17 years who have been adjudicated for sexual offences. GYFS is funded by the Queensland Department of Justice and Attorney General (Youth Justice Services) with in-kind support from Griffith University.

   GYFS has a primary office at Griffith University's Mt Gravatt campus. GYFS clinical staff travel throughout the state, including to regional and remote locations, to conduct comprehensive assessments, prepare psychological pre-sentence assessment reports for the

courts, and to deliver specialised and individualised treatment interventions in collaboration with local community partners. Fundamental to GYFS service delivery model is the need to understand people who commit sexual offences in the context of their development, their natural ecosystem and the immediate environment in which the offence/s occurred.

GYFS operates as part of a broader program of research and practice at Griffith University concerned with understanding and preventing sexual violence and abuse. Applied research activities include investigations of developmental pathways of adolescent and adult sexual offenders; onset, progression and desistance among offenders; risk prediction; prevention of child sexual abuse; and clinical and forensic psychological interventions with adolescent and adult offenders.

Given the assessment and treatment of people who engage in sexually abusive behaviour is a specialist area of psychological practice, GYFS has financially supported the ongoing professional development activities of staff. GYFS and Griffith University have funded staff to travel to both national and international conferences to increase practitioners' skills and capabilities in this specialist area of practice.

3. What kind of clients and referrals do you usually get?

GYFS accepts referrals for clients (primarily male) throughout Queensland, who have been found guilty in court in relation to sexual (or sexually motivated) offences. GYFS receives referrals exclusively from Department of Justice and Attorney General, Youth Justice Services. Youth referred to the GYFS service can be located anywhere in the state of Queensland and GYFS clinicians will travel to their community to provide assessments and treatment interventions.

4. Do you use psychological tests in your practice?

GYFS assessments include the use of psychological tests. Psychometric information is integral to the comprehensive assessment of youth who engage in sexually abusive behaviour. GYFS regularly engage a test battery which is age appropriate, time efficient and demonstrates sound reliability, validity and utility. Additional specialised psychometric assessments are engaged when required, depending on the needs of the young person.

A variety of types of standardised tests and procedures are available and are of value in assessing youth who engage in sexually abusive behaviour. These include personality tests, behavioural checklists and rating scales, structured interview schedules, test measures of cognitive and academic competencies, and attitude measures. The GYFS assessment test battery focuses on adaptive functioning; behavioural, emotional and social problems; personality functioning and psychopathology; and experiences of trauma.

Standardised risk/needs instruments constitute another category of assessment tools. Risk/needs instruments are designed to evaluate the youth's risk of reoffending and to identify his or her needs (dynamic risk factors) to aid in treatment planning. GYFS assessment battery includes comprehensive risk/needs assessments. The assessment of risk of both sexual and non-sexual offence recidivism among adolescents who have engaged in sexually abusive behaviour is a complex task with significant implications for the young person and their communities. Limitations with the existing literature are recognised, including the shortage of validated risk factors associated with reoffending and research, specifically ethnic minorities and female offenders. GYFS engages guides and checklists to aid in the systematic review of risk factors that have been identified in the professional literature as being associated with sexual and non-sexual reoffending.

5. Why do you use psychological tests and in what way do they help you in your practice?

In addition to historical information sourced at referral and clinical interviews, psychometric assessment provides a norm-based reference to assist in understanding what vulnerabilities may have contributed to a young person's offending behaviour. Psychological assessment is an invaluable tool in understanding the connection between psychological functioning and behaviour.

Psychological testing in forensic settings can corroborate clinical impressions and interview data, and collect information of broader psychological complexity. A comprehensive battery of well-researched and standardised tests with highly reliable, valid and reproducible results can assist the forensic clinician in appreciating the complexity of the individual and formulating an understanding of their offending behaviour. This formulation is critical to planning appropriate treatment interventions and guides the development of individualised treatment plans, and allows for ongoing assessment of treatment progress and outcomes. Treatment recommendations also inform the intensity of interventions required to safeguard the individual and ensure community safety.

6. In your opinion, what is the future for psychological testing in your specialisation?

Assessing the probability of recidivism in youth who engage in sexually abusive behaviour is a complex process. Currently, there is a lack of appropriate instruments for predicting youth sexual recidivism, given the low base rate (the majority of youth who engage in sexually abusive behaviour are not charged for new offences once they reach adulthood). Future research is required to improve the understanding of those factors related to offending as well as factors that protect youth from future offending.

Assessment strategies and instruments require development in order to improve forensic practitioners' capacity to identify those youth who are most at risk for sexual recidivism. Given forensic psychologists practice and their assessment findings invariably involve a level of social responsibility, as their actions and recommendations may impact the lives of others, the development of these measures is critical. Once more efficient measures are available, treatment interventions can be more effectively implemented and may lead to further reductions in recidivism and sexual harm.

Chapter summary

The purpose of forensic assessment is to collect relevant data using psychological tests and other assessment techniques to assist decision makers in the legal and criminal justice systems. Compared with therapeutic and other types of assessment, forensic assessment has to follow not just scientific and professional standards and guidelines but also legal standards and requirements. Results and reports of forensic assessment are more likely to be subjected to scrutiny, clarification and challenge because of the adversarial nature of the legal and criminal justice systems. This further highlights the importance for psychologists of writing psychological reports that are empirically sound and based on research findings, and for them to be familiar with ethical and professional guidelines (e.g. confidentiality, informed consent, duty to warn and protect, and record keeping) that relate to psychological assessment. In reviewing some of the more commonly used tests and techniques used for forensic assessment, we hope we have made you more aware of the typical referral questions raised in the legal and criminal justice systems and how they are answered by different types of assessment instruments. Finally, given the relatively short history of forensic psychology, it is

important to be aware of the limitations of forensic assessment and some of the latest developments in the area.

Questions

1. What is forensic psychological testing and assessment (forensic assessment)?
2. Can a clinical psychologist conduct forensic assessment? Provide reasons for your answer.
3. What are some of the common settings in which forensic assessment is conducted?
4. Describe a psychological test designed for forensic assessment and evaluate its psychometric properties.
5. Compare and contrast the three types of assessment techniques that can be used in forensic assessment.
6. What are some of the issues that one needs to take into consideration when assessing a person's competency to stand trial?
7. What is malingering? Why does it happen? What techniques have been developed to assess malingering? Briefly discuss the validity of these techniques.
8. 'Forensic assessment does not assist decision making in the legal and criminal justice systems.' Do you agree? Provide reasons for your answer.

Further reading

Ackerman, M J (2010). *Essentials of forensic psychological assessment* (2nd ed.). Hoboken, NJ: Wiley.

Archer, R P, Buffington-Vollum, J C, Vauter Stredny, R & Handel, R W (2006). A survey of psychological test usage patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84–94.

Martin, M, Allan, A & Allan, M M (2001). The use of psychological tests by Australian psychologists who do assessments for the courts. *Australian Journal of Psychology, 53*, 77–82.

Ogloff, J R P & Douglas, K S (2013). Psychological assessment in forensic settings. In J R Graham & J A Naglieri (Eds.), *Handbook of Psychology: Vol. 10 Assessment Psychology* (pp. 373–93). Hoboken, NJ: John Wiley & Sons.

Otto, R K & Heilbrun, K (2002). The practice of forensic psychology. *American Psychologist, 57*, 5–18.

Useful websites

Australian Standards of Practice for Family Assessments and Reporting—February 2015: http://www.familycourt.gov.au/wps/wcm/connect/fcoaweb/about/policies-and-procedures/asp-family-assessments-reporting
Best practice in psychological assessment of capacity in legal settings (*InPsych* 2015): https://www.psychology.org.au/inpsych/2015/august/dear/
Forensic psychology (Australian Psychological Society): www.psychology.org.au/community/specialist/forensic
Psychologists as expert witnesses in courts and tribunals (*InPsych* 2010): www.psychology.org.au/publications/inpsych/2010/august/allan
Specialty guidelines for forensic psychology (American Psychological Association): www.apa.org/practice/guidelines/forensic-psychology.aspx
The violent client: Advances in violent risk assessment (*InPsych* 2006): www.psychology.org.au/publications/inpsych/risk

13 Educational Testing and Assessment

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. understand the role of psychological testing and assessment in different educational contexts
2. give examples of aptitude tests, achievement tests and rating scales used in educational assessment
3. understand major criticisms of educational tests
4. be aware of some of the social implications of testing and assessment in educational settings

KEY TERMS

achievement test
aptitude test
constructed response test
formative assessment
high-stakes test
multiple choice test
standard
standardised test
summative assessment

Setting the scene

- A school teacher wants to know the standing of her class on the three Rs so she gives her students a reading test, a writing test and a test of arithmetic.
- A child is being very disruptive in class. In order to help diagnose his behaviour problems he is referred to an educational psychologist for assessment.
- A mother believes that her child is gifted. Wanting to provide her child with the best opportunities at an early age, she has his intelligence assessed in order to determine how well he will cope with an accelerated curriculum.
- A policy maker wants to know whether the literacy and numeracy of students in the state education system are improving or declining.

Introduction

Assessment in educational contexts—be it in the classroom, in a special education setting, for university admission or for professional accreditation—involves the basics of assessment: devising opportunities to gather information, collecting the information, interpreting it, and acting on the interpretation (Bennett, 2011). Devising opportunities might be by choosing already available tests that purport to tap the capacities of interest, or it might mean (as with a classroom teacher) developing class activities, formulating questions or setting homework that will draw out the students' knowledge or understanding. Once gathered, the information must be set against what else is known about the person and what inferences are therefore reasonable. Bias in this process is possible and insight into this is necessary if proper inferences are to be drawn. For example, is a wrong answer on a classroom test a careless mistake or a sign of a confusion of concepts, or is it based on a more fundamental misunderstanding? And to what extent does the teacher's overall view of the student influence the inference that is drawn? Effective action relies on the dependability of the inference: dismissing a mistake as a slip because it was made by a 'good' student might impede the student's progress if the real basis of the mistake was a significant misunderstanding. In this case, action—such as having the student revise more basic material—is required. The hallmarks of good assessment practice apply in education as in other contexts.

In educational contexts, an important distinction is often made between summative and formative assessment. The distinction originated with Scriven (1967) in the area of program evaluation, but was used by the educationist Bloom (1969) and has been elaborated by Australian researchers (see, for example, Sadler, 1989, 1998). Summative, or evaluative, assessment, particularly when it involves standardised testing (discussed below), is seen in a negative light in

some educational circles, whereas **formative assessment** (i.e. which facilitates learning) is viewed positively. This is not, however, a critical feature of the difference, and in fact in some circumstances a hard and fast distinction cannot be drawn. For example, items of formative assessment might be included for reaching a **summative assessment**, or summative assessment might be used to inform subsequent teaching.

formative assessment
an assessment aimed at facilitating learning as well as evaluating it

summative assessment
an assessment that has a purely evaluative function

A further distinction that is made in educational assessment, although by no means unique to it, is between achievement and aptitude tests. **Achievement tests** assess past learning—that is, learning that has already taken place—whereas **aptitude tests** assess future learning potential. Probably no other form of assessment is more common than the teacher-constructed achievement test administered at the end of a course. We are all familiar with this type of class test or exam at school. Although the vast majority of achievement tests are specific in nature, it is possible to develop general achievement tests aimed at assessing basic skills developed during any mode of instruction. Like achievement tests, aptitude tests might be general or specific, although general aptitude tests are more common. General aptitude tests are virtually indistinguishable from the ability or intelligence tests described in Chapter 7. Well-constructed general aptitude tests correlate with academic performance about 0.3 to 0.7 depending on the criterion used (e.g. Roth et al., 2015). Specific aptitude tests can be constructed for particular skills and abilities if a more focused assessment is required.

achievement test
a test to assess past learning

aptitude test
a test to assess future learning potential

During primary and secondary school, an important concern in educational assessment is identifying students with special needs. The sooner such needs are

recognised, the sooner remedial work can begin. Achievement tests can be used to assess a student's progress through the standard curriculum, whereas aptitude tests can be used to diagnose deeper problems with their learning and reasoning abilities. Is a learning or a behaviour problem the result of visual, hearing or motor impairment, or is it essentially cognitive in nature? What role if any do social, economic or environmental factors play? Rating scales and checklists, which parents or teachers can use to record the frequency and occurrence of certain behaviours, are readily available (Kamphaus, Petoskey & Rowe, 2000). Assessment of special needs lies at the boundary of educational and clinical psychology.

More recently, interest has also developed at the other end of the ability continuum: identifying gifted students who might benefit from a more enriched or accelerated curriculum. Giftedness is invariably assessed using intelligence or aptitude tests, although peer ratings and teacher nominations have also been used. The needs of gifted students have been highlighted because it is believed that, if not catered for, these students might lose motivation and fail to reach their full potential, becoming bored and frustrated with the slow pace of education around them. In extreme cases, boredom can lead to misbehaviour in the classroom. Although giftedness might sound highly desirable, growing up gifted presents its own difficulties, such as loneliness and isolation through difficulties in fitting in with one's peer group (Clark, 1988). As such, it is important to identify gifted children to ensure that they remain stimulated and able to develop their talents to the fullest.

In the following sections we outline some of the commonly used tests in educational assessment and the contexts in which they are applied.

Group-administered achievement tests

Achievement tests are traditionally thought of as tests of what the person knows or can do as a result of an education or training program, such as exposure to an primary school curriculum or a first-year course in psychology. The program is usually common to a cohort of students, as when students in different schools complete a specified curriculum. The achievement test is administered to the group at a point in time that is meaningful in terms of the objectives of the training program, such as the end of an instruction period, and samples the content appropriate to the program. The term summative assessment is commonly used when the achievement test result is compared with some standard or benchmark of performance to judge the success or otherwise of the training program.

Achievement tests are often standardised to permit the comparison of scores among candidates. For example, the SAT (originally, the Scholastic Aptitude

Test) has been used with relatively few modifications by many universities in the USA since 1926 to assess applicants for entry. It currently has a vocabulary section and a mathematics section and assesses writing skill separately. With almost as long a pedigree is the ACT (originally American College Testing) for university entry. Competition for entry, particularly for the more prestigious universities, is such that the grounds for decision making need to be quite transparent and this purpose is met by **standardised tests**. See the section 'Admissions decisions' later in this chapter for a discussion of this concept in the Australian context.

standardised test
a test administered and scored in a set way

The terms standardised and **standard** should not be confused. Standardised tests are those where the 'test content is equivalent across administrations and the conditions under which the test is administered are the same for all test takers' (Sireci, 2005, p. 113). The rationale is that, with conditions constant, the only source of difference is the characteristic (knowledge, skill or proficiency) being measured. Standardised tests are designed according to a test specification (see Chapter 6) and are administered under uniform conditions, the scoring is the same for everyone, and if there are different forms of the test they are statistically and qualitatively equivalent. The **multiple choice test** (MCT), in which a question is followed by four or five options from which the candidate selects the correct answer, has been frequently used as the format in standardised tests since this format was pioneered in the US military during the First World War (Jones & Thissen, 2007).

standard
fixed level of attainment

multiple choice test
(MCT) a test where each question has a number of options, of which only one is correct

A standard, in contrast to a standardised test, is a basis for comparison in terms of level of attainment. A standard can be applied whether or not a standardised test is used for assessment. There are various ways in which a standard can be set. One is in terms of the average performance of a group of interest (e.g. students in third grade in Australian schools). Alternatively, one

might opt for an absolute standard. Absolute standards are predetermined levels of achievement that are maintained irrespective of student performance. For example, a typing rate of 50 words per minute might be set for a secretarial position, even though none in an applicant group might achieve that speed. Absolute standards are difficult to implement because the definition of what constitutes proficiency depends on the perspective of the person setting the standard. A maths teacher is likely to expect more of students in maths than, say, a history teacher. Although it is generally seen to be a good thing to set high educational standards, they can lead to rates of failure that are considered unacceptable and to questions about what to do with those who fail. Absolute standards often have to be tempered by reference to what people can actually do (see Box 13.1).

Box 13.1

Are our educational standards declining?

From time to time in the media there are reports of poor or declining literacy standards among Australian workers. 'Can't spell, can't count: Bosses lash out at workers' lack of skills' ran the headline in the *Herald Sun* on 18 January 2016, which introduced a report on the claim by the Australian Industry Group (AIG) that nine out of 10 employers have workers who 'prepare work riddled with errors', among other limitations. The AIG convened round-table discussions of fifty-eight employers and surveyed 338 companies about major literacy and numeracy problems in the general population and the workforce. Poor literacy is obviously a major concern in a democratic society and is of particular significance at a time of increased global competitiveness.

The judgments of employer groups might be thought of as an absolute standard against which performance can be judged. An alternative norm-based standard is drawn from the Survey of Adult Skills, an international survey conducted in 33 countries as part of the Programme for the International Assessment of Adult Competencies (PIAAC). It surveys 5000 adults in each country and measures the key cognitive and workplace skills needed for individuals to participate in society and for economies to prosper. In 2013, the most recent data available, Australia's performance in numeracy was at about average for the twenty-one countries with which it was compared, including countries of Western Europe, the UK, Canada, the USA and Japan; while for literacy Australia was placed in the top four countries. It might be that literacy and numeracy in these OECD countries is substandard or that attempting to set

standards without some reference to how people actually perform is not as informative as it might appear to be. Robert Linn, a specialist in educational measurement, espoused the need for an 'existence proof' before a standard is set; that is, that someone, somewhere has been shown to meet that standard (Shepard & Baker, 2016).

When standards are not met, someone is often held accountable. It might be the government of the day, the school system, individual schools and their principals or teachers, and sometimes the students themselves. Achievement tests are often 'high stakes' tests. Mislevy (2012) defines high-stakes tests as 'ones for which results have important consequences for someone'. Standardised tests can be high stakes but they need not be, and not all **high-stakes tests** are standardised. Mislevy gives the example of a PhD thesis examination, which is high stakes for the candidate but is not standardised—rather it is tailored to the work the candidate has produced.

high-stakes test
a test where the results have important consequences for the test taker

NAPLAN

In the Australian context, the debate about high stakes testing has centred on NAPLAN (National Assessment Program—Literacy and Numeracy) (see, for example, Klenowski & Wyatt-Smith, 2011). NAPLAN is a standardised test or, more accurately, a suite of standardised tests. Since 2008, all school students in Years 3, 5, 7 and 9 in May each year sit NAPLAN tests in reading, writing, language conventions and numeracy, unless intellectually or functionally impaired or from a non-English-speaking background and have been in Australia less than 12 months. The purpose is to develop 'consistency, comparability, and transferability of information on students' literacy and numeracy performance nationally' (Australian Curriculum Assessment and Reporting Authority (ACARA), 2010, p. 3). From 2010, the test results have been available on the MySchool website, which is managed by the federal government and provides parents and the public at large with the opportunity to compare the performance of 'like' schools across Australia—where likeness is based on an index of socioeconomic status. Teachers' organisations (e.g. Australian Primary Principals Association, 2009, 2010) have been vocal critics of NAPLAN. The Australian Senate undertook an inquiry into NAPLAN in 2013 (Senate Standing Committee on Education and Employment, 2014), and the Whitlam Institute has amassed

research information and data on NAPLAN and its perceived effects on teachers, principals, parents and children (Dulfer, Polesel & Rice, 2012).

Brady (2013) reviewed the criticisms that have been made of NAPLAN and summarised them in this way. NAPLAN:

- has resulted in teaching to the test and a narrowing of the curriculum
- has led to a 'dumbing down' of learning
- has led to high levels of teacher stress
- has led to high levels of student stress
- has disadvantaged certain groups of students
- is not sufficient as a diagnostic tool.

Brady sees merit in all of these criticisms, but (with the possible exception of the sixth), these are more about the way the test is used rather than the test itself. Other standardised tests of achievement used in Australia have not resulted in the degree of notoriety of NAPLAN. Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA) and Progress in International Reading Literacy Study (PIRLS) are used in Australia as part of international testing programs (Meeks, Kemp & Stephenson, 2014), but the consequences for poor performance on these lie with the state and federal governments and not with individual principals or teachers. It is important to note that the perceived problems with standardised tests are largely those perceived by teachers. In this country, as in the USA, parents and the general public have not been the major critics of standardised tests (Cizek, 2001; Dulfer, Polesel & Rice, 2012).

Constructed response tests

The issues surrounding standardised tests in education need to be separated from issues to do with the application of standards and the issues of accountability. When this is done the question remains whether standardised tests are the best way to assess achievement. The traditional alternative has been the essay question where the examinee is asked to write on a set topic. It might be a long essay of several pages completed under examination conditions, but is often much shorter than this. It is sometimes referred to as a **constructed response test** (CRT), to contrast it with the MCT. The specific format allows for variations in length, but the essential feature is that the examinee must construct the answer rather than recognise the correct option, as in the MCT. The essay format is said to engage higher level cognitive processes of selecting, relating and organising knowledge in contrast to the more basic processes of memory for

facts that critics say characterise standardised tests. Bloom's taxonomy (see Anderson & Krathwohl, 2001) is often appealed to in this discussion because it orders the processes involved in learning from higher to lower levels. Methods that involve the higher levels are usually thought of as the goal of educational endeavours and these, it is said, are best assessed, or can only be assessed, using a constructed response format.

constructed response test
(CRT) a test that requires the test taker to construct the answer in response to the question; no options are provided (as are in multiple choice tests)

It is not necessarily the case that standardised tests are unable to engage higher level processes (see, for example, Williams, 2006), although unskilled test developers might not know how to do this. It is also the case that essay questions have some difficulties of their own that have been known for the best part of 100 years. One is the time taken to complete and mark essay questions. Several MCT questions can be administered and scored in the time taken to complete one constructed test item. A further problem is the reliability of essay marking. Different markers do not agree on the quality of written essays and the judgments of the same marker from one occasion to another shows considerable variation. This was reported in the early years of the twentieth century (e.g. Starch & Elliot, 1912) and continues to be reported in the early years of this century (see Meadows & Billington, 2005; Tisi et al., 2013). The irony is that reliability can be improved, but this has to be done by constraining the constructed response test to one that more closely resembles the MCT.

Individually administered achievement tests

The tests considered to this point have been tests administered to groups of students at the one time. As well as these, there are a number of tests that can only be administered to one person at a time and these are relevant when the performance of the individual is the central concern.

Wechsler Individual Achievement Tests

As well as group tests administered to numbers of people at a time, there are individually administered achievement tests. A good example is the Wechsler Individual Achievement Test–Third Edition (WIAT–III; Wechsler, 2009a). The first version of the WIAT was published in 1992 and the WIAT–II in 2005. An updated version for Australia and New Zealand was published in late 2016.

The WIAT–III assesses basic academic skills in oral language, reading, written expression and mathematics. It is designed to be used with children from as young as 4 years to adults aged 19 years and 11 months. There are separate norms for adults aged 20 to 50 years. Administration time varies from 30 to 145 minutes, depending on the subtests used and the grade level of the child. The WIAT–III can be used to assist diagnosis of learning difficulties, eligibility for placement in special education programs, and other intervention decisions. It is not designed to assess giftedness.

Table 13.1 lists the subtests and the composite to which they belong, and an example of each. There is also the composite score—a total achievement score made up of scores on all the subtests. Three of the subtests are new (Oral Reading, Math Fluency and Early Reading Skills) and several subtests have been revised (Listening Comprehension, Oral Expression, Written Expression and Reading Comprehension). More detail is provided in the test manual and in reviews of the test.

The norms for the 4–19-year-olds were based on a sample of 2775 students stratified by age, grade, gender, race/ethnicity, parents' education and geographical region. Special education students were included in the sample. Split-half reliabilities for the subtests range from 0.83 to 0.97 and for composite scores from 0.91 to 0.98. Retest reliabilities ranged from 0.82 to 0.94 for subtests and 0.87 to 0.96 for composites. Subtests requiring the exercise of judgment by the scorer showed inter-rater agreement of 91 per cent to 99 per cent. Validity evidence included a discussion of content validity, an analysis of factor structure, and correlations with other tests.

Table 13.1: WIAT–III subtests and composites

Subtest	Example	Composite
Listening Comprehension	Pointing to pictures that show meaning of presented words	Oral Language
Oral Expression	Generating words to best describe pictures	Oral Language
Early Reading skills	Naming letters	(no composite)
Reading Comprehension	Responding to questions about a passage	Total Reading, Reading Comprehension and Fluency
Word Reading	Reading aloud from a word list	Total Reading
Pseudoword Decoding	Decoding nonsense words	Total Reading, Basic Reading

Subtest	Example	Composite
Oral Reading Fluency	Reading passages aloud	Total Reading, Reading Comprehension and Fluency
Alphabet Writing	Writing letters	Written Expression
Sentence Composition	Formulating sentences	Written Expression
Essay Composition	Writing an essay on a given topic	Written Expression
Spelling	Writing words presented	Written Expression
Maths Problem Solving	Problems in geometry and algebra	Mathematics
Numerical Operations	Calculation skill with numbers	Mathematics
Math Fluency—addition	Speed and accuracy of addition	Math Fluency
Math Fluency—subtraction	Speed and accuracy of subtraction	Math Fluency
Math Fluency—multiplication	Speed and accuracy of multiplication	Math Fluency

Woodcock-Johnson Tests of Achievement

The first version of the Woodcock-Johnson in 1977 included tests of achievement, cognitive abilities and interests, and was the first comprehensive battery for psychoeducational assessment. It was revised in 1989. The interests section was dropped and the cognitive sections restructured to reflect the Cattell-Horn theory of intelligence. This was revised in the Woodcock-Johnson Cognitive and Achievement Battery (WJ III COG) published in 2001 (Woodcock, McGrew & Mather, 2001), which used the Cattell-Horn-Carroll (CHC) theory (see Chapter 7) as the basis for subtest selection. There was a re-norming in 2007 using the 2005 US Census, and in 2014 the Woodcock-Johnson IV Tests of Achievement (WJ IV ACH; Schrank, Mather & McGrew, 2014) was released. It added seven new achievement tests and separated the oral language tests into their own battery (WJ IV OL). The WJ IV COG was also published in 2014 (Schrank, McGrew & Mather, 2014).

The WJ IV ACH was co-normed with the WJ IV OL and the WJ IV Cog test batteries. Co-normed means that they were normed on the same stratified sample of children and adults to allow comparison between achievement and aptitude scores with greater accuracy than if the tests were normed separately.

The WJ IV ACH comprises a standard set of eleven subtests that yield fifteen cluster scores and an extended set of nine subtests that yield an additional seven cluster scores. The clusters are described in Table 13.2. Inspection indicates that they cover a variety of aspects of reading, writing, mathematics and general academic achievement. The extended set provides more in-depth information and the assessment of specific strengths and weaknesses. The standard set can be used alone or administered with the extended set.

There are age- and grade-equivalent scores, percentile ranks, relative proficiency index (RPI) scores, W scores (the Rasch model was used in test construction), and standard scores, stanine scores, T-scores and z scores for both tests and clusters (see the Technical Appendix).

Table 13.2: Areas of academic achievement covered in the Woodcock-Johnson IV Tests of Achievement

Cluster	Description
Reading	reading decoding, reading comprehension
Broad Reading	reading decoding, reading speed
Basic Reading Skills	sight vocabulary, phonics, structural analysis
Reading Comprehension	comprehension, reasoning, vocabulary
Reading Comprehension	comprehension, reasoning, vocabulary
Extended Reading Fluency	automaticity, accuracy
Reading Rate	automaticity
Mathematics	problem solving, computational skill
Broad Mathematics	number facility, reasoning, problem solving
Math Calculation Skills	skills with basic maths facts
Math Problem Solving	mathematical knowledge and reasoning
Written Language	spelling and quality of expression
Broad Written Language	spelling, writing fluency, quality of expression

Cluster	Description
Basic Writing Skills	identifying and correcting errors in spelling, punctuation, word usage
Written Expression	meaningfulness and fluency
Brief Achievement	proficiency in reading, writing, mathematics
Broad Achievement	proficiency in reading, writing, mathematics
Academic Skills	basic academic achievement
Academic Fluency	fluency in use of academic skills
Academic Applications	application of skills to academic problems
Academic Knowledge	knowledge of science, social studies, humanities
Phoneme-Grapheme Knowledge	knowledge of sound-symbol relations

Villarreal (2015)

The WJ IV ACH was normed on a stratified sample based on the 2010 US Census. Stratification was based on region, sex, country of birth, race, ethnicity, community type, parent education, type of school, type of college, educational attainment, employment status and occupational level. A total of 7416 individuals were tested from 2 years to 80 plus years.

Median internal consistency reliabilities (split-half) across different age groups of the norming sample ranged from 0.84 to 0.94 for test scores and 0.90 to 0.96 for cluster scores. Test-retest reliabilities over a one-day period were 0.83 to 0.95. Validity evidence includes content evaluation, factor structure information and the ability of the test to identify clinical groups, including those with learning difficulties.

The literature on use of the Woodcock-Johnson III, the predecessor of the WJ IV ACH, with special populations was reviewed by Abu-Hamour et al. (2012). They concluded:

> [T]he WJ III proves to be a valuable diagnostic tool to be used to identify exceptional children including: high incidence disabilities such as ADHD, language impairment, mild intellectual disability, specific reading, math, and written language disabilities, and traumatic brain injury; and low incidence disabilities such as hearing impairment, visual impairment, and autism; and gifted students including those with a learning disability. (Abu-Hamour et al., 2012, p. 671)

In Australia, however, the test is not widely used, to judge from a report by Meteyard and Gilmore (2015). They surveyed professionals in the area of psychoeducational assessment in Australia to determine how they assessed specific learning difficulties and the tests they used for this purpose. Respondents held qualifications in psychology or education and worked predominantly as school psychologists or guidance counsellors in public and private systems. The most widely used tests were the Wechsler Scale of Intelligence for Children–Fourth Edition (WISC–IV), with 84 per cent of respondents using it 'Often or Always'; and the Wechsler Preschool and Primary Scale of Intelligence (WPSSI; 65 per cent), followed by the Wechsler Individual Achievement Test–Second Edition (WIAT–II; 59 per cent). Only 16 per cent reported using the WJ III Ach and only 9 per cent reported using the WJ III COG 'Often or Always', despite there being Australian norms available (McGrew, 2008). Experience with a test and its psychometric properties were indicated as the most frequent reasons respondents chose a test, but the authors noted that another possible reason was that in some systems the Wechsler tests are mandated.

Case study 13.1

Learning disability and the IQ–achievement discrepancy

The concept of learning disability (LD) was introduced to the literature to describe those children whose achievement level was substantially below what was expected in terms of their general ability or, as one scholar put it, 'unexpected learning problems in a seemingly capable child' (see Lyon, 1996).

In the USA, LD became an important referral question for school psychologists because of federal legislation that funded children who were so assessed. The idea of a discrepancy between ability and achievement became associated with a difference in scores on an IQ and an achievement test, although this was not mandated by federal legislation and there was no formal statement about which IQ or achievement tests were to be used or which level of difference constituted an important enough difference to invoke a description of LD. (One of the motivations in revising the Woodcock tests reviewed in this chapter was the accurate measurement of this discrepancy.)

The number of children with LD increased substantially in the USA from the 1970s to the 1990s and several high-level committees were commissioned to inquire into the matter. Fletcher et al. (2004) describe the background to this public concern and the several issues that bedevilled the field. Their conclusion was that the IQ–achievement discrepancy be abandoned in describing LD, and that IQ tests play a limited role in assessment. Attention instead should focus on

the criterion of low achievement and the clear differentiation of LD from other problems (e.g. mental retardation or behaviour problems). The defining feature of LD, in their view, should be a poor response to instruction directed to remediating low achievement. Such an approach was seen to avoid the problems in defining a discrepancy and, importantly, focus efforts and resources on remediation.

In Australia, a strict discrepancy criterion was not adopted by state education departments and instead what was considered LD in the USA was described in terms of specific learning difficulties that could arise against a background of underachievement. For example, about 80 per cent of children classed as LDs in the USA were found to have problems with reading. Early identification of a reading problem can lead to intervention before it becomes a significant impediment to future school success. Although the discrepancy criterion has no official status in Australia, a survey of school psychologists in Western Australia by Klassen, Neufield and Munro (2005) found that 90 per cent considered that 'learning disabilities should be distinguished from other forms of low achievement', while 81 per cent believed that 'IQ tests are useful in the identification of learning disabilities'. Together these beliefs imply a discrepancy criterion.

Discussion questions

1. Why do you think the discrepancy definition endured for so long in the USA and is still defended by some professionals and clients?
2. Could a discrepancy definition be used alongside the other definitions proposed for learning disability?
3. What are some of the factors to be considered when a child is not achieving well at school?

Wide Range Achievement Test

The Wide Range Achievement Test–Fourth Edition (WRAT–4; Wilkinson & Robertson, 2006) continues a test first published in 1946. It measures the basic academic skills of reading, spelling and mathematical computation, and in the latest version provides a measure of sentence comprehension in addition to word reading.

The Word Reading and Sentence Comprehension subtests must be individually administered, but the other two can be administered to individuals or groups. There are alternate forms (Blue and Green) that can be used interchangeably as before and after measures for assessment of an intervention or that can be combined for a more comprehensive evaluation. There are also grade-based norms that can be used for assessment in grades K to 12 (see

Chapter 3 on the use of grade-based norms). Age-based norms (up to 94 years) allow assessment of basic literacy with older adults.

The normative sample included 3000 US citizens. Alternate forms reliability (30 days) is reported as 0.78 to 0.89, with little practice effect. Validity information is drawn mainly from content analysis and correlations with other tests.

Teacher-constructed tests

Classroom teachers in their everyday work are more interested in formative than in summative assessment. That is, they are more concerned with how the student is progressing towards an outcome rather than in the outcome itself; that is, what is the gap between where the student is now and where they could be. Assessment in this context is concerned with: (a) understanding the student's misunderstanding and so planning and adjusting future instruction; (b) providing prompt feedback to the student to assist learning and maintain motivation; and (c) informing parents and caretakers of strengths and areas of challenge for the student. Standardised assessment items or essay items can be used in this process, but it is far more dynamic than the fixed item or set of items can capture and calls for a range of tasks and often individually tailored questions asked of a student.

In such a context, criterion referencing rather than norm referencing becomes important. That is, for competence in a particular subject domain or segment of that domain, what are the tasks the student can complete and what are those they cannot? To use a primary school example, what are the steps in mastering the operation of division in arithmetic? Sophisticated theoretical models of 'dynamic assessment' have been developed to facilitate this process (Robinson-Zañartu & Carlson, 2013). These models focus on the interaction of teacher and student and on the change that is demonstrated with specific interventions. They are, however, relatively complicated models and require special training and to date do not seem to have been taken up by classroom teachers. Less formal approaches are more typical of classroom practice, such as asking questions (e.g. Heritage & Heritage, 2013).

Black and Dylan (1998) in a comprehensive review of assessment in the classroom noted that actual practice, perhaps not surprisingly, often falls short of the ideal. They noted the following key weaknesses:

- Classroom evaluation practices generally encourage superficial and rote learning, concentrating on recall of isolated details, usually items of knowledge which pupils soon forget.

- Teachers do not generally review the assessment questions that they use and do not discuss them critically with peers, so there is little reflection on what is being assessed.
- The grading function is over-emphasised and the learning function under-emphasised.
- There is a tendency to use a normative rather than a criterion approach, which emphasises competition among pupils rather than personal improvement of all. The evidence is that with such practices the effect of feedback is to teach the weaker pupils that they lack ability, so that they are demotivated and lose confidence in their own capacity to learn. (Black & Dylan, 1998, p. 10)

You may see some parallels in these criticisms with those of standardised achievement tests such as NAPLAN. Time constraints and the focus on pedagogy mean that the classroom teacher is unlikely to spend much time in formally testing students, but the collecting of evidence, sifting it and forming judgments—the hallmarks of good assessment practice—apply here as in other educational contexts.

Aptitude tests

As well as individual tests of achievement, there are a number of individual tests of aptitude that are used in educational contexts.

Stanford-Binet Intelligence Scales

The origins of practical assessment of intelligence lie very much in the work of Alfred Binet and Lewis Terman, as outlined in Chapter 7. The Stanford-Binet scales that these researchers pioneered have seen several revisions over the years since the first edition was published in 1916. The most recent version, Stanford-Binet Intelligence Scale–Fifth Edition, was published by Gale Roid (2003) and is referred to as the SB5. A number of the items are retained from earlier versions but were included after a rigorous selection process that involved item fit to a Rasch model (see Chapter 6). There is a new structure to the test based on the CHC model of cognitive abilities (see Chapter 7). The ten subtests, including both verbal and non-verbal items, are scored for five of the CHC factors: Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual-spatial Processing and Working Memory. There are also scores for the subtests, four intelligence composites and an overall measure of general intelligence.

The test was normed using a sample of 4800 US citizens covering the age range of 2 years to 85 years plus, thus providing comparison points for a wide range of ages. Apart from its historical significance, the test offers much in terms of the assessment of abilities in special populations. In the case of the gifted, for example, it has a higher ceiling of difficulty compared with other comparable intelligence tests. A suggested classification based on the full scale score (mean = 100, standard deviation = 15) is that those with scores in the range 145 to 160 are 'very gifted' or 'highly advanced' and those in the range 130 to 144 are 'gifted' or 'advanced'.

Internal consistency is high, ranging from 0.95 to 0.98 for full scale, verbal and non-verbal scales; 0.90 to 0.92 for the factor scores; and 0.84 to 0.89 for the ten subtests. A variety of methods was used to establish validity, including professional judgment, intercorrelations with other aptitude tests, examination of age changes, and confirmatory factor analysis. The test has been adapted for use in Australia, but is not as widely used as the WISC.

Wechsler Intelligence Scale for Children

As discussed earlier in this chapter, the Wechsler Intelligence Scale for Children (WISC) is one of the most frequently used general ability tests for school-aged children (Kamphaus, Petosky & Rowe, 2000). Now in its fifth edition (WISC–V; Wechsler, 2014), the test was first introduced in 1949. It is an individually administered test for children aged 6 years to 16 years 11 months and is used to identify a child's intellectual strengths and weaknesses, as well as diagnose giftedness and mental retardation. It can be used for planning treatment and making placement decisions in clinical and educational settings, and to assist in neuropsychological evaluation.

The WISC–IV (Wechsler, 2003) introduced a number of changes from the editions that preceded it, notably in dropping the Verbal IQ and Performance IQ concepts, which had been part of the Wechsler suite of tests since the outset. These were replaced with four index scores (Verbal Comprehension, Perceptual Reasoning, Working Memory and Processing Speed), which was more in line with the results of factor analyses of the subtests and theoretical ideas about intelligence. The changes continue with the WISC–V. Two primary subtests (Word Reasoning and Picture Completion) have been removed and three new ones added (Visual Puzzles, Figure Weights and Picture Span). The Perceptual Reasoning Index has been split into two (Visual Spatial and Fluid Reasoning) to better reflect a five-factor solution, with the Arithmetic subtest being transferred from Working Memory to the new Fluid Reasoning Scale. There are also a number of new complementary subtests. For the first time a digital version as

well as a pencil-and-paper version of the test is provided for administration on an iPad.

The primary structure of the WISC–V is shown in Table 13.3. Not shown are the Ancillary (5) and Complementary (3) scales that are of value in clinical work. The primary subtests take just over an hour on average to administer. A number of scores can be obtained apart from the Full Scale Intelligence Quotient (FSIQ), including subtest and primary index scores. The FSIQ is now based on seven subtests, whereas in the previous edition it was based on ten subtest scores.

Table 13.3: Structure of the Wechsler Intelligence Scale for Children–Fifth Edition, with primary subtests arranged under the Primary Index Scales

Primary Index Scales				
Verbal Comprehension	Visual Spatial	Fluid Reasoning	Working Memory	Processing Speed
Subtests				
Similarities*	**Block Design***	**Matrix Reasoning***	**Digit Span***	**Coding***
Vocabulary*	**Visual Puzzles**	**Figure Weights***	**Picture Span**	**Symbol Search**
Information Comprehension		Picture Concepts	Letter-Number Sequencing	Cancellation
		Arithmetic		

Note: Subtests in bold are those used for computing Primary Index Scale scores. Subtests marked with an asterisk are the seven subtests contributing to the calculation of the Full Scale IQ (FSIQ) or g.

The WISC–V was normed on a stratified sample of 2200 children. Stratification within eleven age groups was based on the 2012 US Census and was by gender, ethnicity, parents' education and geographic region. Across age groups, internal consistency (split-half) estimates for subtests ranged from 0.81 to 0.94, for primary index scores from 0.91 to 0.96, and for FSIQ from 0.96 to 0.97. Stability coefficients (over an average 26 days) were 0.91 for the FSIQ, 0.68 to 0.91 for primary index scores and 0.76 to 0.89 for subtests. Inter-scorer reliability varied from 0.97 to 0.99, because scoring is for the most part objective. A variety of validity evidence is reported, including content evaluation, factor structure, correlations with other tests and discrimination of clinical groups.

A technical review of the WISC–V was provided by Canivez and Watkins (2016) and an overview with a more clinical focus by Greathouse and Shaughnessy (2016).

The Wechsler Preschool and Primary Scale of Intelligence

This Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 2012a, 2012b) is part of the Wechsler suite of tests and is designed to measure general cognitive functioning in children aged 2 years 6 months to 7 years 7 months. First developed in 1967, it was revised in 1989 and 2002, and can be considered a downward extension of the earlier developed WISC. The fourth edition, WPPSI–IV, was released in 2012 with some modifications, including new subtests for assessing working memory and processing speed and a closer integration with the CHC model of intelligence (see Chapter 7). The concepts of Verbal IQ and Performance IQ, used since the original WAIS, are no longer used. A version with Australian and New Zealand norms appeared in 2014. The WPPSI was third in Oakland, Douglas and Kane's (2016) list of most used tests by school psychologists in a survey of sixty-four countries.

On the WPPSI–IV, children aged 2 years 6 months to 3 years 11 months complete five core subtests: Receptive Vocabulary, Information, Block Design, Object Assembly and Picture Memory. Picture Naming and Zoo Locations are supplementary subtests that can be administered. Three composite scores (Primary Index Scales) are calculated: Verbal Comprehension (VCI), Visual Spatial (VSI) and Working Memory (WMI), which are combined to provide a Full-Scale IQ (FSIQ). There are also three Ancillary Index Scales that can be calculated, which have particular clinical use.

Children aged 4 years to 7 years 7 months complete six core subtests: Information, Similarities, Block Design, Matrix Reasoning, Picture Memory and Bug Search. Seven supplemental subtests are available: Vocabulary, Object Assembly, Picture Concepts, Zoo Locations, Animal Coding and Cancellation. As well as the three Primary Index Scales calculated for children in the earlier age band, two additional Primary Index Scales are calculated: Fluid Reasoning (FRI) and Processing Speed (PSI). All five are used to calculate the FSIQ in this age band, and there is an additional Ancillary Index Scale. The core subtests take approximately 30 minutes for most children in both age bands to complete.

Standardised scores (mean = 10, standard deviation = 3) are calculated from raw scores for each subtest and these are combined to calculate index scores and FSIQ, which each have a mean of 100 and a standard deviation of 15. Age-based percentile ranks are also available as well as scores for various types of comparisons.

The US norming sample included 1700 children with equal numbers of males and females, stratified by ethnicity. The sample included special groups, including intellectual disability, autism and Attention Deficit/Hyperactivity Disorder. Internal consistency (split-half) at the level of the subtests ranges from 0.75 to better than 0.90. At the composite level the lowest reliability was from 0.86, with the FSIQ showing reliability of 0.96. Subtest test-retest reliability ranged from 0.75 to 0.87, with values for the composites from 0.84 to 0.89. For FSIQ ,the test-retest reliability was 0.93. Validity evidence includes content evaluation, factor structure (three factors for the early age band and five for the later age band) and correlation with other tests. A useful review of the test was provided by Syeda and Climie (2014).

Kaufman Assessment Battery for Children

The Kaufman Assessment Battery for Children (K–ABC) was developed by Alan S Kaufman and Nadeen L Kaufman. The first edition appeared in 1983 (Kaufman & Kaufman, 1983) and offered two important advantages over tests available at the time for assessing children's cognitive development such as the Stanford-Binet and the WISC. One was that it purported to assess thinking in terms of a neuropsychological model that had several antecedents, but most notably the work of AR Luria (1966). The model distinguishes between the style of problem solving the child adopts and the knowledge, facts and skills the child has actually acquired. The style of problem solving involves integrating stimulus materials to achieve a result (e.g. in solving a jigsaw puzzle) or arranging material in a series (e.g. repeating numbers in the order they are presented). The former is an example of what is termed simultaneous processing and the later serial processing. The K–ABC was designed to assess these thinking styles, which could then be put together to form a Mental Processing Composite (MPC), and to assess separately in an Achievement Scale what knowledge the child has acquired.

The second advantage of the test was that a number of subtests in the battery can be administered without the use of language; that is, by pantomime presentation of the requirement of a task and by having the child answer with a motor response. This was a distinct advantage for testing children who were non-English speakers or who had a language or hearing disorder.

The test was revised in 2004 (K–ABC II; Kaufman & Kaufman, 2004), with some subtests removed and new ones added to provide a total of eighteen subtests to cover a somewhat wider age range (from 3 to 18 years). The simultaneous/successive processing model (the Luria model) was retained but an alternative in terms of CHC theory added, with the choice of model being left to the test user. For example, for children from mainstream culture the decision

might be that the CHC model be used, whereas for children with a disability or from an ethnic background the Luria model might be more appropriate. Using the CHC model, the test can take up to 70 minutes to administer, depending on the age level and subtests selected. With the Luria model chosen, administration typically takes less than this (maximum 55 minutes) because there are two fewer subtests.

Scores with the Luria model are provided on the following scales: the Sequential Processing Scale, the Simultaneous Processing Scale, Learning Ability Scale and Planning Ability Scale. With the CHC model the scales are termed: Short Term Memory (Gsm), Visual Processing (Gv), Long Term Storage and Retrieval (Glr) and Fluid Reasoning (Gf), respectively. A Crystallised Ability (Gc) scale is the Knowledge scale in the Luria model. As with the original edition of the test, Knowledge is not added to scores on the other four scales to form the Mental Composite Index (MCI). It is, however, combined with the four scale scores to form a composite under the CHC model, termed the Fluid–Crystallised Index (FCI). Scoring provides for age-based standard scores, age equivalents and percentile ranks.

The standardisation sample included 3025 participants aged 3 to 18 years stratified by socio-demographic factors according to the 2001 US Census. Reported internal consistency coefficients for scale scores vary from 0.81 to 0.95, with composite scores somewhat higher. Test-retest reliability for subtest scores vary from 0.5 to the mid-0.80s, and from 0.86 to 0.94 for the composite scales. Evidence of validity is offered in terms of the outcomes of confirmatory factor analysis of the structure of the test, the correlations of the test with other tests of ability and achievement such as the WISC–IV and WIAT–II and the Woodcock-Johnson III ability and achievement batteries, and in terms of the capacity of test scores to identify particular clinical groups, such as children with learning disabilities or ADHD.

Cross-battery assessment

The brief review of individual aptitude tests indicates a substantial number are available, but which should be chosen? Tradition has favoured the Wechsler tests and these are still the most widely used, but understanding of cognitive functioning has increased since the basic model for these tests (verbal versus non-verbal abilities) was first established. For example, CHC identifies ten factors in the cognitive domain and none of the tests currently in use provides measures for all of these. For this reason, some psychologists have recommended using subtests from more than one test to cover the domain more adequately (McGrew & Flanagan, 1995, 1996; Woodcock, 1990). A ‘cross-battery’ (XBA) approach, as it is termed, is based on the findings of factor-analysis and seeks to assess as

validly as possible the basic constructs identified in the factor-analytic work. This involves choosing subtests from ability batteries that have the strongest loadings on the CHC factors that are relevant to the particular assessment of aptitude. Guidelines for the application of this approach were published by Flanagan, Ortiz and Alfonso (2013). The approach demands much of the test administrator, not only in terms of increased testing time and the provision of resources (several test batteries rather than just one) but also in terms of a thorough understanding of modern factor theory of mental ability. A case example illustrating the use of XBA was reported by Jacobs, Watt and Roodenburg (2013).

Behaviour rating scales

To this point we have been considering psychological tests composed of items that the individual being tested completes. As well as these sorts of tests, a method of assessment that has become widely used is the rating scale completed by someone who observes the behaviour of the person being assessed. Observation of behaviour in context is a rich source of information about adequacy of functioning. This can be done in a number of ways, from direct monitoring of particular behaviours in terms of their frequency and the conditions that precede and follow them—referred to as functional behavioural assessment—to the use of rating scales directed to the occurrence of specific behaviours. Rating scales are widely used because of their efficiency in collecting information from a number of sources. For example, a child might be rated by their parents, by their teacher and, depending on their age, by their peers. Scales can be used to evaluate the presence or absence of a particular behaviour (e.g. aggression in the playground) as well as its frequency or intensity. Although efficient, the scales capture a perception of the subject's behaviour, and different observers will have different perceptions, partly because of the context in which they make the observation and partly due to inter-rater variability. The following sections summarise three scales used for assessing behaviour in children and adolescents.

Achenbach Child Behaviour Checklist

Research on the Achenbach scales began in the 1960s and the first, the Child Behaviour Checklist (CBCL) for the rating of boys by their parents, appeared in the late 1970s (Achenbach, 1978). Since then a number of rating scales have been published to cover boys and girls from ages 1 to 18 years. Collectively they are referred to as the Achenbach System of Empirically Based Assessment (Achenbach & Rescorla, 2001). Their purpose is to help identify adaptive and

maladapative functioning in children (Achenbach & Rescorla, 2001). There are scales to cover the age range 1 to 5 years and the age range 6 to 18 years, and a self-report form for those aged 11 to 18 years. Areas covered and the number of items vary somewhat depending on the age group. For those aged 6 to 18 years, areas include academic performance, working hard and behaving appropriately, as well as internalising and externalising scales. The latter cover a range of behaviours such as depressed, withdrawn, nervous and obsessive (internalising) and hyperactive, attention demanding and aggressive (externalising).

The original norming sample consisted of 3943 US children. A T-score transformation is used to express scores. The scales are now used round the world (Ivanova et al., 2007), including Australia, but there are no Australian norms.

The manual (Achenbach & Rescorla, 2001) provides a good deal of information on reliability and validity of the scores on the scales from the item level to the composite level. For reliability, internal consistency (alpha), test-retest and between-rater coefficients are reported. The latter are generally lower than the stability and consistency coefficients (0.69 to 0.93). For validity, several sources of data are provided, including factor structure and criterion prediction. In the latter case, children referred to guidance centres and other agencies for evaluation are compared with children not so referred. The results of validity examinations in most cases are strongly supportive.

Conners Rating Scales

Work on developing the Conners Rating Scales began in the 1960s and the first of these was published in 1989 as the Conners Rating Scales (CRS; Sparrow, 2010). Its purpose was the assessment of behaviours of clinical significance in diagnosis, particularly Attention Deficit Hyperactivity Disorder (ADHD), and it came in time to be one of the most widely used devices for this purpose. The original scale was revised in 1997 (CSR–R) and again in 2008 (Conners Third Edition; Conners–3). At the time of the third revision, two new scales of wider applicability than the Conners–3 were developed and published in 2008 (Conners Comprehensive Behavior Rating Scales; CBRS) and in 2009 (Conners Early Childhood; Conners EC). The Conners–3 and the CBRS cover the age range 6 to 18 years and the Conners EC the age range 2 to 6 years.

Taken together, the scales cover a wide variety of behavioural, emotional, social and academic domains (e.g. aggressive/oppositional behaviour, irritability, anxiety, depression, social skills and interests, subject-specific difficulties and inattention), as well as information used to make specific predictions (e.g. potential for violence and self-harm). The Conners-3 and the CBRS were aligned with the Diagnostic and Statistical Manual of Mental Disorders–Fourth Edition–

Text Revision (DSM–IV–TR), and the most recent updates provide an option to use DSM–5. From the outset, the scales were designed to be used by teachers and parents and, to a limited extent, for self-rating. The latter use was extended over the years so that there is now the option for respondents aged 8 to 18 years to complete a self-report form (Conners-3 and Conners CBRS).

Each of the tests comes in forms of different lengths (e.g. the short form of the Conners–3 has 99 items and takes about 25 minutes to complete, and the CBRS teacher version has 204 items). Items are rated on four-point scales from 0 to 3, except for milestone items on the Conners EC. The number of scale scores that can be derived depends on the test and the form used. There are also validity scales to assess Positive Impression, Negative Impression and Consistency in Rating. Scores are expressed as T-scores with an option for percentiles. These are based on the standardisation samples for each test. The Conners–3 and the CBRS were co-normed using data from 1200 teachers, 1200 parents and 1000 self-reports for equal numbers of boys and girls at each age from 6 to 18 years. As well, there were large clinical samples (2143 for the Conners–3 and 2076 for the CBRS). For the Conners EC the standardisation sample of 1600 included equal numbers of teachers and parents and equal numbers of children in the age range 2 to 6 years. There are currently no Australian norms.

Average internal consistency across scales ranged from 0.84 for the Conners CBRS to 0.80 for the Conners–3. Test-retest reliability ranged from 0.82 for the CBRS to 0.90 for the Conners EC, and inter-rater reliability from 0.73 for the CBRS to 0.78 for the Conners–3. Validity evidence includes the ability of the scales to discriminate clinical and healthy groups, with overall correct classification rates averaging 75 per cent for the Conners–3, 78 per cent for the CBRS and 86 per cent for the Conners EC. Information on sensitivity and specificity (see the Technical Appendix) of the scales is also provided. For example, for the ADHD index on the Conners–3, the sensitivity and specificity for parent ratings are 80 per cent and 87 per cent respectively, and for the teacher ratings are 75 per cent and 83 per cent. Data are also provided on the convergent and discriminant validity of the scales using scales from other measures of childhood psychopathology.

Vineland Adaptive Behavior Scales

The first version of the Vineland Scales (VSMS) appeared in 1935 and was the first attempt to assess the social maturity and competence necessary for personal independence. Up until then, judgments of competence and maturity were based on tests of mental ability such as the Stanford-Binet. Doll (1935) argued that assessments of mental ability need to be supplemented with information about what the person can do in managing their daily affairs; that is, their behaviour as

it is observed by people who know the person well and who observe their behaviour across a range of tasks. Although there have been changes to Doll's original test, the basic ideas of observations of behaviours in different settings remains.

The original form of the test was revised in 1984 and a classroom edition added in 1985. The Vineland Adaptive Behavior Scales II (Vineland II; Sparrow, Cicchetti & Balla, 2005) appeared first in 2005. A Vineland–3 has recently been released (Sparrow, Cicchetti & Sauliner, 2016). The Vineland II covers the age range birth to 90 years and can be completed as a set of rating scales (0 'never' to 2 'usually' with a 'don't know' option provided) or as an interview. The interview form is slightly shorter (413 items versus 433) and covers five domains: Communication, Daily Living Skills, Socialisation, Motor Skills and Maladaptive Behaviour. The rating form adds a further domain of Problem Behaviours. Domain scores are calculated using standard scores and expressed as an Adaptive Behavior Composite.

Norms are based on a sample of 3695 citizens of the USA selected to be representative in terms of education, ethnicity and geographic residence within a number of age groups. The same norms are used for interpretation of both the interview and rating forms. No Australian norms are available.

Reliability data are provided for internal consistency, test-retest and inter-rater. Split-half reliabilities across age groups for the Adaptive Behavior Composite are reported as 0.90 or better across age groups, with domain scores somewhat lower. The same is true for test-retest reliability, although that for the 14 to 21 year age group was only 0.81. Inter-interviewer correlations range from 0.4 to 0.9, with most in the 0.4 to 0.6 range. Inter-rater reliability was reported as 0.32 to 0.81. A variety of validity information is provided, including the capacity of scores to identify a range of disabilities including mental retardation, autism, ADHD and learning disability. Interview and rating forms of the Vineland II correlate at a maximum of 0.8.

Admissions decisions

A further area of application of educational testing is the assessment of potential candidates for places within particular programs at universities. Admissions decisions involve deciding who gets into a course or program of study. This is somewhat analogous to personnel selection, as described in Chapter 10, and can be an extremely difficult decision for administrators to make. Many see value in an independent, objective, standardised assessment as providing a common metric along which to compare applicants.

Earlier we noted the use of the SAT and the ACT in the management of university admission in the USA. In Australia, however, how well a person

performed at school determines admission to university for most institutions. Performance on achievement tests such as the NSW Higher School Certificate or the Victorian Certificate of Education are used in admissions decisions. In practice, school marks are transformed to a score (a percentile rank) that indicates a student's position in relation to all other students in the state or territory and allows comparison across state-based school systems. The score in current use is the ATAR, which replaced the Universities Admission Index (UAI) in New South Wales in 2009 and similar indexes used in other states, such as the ENTER in Victoria, in 2010. Queensland will use the ATAR rather than the Overall Position (OP) from 2018.

One of the issues in such debates is the extent to which university admission based on school results serves to perpetuate socioeconomic differences in the society. If the more socially advantaged compared with the less receive better schooling, then they have an advantage in entry to university and their time there reinforces that advantage. For example, less than 20 per cent of first year places in Australian universities are filled by students from those in the lowest socioeconomic category, and this has changed little over the four years for which most recent data are available (Department of Education and Training, 2014). Moves in recent times to increase enrolment numbers in Australian universities have meant lower admission requirements for some courses at some universities, and this might redress the situation in time. However, as long as the more prestigious courses at the more prestigious universities are subject to stringent admission requirements based on school marks, the argument about disadvantage accruing from the use of these selection procedures as a predictor will continue.

A second issue is the role of so-called 'soft skills' (e.g. teamwork, leadership and emotional intelligence) in career and life success. These, it is argued, are not well captured by school grades or university results, and yet are as important (and possibly more so) in determining success outside the academic domain (e.g. Heckman & Kautz, 2012). Achievement and aptitude—as they have been traditionally measured in the tests reviewed in this chapter, and as reflected in school results—have their origins in the classroom and perform best in predicting academic criteria. Why then, it is argued, make career decisions on such a narrow basis? One problem is that it has proved difficult to identify measures of soft skills that predict 'real world' criteria independently of measures of academic achievement or aptitude.

A case in point is the UMAT, an aptitude test devised by the Australian Council for Educational Research and used for student selection by the majority of medical schools in Australia, in conjunction with other methods including school grades and interviews. UMAT is designed to assess reasoning skills and, importantly, the ability to understand people. The data currently available indicate that the test does no better than school grades in predicting outcomes in

medical training (Wilkinson, Zang & Parker, 2011). Whether UMAT scores will predict longer-term career outcomes remains to be seen, but as these criteria are assessed a long time after taking the test, success at predicting them is likely to be poor. It is not surprising that school grades, which represent an amalgam of scholastic ability and motivation—at least at the general level if not in every individual case—do a better job than other variables in predicting outcomes in academic settings.

Tests are also used by some professional bodies in deciding fitness to practise. The Psychology Board of Australia, for example, has introduced an examination for those seeking registration as psychologists. The examination resembles a standardised achievement test for knowledge of psychology and rests on the assumption that competent practice requires a sound understanding of the subject. Unlike admission tests where predictive validity can be tested against a range of criteria, licensing or certification tests rely solely on content validity for their justification.

Practitioner profile

Associate Professor Tim Hannan

1. **How long have you been a psychologist?**
   I completed the requirements for registration as a psychologist in 1991, and since that time have worked in diverse areas of psychology, including public health, private practice and tertiary education. The majority of my career has been in the field of child and adolescent psychology, and over the years I developed a keen interest in the assessment of developmental and acquired cognitive disorders, and the importance of assessment to the delivery of effective, evidence-based therapeutic services to children and their families.

2. **What is your specialisation and how did you get the training and experience to do this job?**
   My educational background includes postgraduate qualifications in clinical psychology (University of Sydney) and clinical neuropsychology (Macquarie University), which provided a firm foundation for developing expertise in assessing children, adolescents and adults. By undertaking further professional development provided by the Australian Psychological Society, I acquired specific knowledge and skills in cognitive assessment and understanding childhood disorders.

3. **What kind of clients and referrals do you usually get?**
   The majority of children referred for psychological assessment are presenting with a developmental learning disorder or other developmental cognitive disorder, such as attention deficit hyperactivity disorder, intellectual disability or autistic spectrum disorder. Generally, children are referred by parents, teachers or caregivers due to concerns over academic progress, or other indications of a cognitive disorder.

4. **Do you use psychological tests in your practice?**
   Yes I do use them. Standardised tests are an essential component in a comprehensive psychological assessment, in combination with a clinical interview, questionnaires and

structured observation. Useful tests include measures of intelligence, oral and written language, and abilities in other specific areas of cognitive functioning. Competent practice in this field requires the knowledge and skills to integrate test results with information gained from other sources, in order to diagnose the specific condition with which the child is presenting, and to provide informed advice or recommendations concerning interventions for some aspect of the child's educational or psychosocial functioning.

5. **In your opinion, what is the future for psychological testing in your specialisation?**

   As accurate diagnosis of the presence, nature and severity of developmental cognitive disorders requires careful measurement of cognitive abilities, it is certain that psychological tests will continue to play a critical role in assessment.

   With developments in touchscreen technologies and online delivery of tests, we are likely to see considerable changes in the way that testing is conducted. Computer-based tests offer the promise of a greater degree of consistency in test administration and the quantification of responses, as well as the potential to perform rapid and complex analyses of results.

Chapter summary

Testing and assessment have figured strongly in education for over a hundred years. Achievement tests are used extensively to assess the level of learning achieved by each student in a course or achievement in education more generally. Aptitude tests are used to assess giftedness or special needs, and rating scales are used to assess aspects of non-cognitive functioning. Revision and refinement of these tests over the years have increased their effectiveness at a technical level. However, the use of tests in educational contexts—primary, secondary or tertiary—remains controversial because of the differing value positions of the stakeholders involved.

Questions

1. Is a take-home exam an example of formative or summative assessment?
2. Should criteria for good assessment practice be applied to classroom tests?
3. Give an example of a standardised test and justify your choice.
4. What are some of the instrumental uses of testing in our society?
5. Are educational standards created or discovered?
6. Should NAPLAN be abolished? Justify your answer.
7. Which commonly used aptitude tests use CHC theory in their interpretation?
8. How are the Wechsler tests alike?
9. What are the typical reliabilities of aptitude tests at the subtest and composite levels?

10. Professional licensing tests rely on content validity. Is this a problem?

Further reading

Kubiszyn, T & Borich, G (2007). *Educational testing and measurement: Classroom application and practice* (8th ed.). New York, NY: Wiley.
Nitko, A J (2004). *Educational assessment of students* (4th ed.). Upper Saddle River, NJ: Merrill.
Overton, T (2000). Assessment in special education: An applied approach (3rd ed.). Upper Saddle River, NJ: Merrill.
Reynolds, C R & Kamphaus, R W (Eds.). (2003). *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement*. New York, NY: Guilford.

Useful websites

School assessment (Australian Council for Educational Research): www.acer.edu.au/assessment/school-assessments
My School (Australian Curriculum, Assessment and Reporting Authority): www.myschool.edu.au
OECD Programme for International Student Assessment (PISA): www.oecd.org/pisa

PART 5 PROSPECTS AND ISSUES

14 The Future of Testing and Assessment

CHAPTER OBJECTIVES

By the end of this chapter you should be able to:

1. understand how psychological construct development (e.g. in areas of intelligence, personality, integrity, and neurosciences) is contributing to driving future test innovation and development
2. understand how technical (e.g. computer, online, gaming applications) and methodological advances (e.g. item response theory) are changing the way we think about, deliver and interpret psychological tests
3. explain the basic mechanics of CAT and MAT
4. discuss the advantages and disadvantages of internet testing.

KEY TERMS

artificial intelligence
computerised adaptive testing
construct
emotional intelligence
integrity test
item response theory
item-generation technology
latent factor- centred design
multidimensional adaptive technology
time-parameterised testing
virtual reality

Introduction

Discussion of the future of any field necessarily involves an element of speculation. One can ponder trends that seem to be on the horizon, but there is always the possibility of some unexpected development. Our discussion of the future of testing and assessment is no different and is organised into three sections: content developments, technical and methodological developments, and contextual changes.

Construct development

Advances in psychological theory, such as new **constructs** emerging in the literature, might give some idea about new psychological tests and procedures likely to be developed in the future. This seems especially true in the fields of cognitive abilities and personality theory.

construct
a specific idea or concept about a psychological process or underlying trait that is hypothesised on the basis of a psychological theory

Criticisms of traditional formulations of intelligence are well known (see Chapter 7). Several authors have sought to expand the concept of intelligence over the past few decades, and research along these lines is likely to continue. Two prominent theorists in this regard are Gardner (1983) and Sternberg (1985a, 1997). Gardner's theory of multiple intelligences includes dimensions such as interpersonal and intrapersonal intelligences, in addition to traditional psychometric intelligence (Gardner, 2006; also see Chapter 7). One of the difficulties Gardner and his colleagues have faced is the development of adequate measures that tap individual differences in some of the new intelligences he proposes. When such tests are available, they will no doubt have an impact on psychological testing and assessment.

In his triarchic theory of intelligence, Sternberg (1985a, 1985b) introduced the idea of practical intelligence, partly to explain success in non-academic pursuits. Many people seem to survive on their wits, using what might be called 'street smarts' rather than any kind of analytic problem solving or deep analysis (Wagner & Sternberg, 1991). Intelligence measured by traditional tests does not seem to capture this, explaining less than half of the variance in any relevant outcome measure. Sternberg regarded practical intelligence as consisting of knowledge of process and procedures rather than knowledge of content, facts and figures: 'knowing how' rather than 'knowing that'. Practical intelligence is context-based, pragmatically useful and acquired through experience rather than formal instruction (see Chapter 7). Importantly, from an applied perspective,

useful measures of practical intelligence have been developed in the form of tacit knowledge tests (e.g. Sternberg, 1997; Sternberg & Horvath, 1999; Sternberg et al., 1995). Methods of assessing practical intelligence include simulations, analysis of critical incidents and situational judgment tests. Proponents of practical intelligence claim their measures are uncorrelated with traditional measures of intelligence, yet are as predictive of important outcomes as more traditional ability tests.

New constructs have also appeared in the area of personality and the ability domain. Probably the most noteworthy is the emergence of the Big Five model of personality discussed in Chapter 8. Another important development within personality theory is the rise of the concept of integrity (Van Iddekinge et al., 2012; see Chapter 10). **Integrity tests** attempt to measure concepts like dependability, theft proneness and counterproductive work behaviour. In the current age of heightened security, there will no doubt be a growing interest in ideas around constructs like integrity. Another interesting new construct is **emotional intelligence** (EQ; Matthews, Zeidner & Roberts, 2002). Many instruments have been developed to measure emotional intelligence over the past decade and interest in this construct looks to continue. At this stage theorists are still uncertain about exactly where to locate emotional intelligence within existing theory. Some researchers have dismissed it as an amalgamation of existing personality traits (Joseph & Newman, 2010), but many EQ theorists put the emphasis on 'intelligence' and see it as a new type of ability. Indeed, the whole area of emotions remains largely untapped by psychological tests, so we could be witnessing the emergence of whole new domains of individual differences. New developments in the psychology of emotions will no doubt strongly influence these trends.

integrity test
either a specific type of personality test or a direct measure to assess a job applicant's honesty, trustworthiness and reliability

emotional intelligence
(EQ) a controversial construct (considered by many to not be an 'intelligence') that refers to the person's capacity to monitor and manage their own emotions and to understand the emotions of others, and to use these insights to function better interpersonally

The burgeoning neurosciences continue to throw light on brain function through an array of imaging techniques and these are having a growing impact on psychology. It is even possible that psychological interpretations might be found for particular image patterns, blurring the lines between physiological and

psychological assessment. As yet, however, psychological constructs have not been identified with particular brain locations or specific brain activity and it is unlikely that this degree of correspondence will ever occur. Most psychological constructs are normally distributed in the population, suggesting, in line with the central limit theorem of theoretical statistics, that they are the product of many underlying processes rather than the result of specific physiological, biological or evolutionary events.

Technical and methodological developments

As has happened throughout the history of testing and assessment, new technology and developments in psychometric theory will continue to influence the field. The most obvious innovation here is the use of computers and tablets to administer and score tests. Most of the tests discussed in previous chapters have been of the traditional pencil-and-paper variety, limited as they are to static two-dimensional graphics and text, whereas a computer can present almost any kind of stimulus imaginable. Much has already been achieved using a standard screen, mouse and keyboard, including the presentation of animation and video; and the collection of simple responses, reactions and even complete prose. Sound is also being incorporated, including spoken items and instructions. Speech recognition technology might eventually permit free-flowing, verbal responses as input. In spite of these possibilities, many computerised tests use the computer as little more than an automatic page turner (Bartram & Hambleton, 2006), reflecting the legacy of pencil-and-paper tests. However, it might not be long before simple, dyadic, question-and-answer sessions give way to fully immersive, interactive experiences. These might incorporate **virtual reality** (a computer-based technology that mimics a real or constructed environment via a computer screen or virtual reality headset, although some programs can deliver more specific experiences, such as via wired gloves; also known as immersive multimedia, computer-simulated reality and remote communication environment) and **artificial intelligence**, which is still largely undeveloped, but eventually might parallel or surpass human cognitive functioning, including in areas of perception, problem solving and learning. Complexity of scoring and presentation rules are no longer an obstacle once a computer takes over. Computers can score tests and write reports, all within a few seconds of test completion. Further, a computerised test displays the same stimulus in exactly the same way to each examinee, greatly improving standardisation of presentation.

virtual reality
a computer-based technology that mimics a real (e.g. for pilot training) or constructed environment (e.g. game technologies) for the user, who is placed in,

and can interact with, the environment; advanced packages also include touch and smell

artificial intelligence
(AI) a technology-based intelligence that attempts to mimic human intelligence; recent expressions are sophisticated chess and GO playing programs, and self-driving vehicles

Three main periods have been identified in the application of computers to psychological testing and assessment (Ways et al., 2016). Starting in the 1950s, when computers first became available, the idea of computerised adaptive testing (CAT) was conceived, and new developments in test theory, such as **item response theory** (IRT; see Chapter 6), seemed ready to make this a possibility. However, the cost and specialised skills required to program large, expensive mainframes kept this technology out of the mainstream of test development until the second period, which began in the 1980s with the widespread proliferation of cheap personal micro-computers. From that time, test developers had ready access to affordable computing power, and the development of computerised testing began in earnest (Scheu & Lawrence, 2013).

item response theory
(IRT) a family of theories that specifies the functional relationship between a response to a single test item and the strength of the underlying latent trait

Figure 14.1 Virtual reality via headset

A key issue, which slowed the proliferation of computerised testing until research caught up, was the question of equivalence between computerised and pencil-and-paper tests (Scheu & Lawrence, 2013). Did computer presentation fundamentally change the construct being measured? After many equivalence studies, the conclusion has generally been in the negative (Wang et al., 2008). Meta-analysis of this research reported a cross-mode correlation of 0.97 between computerised and pencil-and-paper forms (Mead & Drasgow, 1993). There is not much difference between ticking a box on a questionnaire with a pencil and checking a box on a computer screen with a mouse. The psychological decision-making processes underlying these different motor responses remained the same.

The only exception to this were speeded tests (Way & McClarity, 2012), which are characterised by very simple tasks performed repetitively, as quickly as possible, within a short time limit. Although strongly correlated (the cross-mode correlation is about 0.70; Mead & Drasgow, 1993), speed-based paper-and-pencil tests were not equivalent to similarly constructed computer-based tests. Differences seem to be accounted for by the different psychomotor abilities required for each (e.g. manipulating a pen versus controlling a mouse), although the different contexts (page versus screen) and computer familiarity also play roles. Generally, test takers are faster using paper-and-pencil versions, as using a pencil is usually a lot easier than manipulating a mouse. A simple solution to this problem of non-equivalence was to develop new norms for computerised versions of speed tests.

Finally, the 1990s saw the widespread growth of the internet, which heralded the new era of internet testing (Barak & English, 2002). This time, the belated research process did not hamper the speed of proliferation of technology: online delivery simply had too many advantages, and testing had become big business.

The internet offers many benefits to the test developer, especially the ability to self-publish and reach millions of potential users easily. The distributed delivery of materials afforded by the internet facilitates rapid test development. Data resides with the server, not with the client, improving item security and facilitating rapid dissemination of new versions and updates. Further, examinees have easy access to practice tests and examples, and can access test materials at any convenient location that has an internet connection.

Smart testing

An interesting exercise in speculation about the technological future of testing was provided by Kyllonen's (1997) smart test. The smart test was designed to incorporate 'all current significant technology associated with abilities measurement' (Kyllonen, 1997, p. 347), including computer delivery, item-generation technology, multidimensional adaptive technology, time-parameterised testing and latent factor-centred design. Although Kyllonen emphasised ability testing, many of the techniques readily translate to other domains. The basic idea of computer delivery has already been considered, but the use of computers facilitates a range of other advanced technologies. Understanding these may provide some clue as to what might be occurring inside a computer presenting a computerised test in the not too distant future.

Figure 14.2 Online computerised adaptive testing

Computerised and multidimensional adaptive testing

Multidimensional adaptive testing (MAT) is the multivariate generalisation of **computerised adaptive testing** (CAT; Segall, 2009). In order to understand MAT, first we must understand the basic ideas behind CAT.

multidimensional adaptive technology
(MAT) programs that allow assessment of multiple dimensions of a construct of interest, which allows for a better fit between the theorised, multidimensional construct or model and the obtained data than assessing a single dimension

computerised adaptive testing
(CAT) programs that rapidly identify a test taker's ability level from a small number of items by (a) administering an initial item, (b) administering a more difficult or

easier item depending on whether the initial item was correct or incorrect, (c) again administering a more difficult or easier item depending on the response to the second item, and (d) so on

The early promise of computerised testing was nowhere more evident than in the concept of computerised adaptive (or tailored) testing. This is the idea that the computer can continuously monitor an examinee's performance and refine the trait or ability estimate after each item is presented. Further, the computer can choose, as the next item to present, the one that will provide the most information about the ability or trait being measured. This amounts to choosing the maximally discriminating item, based on the examinee's performance on the test so far. Basically, if you get an item right, the computer presents you with a harder item, but if you get it wrong, the next item is easier. In this way the test adapts to your location on the underlying trait: the point at which you would get about half the items correct and half the items incorrect.

Traditional static, non-adaptive tests need to be composed of a set number of items of varying difficulty, spread across the full range of ability. They must contain easy items for low ability examinees as well as hard items suitable for the most able. An average person taking a static test has to waste considerable time answering the easy items at the beginning of the test. Similarly, they might get bogged down on the difficult items at the end. It would be more efficient if the test could adapt to each person's ability and not waste time on items that were either too difficult or too easy. Adaptive tests 'zero in' on each person's ability level and spend most testing time administering appropriately graded items. This is the basic idea behind CAT, where continuously scoring and selecting of items results in tests that are much shorter than static pencil-and-paper tests, which are scored when all items are completed. CAT tests are highly efficient because time is not wasted on items far removed from each test taker's true score.

One interesting feature of CAT is that each test taker gets a slightly different test. Some examinees might not even be presented with any items in common. On the face of it, this seems unfair because different people are apparently assessed according to different criteria. What happened to the idea of standardisation? However, the theory is that all items need to be unidimensional —that is, tap the same underlying psychological construct—so a test composed of any subset of items from the full set of items ought to be equivalent. Development of CAT requires much work in order to produce a large pool of unidimensional items, graded in difficulty along a single psychological scale. Analysis of the items can also be quite intensive, requiring a sample of thousands of examinees in order to accurately estimate each item's characteristics using item response theory (see Chapter 6). The large pool of items underlying a CAT is known as an item bank.

MAT takes computerised adaptive testing to the next level by applying the basic idea to a whole battery of tests, instead of just to a single test. MAT capitalises on the fact that many of the constructs measured by a battery of tests are correlated. This is especially true of cognitive abilities (see the hierarchical structure of abilities discussed in Chapter 7). Interdependence among the subtests of a battery means that all the items from all subtests in a battery are related to some extent. Thus the score of one item from one subtest could conceivably be related to the score on an item from another subtest. MAT takes advantage of this interdependence and feeds performance on every item in the battery into the score for every subtest in the battery. This allows it to adapt simultaneously over all subtests by selecting the next item to present not from the item bank for a particular subtest, but rather from an item bank of all items in the whole battery. CAT adapts by dynamically estimating the single ability being measured by the test and selecting the next item that optimally improves that measurement; MAT adapts by dynamically estimating all abilities being measured by the battery simultaneously and selecting the next item, from whichever subtest, that optimally improves the measurement of all abilities.

One of the main practical advantages of CAT is its potential to reduce testing time without sacrificing accuracy of measurement due to its selection of maximally informative items. MAT takes this idea one step further and selects the item that will be most informative for the whole battery, thus making testing potentially even more efficient. Many problems are yet to be solved in MAT before it becomes widely applied, but when this occurs, MAT will have many advantages over CAT, as well as over standard paper-and-pencil tests and their computerised versions.

Limitations of CAT and MAT

One of the main difficulties in developing a CAT is the effort required to develop the sufficiently large item bank that is required. Even after several hundred items are written, the item parameters must be estimated. IRT provides the best methods for estimating item parameters (e.g. difficulty and discrimination), but requires data from large samples of examinees, which implies extensive testing during development. This issue is exacerbated in MAT.

In the case of MAT, another potential drawback is the likelihood of frequent chopping and changing between item types, as the system selects items from any subtest in the battery. Examinees could find this confusing. Further, they would need to remember the instructions for every subtest during the whole testing session in order to be able to respond to whichever item was presented. Such memory requirements might be unrealistic for many examinees, as well as being a testing confound.

Item-generation technology

If the biggest obstacle in setting up CAT and MAT lies in developing larger item banks, item-generative technologies might be the answer. In adaptive tests, often several hundred items are needed— as opposed to the usual twenty or thirty in traditional pencil-and-paper tests—and these have to be replenished on a regular basis, are time-consuming and are expensive to create. One recent development is **item-generation technology**, which has the capacity to generate new items automatically by computer according to some underlying rule or algorithm (Irvine & Kyllonen, 2002). The idea is that if the main source of difficulty for the subtest can be captured by a rule or template, the computer can be used to generate a large number of actual items of any desired difficulty by randomly initialising a few underlying variables and applying the rule. This has the potential to produce a vast number of parallel tests. To date, item generation techniques have been used mainly to develop figural ability items. Verbal content seems much more difficult to handle this way. As with all applications of computerised testing, the possibilities of this method are limited only by the imagination and ingenuity of test developers. It is not inconceivable for there to be a highly complex cognitive model of test performance underlying the item generation, although to date most applications have tended to use generic templates rather than full-blown theories of test performance.

item-generation technology
new computer programs that focus on generating an item model or template, from which many individual items can be generated

Time-parameterised testing

The aim of **time–parameterised testing** is to solve the fundamental problem of speed–accuracy trade-off, which is a basic dimension or strategy in solving any difficult task. One can work quickly and less carefully, sacrificing accuracy for speed in the hope of scoring well by doing more items; or one can work slowly and carefully, sacrificing speed for accuracy in the hope of making every completed item count. Another way of thinking about it is in terms of quantity versus quality. Someone emphasising speed is opting for quantity, trying to answer a large number of items; someone emphasising accuracy is aiming more for quality of thought and response. Unfortunately, it is impossible to tell from the final test score which strategy someone has adopted. Yet each represents fundamentally different approaches to problem solving.

time-parameterised testing
seeks to solve the problem of the trade-off between speed of responding on a test and accuracy of responding

Since its inception, computerised testing has made it possible to collect response time information. Computerised tests, therefore, can record the examinee's actual answer and the time to answer, and even record a partial response. It was always thought that this timing information would ultimately provide the solution to the speed–accuracy trade-off issue. The problem is that psychometricians have not agreed on how best to combine time and accuracy data. Invariably, these two pieces of information are analysed separately, but attempts to combine them into a single 'efficiency' type measure—for example, number of correct items divided by time to answer—have never proven completely satisfactory.

Kyllonen (1997) suggests that another way of combining timing information with accuracy scores is to treat time as one of the difficulty dimensions. This suggests a test, such as a digit span test (see Chapter 11), in which the rate of presentation of digits is varied instead of the length of the digit string. A more general way of time parameterising a test would be to introduce time limits or deadlines for each item. In this way, someone preferring an accuracy strategy could be forced to adopt a more speeded one, and the range of strategies used by all test takers would be more uniform and therefore easier to interpret. It could be argued, however, that forcing some examinees to adopt their non-preferred strategy merely introduces yet another confound.

Latent factor-centred design

Kyllonen (1997) argues that test developers need to focus more on the constructs that they want to measure, rather than on the specifics of particular tests. He calls this a construct focus rather than a test focus. Being test focused leaves us too wedded to existing ways of doing things, which could partly explain why so many computerised tests resemble familiar pencil-and-paper tests rather than something radically new. If we can become more construct focused and apply **latent factor-centred design**, then we will be more open to new testing forms. After all, the particular test being used is mere surface detail; what should interest us is the construct or latent factor underlying performance.

latent factor-centred design
the use of underlying, latent constructs to represent both multiple measures (e.g. scores for reading, arithmetic and geography) and single tests (e.g. self-

regulation); latent constructs reflect more 'pure' and efficient representations of a group of tests or a group of items

One thing not considered by Kyllonen is internet testing. Add online delivery to all of the above and you have a truly twenty-first-century instrument.

Internet testing

As already mentioned, the internet has revolutionised testing, although it has currently had more of an impact on distribution (i.e. publication) than on the development of new types of tests. Using the internet, a set of questions can be quickly circulated to psychologists and other test users all around the world. Moreover, the internet version of the test can be kept up to date, with the most recent changes disseminated to all users as soon as they are developed. It is easy to modify the questions presented and even the scoring mechanism involved.

Corresponding to the ease of keeping test users up to date with the most current version of a test, the other significant advantage of internet testing lies in getting information back to test developers. If a test is presented on the internet, it is easy to collect data for norming purposes and this speedy turnaround makes rapid test development possible. There is even the possibility of dynamic norming, in which test norms are continually updated as soon as new data comes in. This could lead to a multidimensional adaptive, item-generative, time-parameterised, latent factor-centred, dynamically normed online test.

One disadvantage of internet testing is the so-called 'digital divide' (Bartram, 2000): the fact that some people have better access to the internet than others, and that those with the best access tend to be the most privileged. There is a strong tradition in testing of trying to avoid forms of discrimination, which the digital divide could entrench. Nevertheless, it is generally believed that gaps in access and use are narrowing as computers and tablets become cheaper and more widespread. One area where internet distribution of tests is likely to alleviate discrimination is in rural areas, where accessing a professional or getting to a testing centre is difficult. Internet test users should be mindful of equity and accessibility problems, even though it is generally believed that they will tend to decline with time (Joiner et al., 2013).

A salient issue associated with internet testing concerns security of information. Given that psychological test results can be highly personal in nature, it is important that any private information be kept confidential and secure. Recent privacy legislation has sought to strengthen individuals' protection in this regard. Corresponding to the need for security of test scores and protection of privacy, there is also security of the test itself to consider. In the past, access to many psychological tests has been restricted in order to maintain

the confidentiality of test items: if the items become well known and the correct or most desirable answers become common knowledge, then the test is rendered useless. How can test security be assured when items are passing from computer to computer through unsecured networks round the world? Once the correct answers are discovered, the internet is also suited to their rapid dissemination. Internet testers concerned about item security try to do things like disable the printing and screen capture functions on the browser presenting the test. Sometimes they can even install a 'security agent' prior to test delivery, but they can never stop someone photographing their computer or tablet screen with an iPhone!

Bandwidth limitations also pose unique difficulties for internet testing. Although computers can time events to the fraction of a second, timing of events across the internet can be difficult because of lag. For example, it may not be possible to be sure of exactly when the question appeared on the examinee's screen after it left the server. This issue is called 'ping latency' in computer parlance. Similar bandwidth limitations can seriously affect CAT or MAT delivered over the internet. For these systems to work, an examinee's answers would typically have to be sent back to the server for scoring so the system can adapt. If there are delays on the network, this could slow down the whole test, resulting in a very unsatisfactory testing experience. One way around this would be for the testing system to reside on the client's computer, but this would involve downloading large portions of the item bank (if not the entire bank) every time, which may further exacerbate issues of test security.

Another difficulty concerns the types of test on offer on the internet. The internet is replete with tests of dubious quality that certainly do not measure up to the high psychometric standards set by the profession and advocated in this book. Many of these tests are pop-psychological or even para-psychological in nature. A major problem now facing the profession is how to differentiate itself from unscientific instruments available online, and educating the public about how to recognise what is reputable and what is not. Further, although the internet seems suited to the delivery of psychological tests, at this stage it is not suited to full-blown psychological assessment. Recall the distinction between psychological testing and psychological assessment introduced in Chapter 2. Psychological assessment is considered to be more extensive than psychological testing and implies the integration of multiple sources of information about someone, including their test scores, personal background information, and information about the circumstances in which they are living and working. In psychological assessment, the emphasis is on answering the referral question rather than simply providing a set of scores (Naglieri et al., 2004). This is especially true of clinical and neuropsychological assessment. It takes an experienced practitioner many years to become skilled at this and it is hard to see how a present-day computer could possibly perform this function.

One area where internet testing is expanding rapidly is industrial and organisational applications (Leivens & Harris, 2003). This is mainly due to the explosive rise of online recruiters and job markets. Many such providers seek to include job selection in their portfolio of services in an attempt to add value and attract applicants to their sites. With the entire internet full of potential readers, online recruitment promises to dramatically reduce the selection ratio, but only if your website gets noticed. A likely scenario is that specialist recruiting sites will emerge to target particular industries. There has even been the suggestion of automatic head hunting in which 'web bots' trawl the web for resumés or other information about potential candidates.

With the need for initial sifting of many applications comes the inevitable temptation of delivering psychological tests and assessments straight to the general public without the intervention of a professional psychologist. Such a method of testing has been called 'unsupervised mode' (Bartram, 2000). The idea is that anyone can log on and complete the test, anywhere and at any time. This involves a fundamental shift in how psychology, as a profession, views tests. A long tradition in psychology equates good testing practice with control over the situation by an appropriately qualified professional. The temptation to access potentially millions of users via the internet has led some entrepreneurial psychologists to challenge this assumption. Is an unsupervised test really compromised by the absence of an invigilator? On the face of it, the answer would seem to be 'yes'. Most people's first experience of testing is a formal exam at school that was invigilated very closely. Concerns about cheating were paramount. Interestingly, however, not all tests seem to need this level of close supervision.

Bartram (2000) has analysed the functions of supervision and considered four levels of supervised testing. He suggests that the main functions of a supervisor include:

1. authenticating the test taker (i.e. making sure they are who they say they are and that someone else hasn't been substituted in their place)
2. establishing rapport with the test taker
3. ensuring the test is administered according to the manual
4. preventing cheating
5. ensuring security of the test itself.

Research suggests that tests of typical performance—for example, personality or interest inventories—are not adversely affected by lack of formal supervision (Bartram & Brown, 2004). Variations in conditions are unlikely to affect an examinee's reaction to these items. However, for maximal performance tests, such as aptitude and achievement tests, answers are likely to be affected by the

presence or absence of a supervisor. In the absence of a supervisor an examinee could phone a friend, or look up the answer in a book or online encyclopaedia. This is an important issue because validity generalisation research strongly supports the use of maximal performance cognitive tests (Schmidt & Hunter, 1998), so the temptation to offer these tests in unsupervised mode is likely to be very strong indeed (Tippins et al., 2006).

In an effort to find a place for unsupervised internet testing, Bartram (2004) considered four levels of supervision. He called these open, controlled, supervised and managed modes, and his analysis was subsequently adopted by the International Test Commission in their recommendations for online testing (International Test Commission, 2005). In open mode, anyone can access the test —there is no user identification and no human supervision; examples include tests published in magazines and books. Many tests available on the internet for personal development are offered in open mode; however, tests that have incurred significant development costs, especially through the collection of good norms, are unlikely to be offered in open mode. This is the most extreme unsupervised mode and only suitable in low-stakes testing situations.

Controlled mode involves users being sent a password and logging on to a testing site. Authentication is still minimal because there is no human supervision. The idea is that open or controlled mode could serve as the first step in a selection process to initially identify unsuitable candidates; however, subsequent use of open or controlled mode results would need to be verified in a second follow-up testing session. Supervised mode involves the presence of a human supervisor, but perhaps in a non-secure environment such as an office or work site. Finally, managed mode is similar to the formal examination conditions as they occurred in school, in which access is highly controlled and the test is kept secure.

Neither open nor managed modes are conducive to carrying out a viable testing business on the internet: open mode assumes no control and managed mode is too restrictive; however, the enormous economies of scale available over the internet leads to tremendous pressure for controlled or supervised mode (Carstairs & Myors, 2009). Not surprisingly, unsupervised controlled mode is therefore the mode of choice for most professionally developed internet tests. In a further development, Bartram (2005) suggested that supervised mode might eventually be split into 'locally supervised' and 'remotely supervised' modes. Locally supervised mode is the aforementioned supervised mode, and remotely supervised mode involves a raft of surveillance technologies to carry out the supervision remotely. These include reaction time and keystroke monitoring, as well as the direct use of webcams watching the test taker. By adding a second layer of technology to the testing process, remotely supervised mode will likely raise additional complexities to those already discussed.

Internet testing also raises a number of familiar ethical issues such as the nature of the relationship between psychologist and client, confidentiality of responses, feedback to the client and how to ensure informed consent (Naglieri et al., 2004). In the first session with any new client, it is good professional practice to spend some time discussing expectations and the nature of the professional relationship. The question is: How will this be handled in the case of internet testing? Screens full of formulaic text and disclaimers are hardly conducive to building rapport or establishing a therapeutic alliance. Feedback to the client is another important issue. A psychologist providing feedback can take into account the examinee's state of mind and readiness to handle the feedback. A computer printing a canned report has no such insight. Finally, informed consent could be an issue because it might not be known whether the person taking the test is capable of giving their consent. This can be true especially for examinees with a disability. In short, the removal of human contact, as implied by internet testing, exacerbates many of the concerns raised by ethical issues in the past.

Serious gaming

'Serious gaming' is based on the development or use of games for purposes other than entertainment (Charsky, 2010). This technology is at the very early stages of being applied to psychological testing, although gaming applications (aka gamification) have been used in educational settings for teaching purposes. For example, flight simulators have long been used to train pilots, and simulations and games are being used with school and university students to allow them to work with expensive or hazardous equipment that their institutions cannot supply. These educational games and simulations also have been paired with adaptive or personalised learning: this is where the delivery platform identifies the level at which the student is performing and provides learning experiences based on the individual's progress (Chen, 2007).

The advantage of serious gaming in the testing area is that assessments are more enjoyable and engagement can be improved. Even the most mundane of tasks can be made more interesting when computer graphics are added and point scoring and badges are allocated for the successful completion of tasks (Tong et al., 2014). Some serious gaming tests have been devised. Tong et al. (2014), for example, applied serious gaming to devise a test of cognitive impairment in the elderly based on the 'whack-a mole' game (see Figure 14.3). In this computerised game/test, patients are asked to wait for a mole to appear in a hole and then tap the screen with their finger (i.e. 'whack it') to make it disappear. Performance is rated on a combination of how fast and how accurate the responses are. More

advanced serious gaming applications will be able to incorporate CAT and MAT applications, providing more sophisticated testing opportunities.

Figure 14.3 The 'whack-a-mole' game

Avatars

Finally, avatar-based technologies (aka multi-user virtual environments and virtual worlds), where provider and client operate computer-generated self-representations to interact, is a potential next step in psychological assessment. Avatar technology is an extension of earlier voice and text interaction tools such as text chatting and instant messaging, with the added benefit that self-representations function like real people in real or created environments (Witt, Oliver & McNichols, 2016). *World of Warcraft* and *Second Life* online games are current examples of this technology: here, self-representations interact with other avatars and their environments in ways similar to the manner that humans interact outside of virtual reality worlds (Bartle, 2010). This observation is supported by research that suggests that avatar interactions parallel human interactions in the real world; for example, in the use of non-verbal communication (Yee et al., 2007). Importantly for the future use of avatars in psychological testing and assessment, other research has demonstrated consistency in behaviour between avatar drivers and their self-representations (Anstadt, Bradley & Burnette, 2013). Avatar technologies have been used in

educational settings (e.g. Makransky et al., 2016) and to treat mental health problems (e.g. Yuen et al., 2013), and have the potential to function as psychological assessment tools for assessing qualities such as social competence, response to stressful situations, leadership skills, problem-solving, vocational interests, values, body image, and so on. While the problems associated with the application of any new technology need to be resolved, we might see the functions of psychological testing and assessment transformed in the coming decades.

Contextual changes

In this section we consider developments in the broader social environment—of which psychological testing and assessment form a part—that might have an effect on how testing progresses in the future. A number of forces are readily apparent. First, counter to the technological wizardry discussed in the previous section, there appears to be an increasing demand for simpler and shorter measures, and measures that can be developed quickly (Kamphaus, Petoskey & Rowe, 2000). Such measures often take the form of behaviour checklists and ratings that utilise observer judgment to document a few gross behaviours indicative of a particular disposition rather than an attempt to accurately measure behaviour to a very fine degree.

Further, the rise of managed care in the clinical domain appears to have brought with it a growing reluctance to utilise psychological assessment (Groth-Marnat, 2000a). Some authors report a drop of 10 per cent in the 30 years from 1970 to 2000. The main reasons for this appear to be concerns about the cost of testing and the apparent weak link between many forms of assessment and useful therapies. There is little doubt that burgeoning health-care costs put pressure on the ability of providers to supply everything that might seem desirable. We could be entering an era in which the costs and benefits of all potential services are compared. In this case, psychological assessment might have to compete with various physical tests, drugs and therapeutic interventions. Utility analysis might help establish the cost–benefit of psychological testing in the future.

The rapid pace of change in communications technology and in business, and the social and political upheaval experienced in many countries have seen the diffusion of ideas, people and capital across national borders in unprecedented ways. Globalisation, as it is sometimes referred to, has meant that professional bodies involved in assessment and testing are having to review their training and licensing requirements to allow for work in a global economy. Closer to the interface with clients, questions are being asked about taken-for-granted assumptions. For example, is what we call depression in Anglophone countries experienced and expressed in the same way for people from Chile or Sri Lanka or

South Sudan, or is the construct culture specific, and if so, what are the implications for clinical assessment? Or again, a firm recruiting for staff to work offshore might be more interested in an applicant's cultural competence than in their conscientiousness, even though we know a lot more about how to assess the latter. Also, what of the relevance of local norms for tests, when recruiting is being done simultaneously in Sydney, Singapore and San Francisco? Just when answers in testing and assessment are maturing, globalisation means that the questions are changing.

Finally, we live in an age of continually rising expectations on the part of the general public. There are increasing demands for accountability and transparency. Probably the best way to meet these demands is through ever more vigilance in terms of ethics and professionalism, and increasing scientific research into the validity of the tests we use.

Chapter summary

Whether or not Kyllonen's smart test becomes commonplace remains to be seen. One thing these considerations make clear, though, is that the test professional of the future will not only need to have expertise in the psychological construct that he or she is trying to measure, but also in a range of other technical and professional areas. While all of the techniques discussed by Kyllonen are feasible using present-day technology, Groth-Marnat speculated on what tests might look like in 50 years' time. His smart test of 2050 was a 'fully integrated assessment instrument using a combination of AI [artificial intelligence], interactive virtual reality (or possibly hologram), physiological measures, massive interlinked internet norms, validity/predictions based on chaos theory, branching strategies, genetic measures, in session as well as time series measures' (Groth-Marnat, 2000b, p. 361). There is no doubt that the field of psychological testing and assessment will continue to offer challenges and opportunities to both theorists and practitioners for many years to come.

Questions

1. What new constructs could emerge to fuel a new generation of tests?
2. How would a smart personality test work?
3. Think of a pencil-and-paper test you are familiar with, perhaps one of those discussed previously in this book, and try to think of a better way of measuring the construct via computer. Try to utilise some of the potential of computerised testing that is not possible on paper.
4. Discuss some of the ethical issues raised by internet testing.
5. How can psychological testing and assessment be justified from a cost–benefit point of view?

Further reading

Clauser, B E, Margolis, M J & Clauser, J C (2015). Issues in simulation-based assessment. In F Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 49–78). New York, NY: Routledge.
Drasgow, F (Ed.) (2015). *Technology and testing: Improving educational and psychological measurement*. New York, NY: Routledge.
Groth-Marnat, G (2000). Visions of clinical assessment: Then, now, and a brief history of the future. *Journal of Clinical Psychology*, *56*, 349–85.
Mislevy, R J, Corrigan, S, Oranje, A, DiCerbo, K E, Bauer, M I, von Davier, A A & John, M (2016). Psychometrics and game-based assessment. In F Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 23–48). New York, NY: Routledge.
Plake, B S & Witt, J C (Eds.). (1986). *The future of testing*. Hillsdale, NJ: Erlbaum.
Popp, E C, Tuzinski, K & Fetzer, M (2015). Actor or avatar? Considerations in selecting appropriate formats for assessment content. In F Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 79–103). New York, NY: Routledge.

Useful websites

ETS: www.ets.org
I/O psychologists get wired (American Psychological Association): www.apa.org/monitor/julaug05/wired.aspx
ITC guidelines on computer-based testing (International Test Commission): www.intestcom.org
New constructs (Educational Testing Service; ETS): www.ets.org/research/capabilities/assessment_research/new_constructs
The rapid rise of psychological testing in selection (Australian Psychological Society): www.psychology.org.au/Content.aspx?ID=4925

Answers to Exercises
Chapter 3

1. Mean of scores = 60, SD of scores = 5.2

Scores	z scores	Transformed scores with a mean of 100 and SD of 15
52	–1.54	77
54	–1.15	83
56	–0.77	89
58	–0.38	94
60	0.00	100
61	0.19	103
61	0.19	103
63	0.58	109
67	1.34	120
68	1.54	123

2. a. Percentage correct scores:

	Hassan	Brett	Zhang Wei
Geography	61.3%	96.0%	80.0%
Spelling	73.3%	66.7%	64.5%
Mathematics	75.0%	82.5%	92.5%

b. z scores

	Hassan	Brett	Zhang Wei
Geography	–1.40	1.20	0.00
Spelling	0.50	0.00	–0.15
Mathematics	1.00	1.60	2.40

c. Percentiles

	Hassan	Brett	Zhang Wei
Geography	8th	88th	50th
Spelling	69th	50th	44th
Mathematics	84th	95th	99th

d. T scores and average T scores

	Hassan	Brett	Zhang Wei
Geography	36	62	50
Spelling	55	50	48.5
Mathematics	60	66	74
Average T score	50.3	59.3	57.5

3.

a.

Scores	z scores
16	–2
18	–1
19	–0.5
20	0
21	0.5
22	1
24	2

b. Percentage of scores that fall between:

18 and 22 = 68.26 per cent
19 and 21 = 38.30 per cent
16 and 24 = 95.44 per cent

4. a. The percentile of a score with a z score of 1.0 is 84.
b. The z score of score at the 98th percentile is 2.05.
c. The T score for a score with a z score of 2.0 is 70.

5. For a test with a mean of 30 and an SD of 10

Raw score	z score	Percentile
40	1	84
35	0.5	69
36.7	0.67	75

6. Using Table A1 in the Technical Appendix, there are 0.3413 of cases between the mean and a z score of 1.00 and 0.3749 cases between the

mean and a z score of 1.15. Therefore there are (0.3749 – 0.3413) cases between a z of 1 and a z of 1.15, or approximately 3 per cent.

7. Tanya shows an improvement of 10 points (520 – 510). For Nehir to show an equal amount of improvement on an equal-interval scale her score must be 500 (490 + 10).

8. W = 500 is the midpoint of the scale with a logit value of 0 [W = 9.1024(logit) + 500]. Therefore values of 1.5 and 0.5 are above the value of an item she has a 50 per cent chance of getting correct and –0.2 is below it. Therefore:

   a. Less likely
   b. Less likely
   c. More likely

9. $$\begin{aligned} &\mathrm{W} = 9.1024(\mathrm{logit}) + 500 \qquad \mathrm{Logit}{=}1.5{:}\mathrm{W}{=}513.65 \approx 514 \\ &\mathrm{Logit}{=}0.5{:}\mathrm{W}{=}504.55 \approx 505 \quad \mathrm{Logit}{=}0.2{:}\mathrm{W}{=}498.18 \approx 498 \end{aligned}$$

   The logit is the log (to the base e) of the odds of getting the item right. The odds of getting the item right is then the antilog of the logit value.
   Odds is the probability of getting the item right divided by 1 minus that probability. Therefore the probability is odds divided by 1 plus the odds.

$$\begin{aligned} \mathrm{p} &= \tfrac{\mathrm{odds}}{1+\mathrm{odds}} \\ \mathrm{odds} &= \tfrac{\mathrm{p}}{1-\mathrm{p}} \end{aligned}$$

   Therefore to find the p of getting the item right knowing the logit value for a given theta (Paula's ability level), first find the antilog of the logit value (using a scientific calculator or an online maths website) and apply the formula p = odds/1 + odds.

$$\begin{aligned} \mathrm{Logit} = 1.5 \quad \mathrm{Antilog} = 4.4816 \quad \mathrm{P} &= \tfrac{4.4816}{1+4.4816} \\ &= 0.82 \\ \mathrm{Logit} = 0.5 \quad \mathrm{Antilog} = 1.6487 \quad \mathrm{P} &= \tfrac{1.6487}{1+1.6487} \\ &= 0.62 \\ \mathrm{Logit} = -0.2 \quad \mathrm{Antilog} = 0.8187 \quad \mathrm{P} &= \tfrac{0.8187}{1+0.8187} \\ &= 0.45 \end{aligned}$$

10.

$$\mathrm{SE} = \frac{\mathrm{SD}}{\sqrt{\mathrm{N}}}$$

$$\mathrm{SD} = \mathrm{SE} \times \sqrt{\mathrm{N}}$$

For $\mathrm{SE} = 0.5$ and $\mathrm{N} = 100$

$$\begin{aligned} \mathrm{SD} &= 0.5 \times 10 \\ &= 5 \end{aligned}$$

Assuming SD for the new sample is approximately the same and the

$$0.25 = \frac{5}{\sqrt{\mathrm{N}}}$$

$$\begin{aligned} \mathrm{N} &= \left(\frac{5}{0.25}\right)^2 \\ &= 400 \end{aligned}$$

Chapter 4

1. Mean for test = 2.0, SD = 1.91
   Variance for test = 1.91^2 = 3.65

Mean for item	SD for item	Variance for item
0.13	0.33	0.11
0.11	0.32	0.10
0.11	0.37	0.14
0.06	0.24	0.06
0.21	0.41	0.17
0.08	0.28	0.08
0.08	0.27	0.07
0.19	0.39	0.15
0.11	0.31	0.10
0.23	0.42	0.18
0.01	0.12	0.01
0.10	0.30	0.09
0.15	0.36	0.13
0.01	0.13	0.02
0.11	0.31	0.10
0.01	0.09	0.01

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum \sigma i^2}{\sigma t^2}\right) = \left(\frac{16}{16-1}\right)\left(1 - \frac{1.50}{3.65}\right) = 1.07 \times 0.59 = 0.63$$

2. a. Answer provided by student.
   b.

Test	Reliability coefficient	SD	SEM
A	0.85	15	5.81
B	0.85	5	1.94
C	0.55	15	10.06
D	0.55	5	3.35

Tests A and B and Tests C and D have the same reliabilities and hence the scores within each pair of tests are of the same accuracy. The SEM is expressed in the raw score metric of the test and it so happens that scores for Tests A and C have a wider range than those for Tests B and D. This does not mean they are less accurate. Put another way, you wouldn't say that scores on the subtests of the WAIS are more accurate than Full Scale IQ, even though the SEMs are smaller for the subtests.

   c. The SE_{diff} should be larger than the SEM of the two subtests.

3. Reliability of ASAT = 0.90, SD of ASAT = 15

   a. The reliability of the ASAT can be considered high.
   b. Despite its high reliability, it should be noted that a score obtained by an individual may not be 100 per cent accurate.

$$\text{SEM of ASAT} = \text{SD}\sqrt{[1 - r]} = 15\ \sqrt{1 - 0.90} = 4.74$$

We would, however, argue that the cut-off of 115 should stand and the student not be admitted as to do otherwise is to shift the cut-off to 112 and then raise the same set of issues for someone with a score of 111. Scores have errors but we have to make decisions oftentimes on what we have got.

4. Inspection of the table indicates that Clinician A consistently rates patients as more improved than does Clinician B, by at least 10 points and up to almost 20. Their rank order of patients agrees, as the product moment correlation indicates, but there is a systematic difference between them. Analysis of variance on these date indicates a statistically significant difference (p = 0.03) between raters and an intraclass correlation of only 0.04. The product moment index ignores this systematic difference.

1. Student responses will vary as to the best cut-off score.

   a. If a cut-off score of 31 is used, the new test will have perfect discrimination; that is, all 10 members of the prison population will be correctly classified. It is not necessary to calculate the validity coefficient. If it is calculated, with a split like that you will obtain a phi coefficient of 1.0. That is, the correlation between test score and classification is 1.0, so the test has perfect predictive validity.

   b. The valid positive rate is 30 per cent (percentage of psychopaths correctly classified).

   c. Although the test seems to have perfect discrimination using a cut-off score of 31, more information (i.e. base rate and selection ratio) is needed to properly evaluate the utility of the test in a prison sample. The problem becomes greater if we try to use it in a non-prison or community sample. Although it looks like a great test, it may lead to a number of misclassifications.

2.

False negative = 0.10	Valid positive = 0.20	Base rate = 0.30
Valid negative = 0.60	False positive = 0.10	1 – base rate = 0.70
1 – selection ratio = 0.70	Selection ratio = 0.30	

3.

		Test		Rating	
		EQ	IQ	EQ	IQ
Test	EQ	0.75			
	IQ	0.6	0.92		
Rating	EQ	**0.8**	0.5	0.45	
	IQ	0.5	**0.8**	0.4	0.45

There is no one correct answer to this question. Students are expected to show their understanding of the principles of the multitrait–multimethod matrix (e.g. criteria for convergent and discriminant validity) by making up values for the rest of the table. The values (in bold) included above are examples that point to the validity of the new test.

4. The Vocabulary and Reasoning Tests correlate but neither correlate to any marked degree with the Dexterity and the Mechanical Reasoning Tests, which do correlate with each other to some degree. Two factors are suggested: a Verbal factor and a Practical Ability factor.

5. If there are two constructs, two factors would be predicted. The factors could be correlated or uncorrelated depending on the theory underlying the test. If correlated, however, the relationship would not be so strong that one factor would provide a more parsimonious model of the test.

Chapter 6

1. a. In analysing the items, students can consider: the variances of the items, by squaring the standard deviations provided in Chapter 4; and item-total correlations. 'Good' items are those that have large variances and large item-total correlations.

   b. Spearman-Brown formula:

$$k = \frac{r_{yy}[1 - r_{xx}]}{r_{xx}[1 - r_{yy}]} = \frac{0.90[1 - 0.63]}{0.63[1 - 0.90]} = 5.29$$

   No. of items required for a desired reliability of 0.90 = 5.29 × 16 = 84.6.
   No. of extra items needed = 85–16 = 69

2. Students can check their own answers in the text

3. a. No, 50/50 chance of correct
   b. No, question may confuse individuals with a language problem
   c. No, double/triple negative
   d. No, too easy

4. a. Random responding (the item is unlikely to be answered positively; if it is it is likely that the question has not been read before giving an answer)
   b. Social desirability (the item is unlikely to be answered positively; if it is, this suggests that the respondent is trying to create a favourable impression)
   c. Anxiety (tenseness is a common symptom of anxiety)

5. a. Unclear what is meant by the question or what any answer to it might indicate
   b. Unlikely to be answered positively by other than a very small percentage of people, and therefore not a discriminating item
   c. A shorter version would be easier to understand

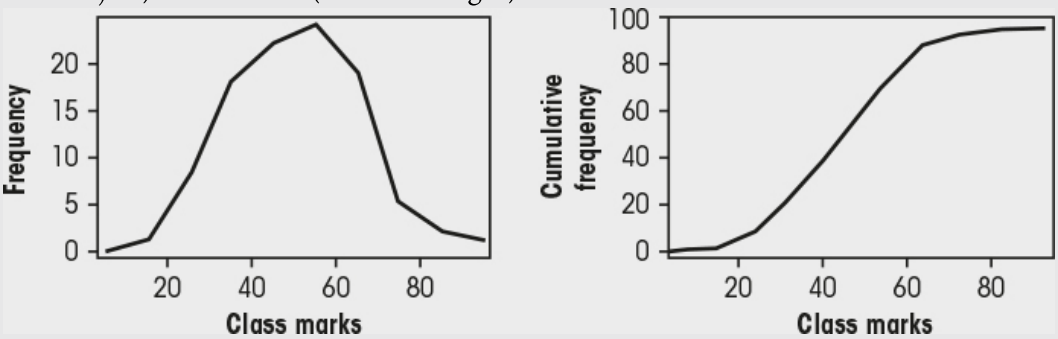
6. a. Answer depends on a good knowledge of geography and may be too specialised
   b. Options differ in difficulty level for the intended audience from (i) quite difficult to (ii) too easy. The third option (iii) is at about the right level of difficulty
   c. A double barrel question: a respondent could answer True to global warming and False to industrial development
   d. Quite specialised knowledge called for with two very attractive distractors (i and ii)

TECHNICAL APPENDIX

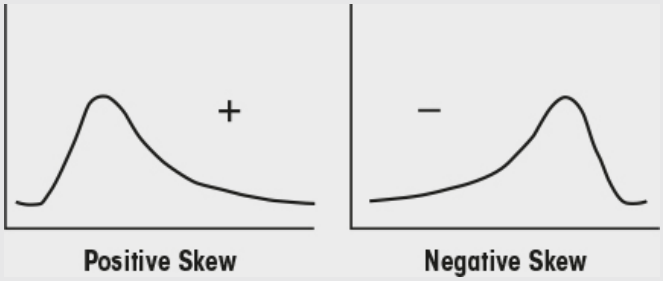
MEASUREMENT STATISTICS

Score Distributions

Population: The entire collection of values on a variable of interest (e.g. test scores of all Australian adult males on an intelligence test).
Sample: A set of the possible values drawn from the population; depending on how the sample is drawn it may or may not be representative of the population.
Frequency distribution: A way of logically organising the values in a sample to reveal certain of its features (e.g. counting the frequency of scores arranged from lowest to highest).
Frequency polygon: A graph (histogram) of the frequency distribution for a set of scores, with the heights connected by line segments (see below left). Where many scores have low frequencies, individual values can be grouped into categories (class intervals) and the upper points of the class intervals joined by lines to make for a more compact graph. When frequencies are cumulated (added from the lowest to the highest value) and plotted, a *cumulative frequency curve or ogive* (pronounced ojive) is the result (see below right).



Uniform score distributions: These have the same frequencies (approximately) across the range of scores.
Symmetric score distributions: These have frequencies that rise from low values to a maximum and then fall as values continue to rise. If there is one maximum (peak) the distribution is unimodal (see above left); if there are two peaks the distribution is bimodal.
Asymmetric distributions: These have high frequencies for one range of scores and low frequencies for another. Where low scores have high frequencies the distribution is positively skewed (see below left); where high values have high frequencies the distribution is negatively skewed (see below right).



Weisstein, E W (n.d.). Cumulative frequency polygon. *Wolfram MathWorld.* Retrieved from http://mathworld.wolfram.com/CumulativeFrequencyPolygon.html

Simple descriptive statistics for single variables

Mean (M)
The average of the N values X_1 ... X_i

$$M = \sum X\iota/N$$

Variance[1]

$$S_X{}^2 = \sum (X_i - M)/N (X_i - M \; is \; the \; deviation \; score \; of \; an \; X \; score \; from \; the \; mean)$$
$$S_X{}^2 = \sum (X_i - M)/N - 1$$

Standard deviation (SD)

$$\begin{aligned} \mathrm{SD} &= \sqrt{s_x^2} \\ \mathrm{SD} &= \sum \sqrt{(X_i - M)^2/N} (\mathrm{sample}) \\ \mathrm{SD} &= \sum \sqrt{(X_i - M)^2 (N-1)} (\mathrm{population\ estimate}) \end{aligned}$$

Median (Q2, 50th percentile)

The middlemost value when the X_i values are ranked.
When N is odd, the (N+1)/2 term
When N is even, the average of the N/2 and (N/2) + 1 terms

Interquartile range (IQR)

The range within which the middle 50 per cent of values lie.

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$$

Q1 (25th percentile), the middlemost value (median) when the values from the median to the lowest value (minimum) are ranked

Q3 (75th percentile), the middlemost value (median) when the values from the highest value (maximum) to the median are ranked

Normal curve

Equation of the normal curve

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ = Mean

σ = Standard Deviation

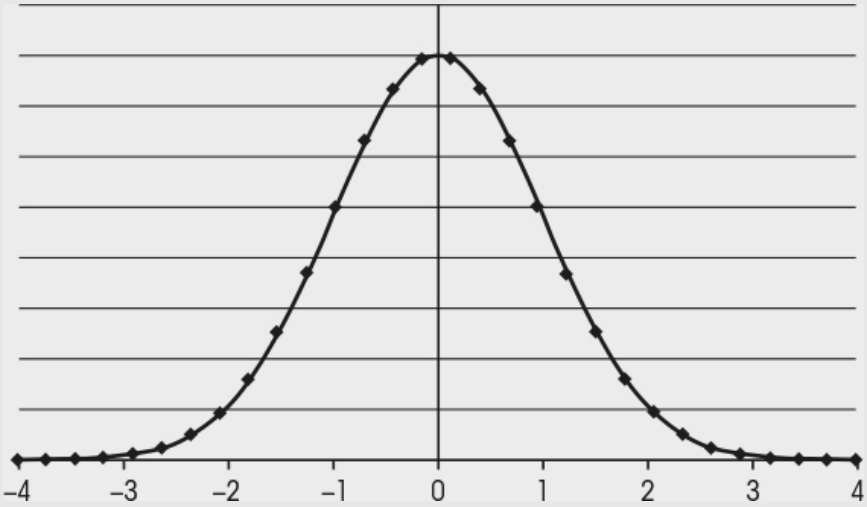
$\pi \approx 3.14159\ldots$

$e \approx 2.71828\ldots$

Taylor, C (n.d.). Formula for the normal distribution or bell curve. *About Education*. Retrieved from http://statistics.about.com/od/Formulas/ss/The-Normal-Distribution-Or-Bell-Curve.htm

Standard normal distribution

Height of the curve (y) at any point (x) is a function of the mean (μ) and the standard deviation (σ) and two constants, Pie (π) and Euler's number (*e*). Hence knowing the mean and standard deviation, y can be determined for any value of x. With mean set to 0 and standard deviation to 1, we have the standard normal distribution (μ=0, σ=1).



Values along the base are not raw scores but z scores; that is, the deviation of the raw score from the mean divided by the standard deviation: $z = (X_i - M)/SD$.

Tables of the normal curve

Rather than solve the equation each time, tables of the normal curve are available (see Table A1 at the end of this appendix).

The table is set up in three columns. The first lists values of z; the second lists the proportion of cases in the distribution that lie between the mean and the value of z; and the third is the proportion of cases that lie beyond that value of z. The tabled values are all positive, but because

the normal curve is symmetrical about the mean, the values for negative zs are the same. Note that the values in columns 2 and 3 sum to 0.5 in all cases because half the scores in a normal distribution lie above the mean and half lie below it. Not included in this table, but included in some versions of the tables of the normal curve, is the height of the curve corresponding to the z score.

To determine how many cases there are in the distribution up to a positive z score, enter the table with the z score, read off from column 2 the number of cases between the mean and the z score and add 0.5 (half the cases lie below the mean). For example, for +0.21, column 2 indicates there are 0.0832 cases between the mean and +0.21. Therefore, there are 0.5 + 0.0832 cases or 0.58 of cases up to that z. That is, if a distribution of scores is normally distributed, the expectation is that 58 per cent lie below the raw score corresponding to a z score of +0.21. A negative z score lies below the mean and hence the third column is the relevant column. For example, for −0.58 there are 0.2810 of the cases to that point in a normal distribution.

Percentiles

Graphing interpolation
Group the raw scores into convenient class intervals, count the frequencies of scores in each of the intervals, cumulate the frequencies beginning from the lowest score intervals, and express the cumulative frequencies as percentages. From a smooth curve fitted to the plot of the cumulative percentages against the midpoints of the class intervals, read the percentile corresponding to any particular raw score. (see Cronbach, 1990, pp. 110–11).

Arithmetic calculation

Plot the cumulative percentage curve as in graphical interpolation, and apply the formula:

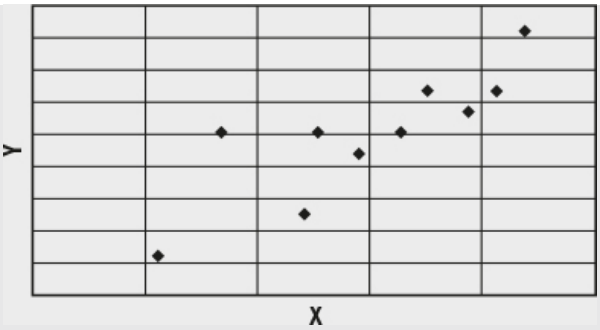
$$\mathrm{PR} = \frac{cf_1 + .5(f_1)}{N} \times 100\%$$

where:
PR is the percentile rank
cf_1 is the cumulative frequency for all scores lower than the score of interest
f_1 is the frequency of scores in the interval of interest
N is the total number of cases in the sample.

Two variables

Scatterplot (X,Y)

A plot of ordered pairs of two variables (X and Y) to show the relation between them.



To summarise the relation

Covariance

Average of the cross products of deviation scores

$$\mathrm{COV}_{(X,Y)} = \frac{\sum[(X_i - M_X)(Y_i - M_Y)]}{N}$$

Correlation

For continuous scores

$$\begin{aligned} r_{(X,Y)} &= \frac{COV_{(X,Y)}}{(SD_X SD_Y)} \\ r_{(X,Y)} &= \frac{\sum(z_X z_Y)}{N} (average\ of\ the\ cross\ products\ of\ z\ scores) \end{aligned}$$

When one variable is continuous and one is dichotomous

$$r_{pb} = \frac{M_{Y1} - M_{Y0}}{SD_x \sqrt{pq}}$$

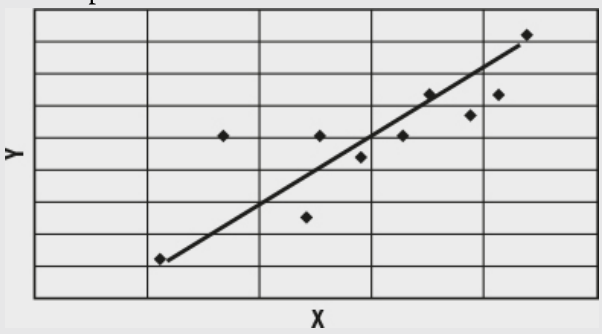
where M_{Y1} is the mean of scores on the continuous variable for one group on the dichotomous variable (those assigned a score of 1) M_{Y0} is the mean for the other group (those assigned a score of 0), SD_X is the SD for the combined groups on the continuous variable, and pq is the product of the proportion of the sample in one group (those assigned a score of 1) and q is the proportion of the sample in the other.
When both variables are dichotomous

$$Phi = \frac{\sqrt{\chi^2}}{N}$$

Where χ^2 is the chi square value from a 2 × 2 contingency table formed from the cross-break of the two variables and N is the number of cases.

Regression

Fitting a straight line to a scatterplot



$$Y = bX + a(\text{equation of a straight line})$$
$$Y_i^{\text{“}} = bX_i + a(\text{linear prediction of a Y score from a score on } X)$$

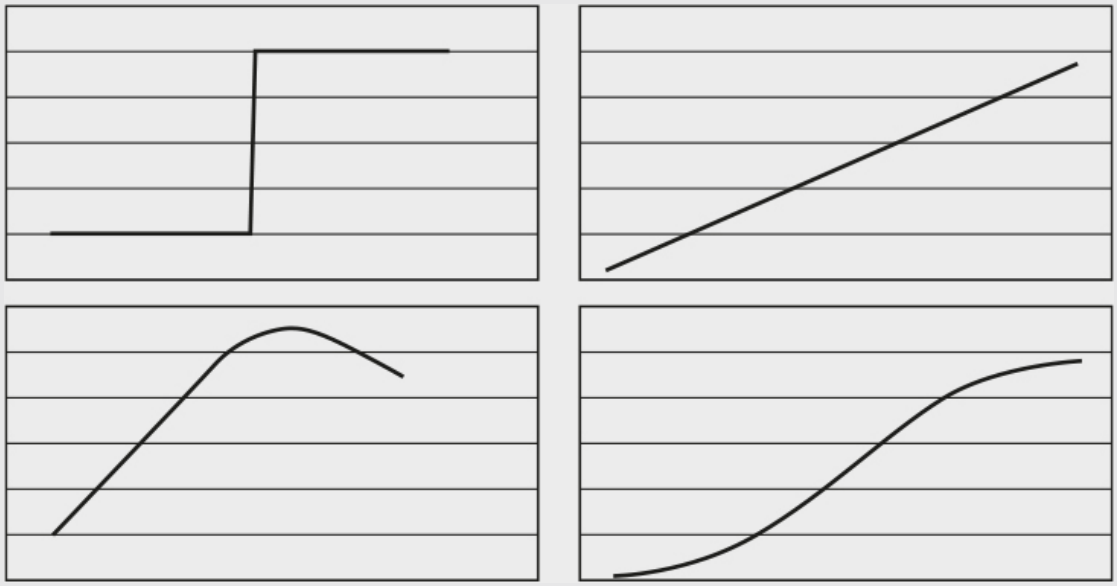
where Y_i' is the predicted score on Y, X_i is the score on X, b is the regression coefficient (slope of the straight line) and a is the point on the Y axis (intercept) where the straight line crosses it.

$$\begin{aligned} b &= \frac{COV_{(X,Y)}}{s^2} \\ a &= M_Y - bM_X \\ Y_i^{\text{“}} &= r_{(X,Y)}X_i(\text{when X and Y are in z score form, } b = r) \end{aligned}$$

Measurement models

A trace line or item characteristic curve (ICC)[2] relates the probability of endorsement of an item (for a dichotomous item, whether the respondent gives the correct answer) to the respondent's position on the underlying attribute[3] of interest.

Possible functions that might be adopted as models of the ICC for psychological measurement scales



Top left: a deterministic model after Guttman (1944). Responding with an incorrect response is constant at zero until some point on the attribute scale is reached, at which point the probability of responding with the correct answer becomes 1 (all those with scores on the attribute below the point of inflexion get the item wrong and all those with scores above get the item right). The model is deterministic in the sense that the probability is either 0 or 1.0 and no intermediate probabilities are possible.

Bottom left: Thurstone's (1929) approach to scaling attitudes. The probability of responding to the item in a particular way increases up to some point on the attribute scale and then begins to decrease. It is difficult to find instances that fit this model.

Top right: The probability of responding in a particular way is a simple linear function of attribute strength. This is a seldom used model for scaling psychological tests, partly because it is not a good fit to most empirical trace lines, and partly because it implies that the probability of responding to an item can be less than 0, which makes no sense.

Bottom right. The probability of responding is a monotonic function of attribute strength. (Monotonic means that the function does not change direction once it begins, unlike the trace lines on the left of the figure above.) Both curves on the right are monotonic, but the one on the top right has the added constraint of being linear. The bottom right model is the most used in psychological measurement, either without further constraints on its form (classical test theory, sometimes called 'weak' true score theory) or with specific requirements about form (item response theory).

Classical test theory (CTT)

The theory began with Spearman in 1904. In 1950 the work to that time was systematised by Harold Gulliksen in *Theory of Mental Tests*. More recent treatments are those by Lord and Novick (1968) and McDonald (1999).

To determine the person's position on the underlying attribute, CTT focuses on the combination (usually linear) of scores for items on a test. That is, the focus is on the sum of the item scores (the observed score) rather than the items themselves and the concern is with the reliability of the total score as an indicator of a person's standing on the attribute (their *true score*).

The theory makes five basic assumptions about a test score:

1. An observed score (the score a person obtains) on a test is the sum of two components: a true score and an error score component.

$$Xo = T + e$$

2. The true score is the population mean of the observed scores. (If, for example, one were able to repeatedly administer a test to a person, then the long-term mean of the scores so obtained would be the true score for that person.)

$$E(Xo) = T$$

where E(Xo) means the expected value (population mean) of the observed scores.

3. The correlation between true score and error score components is zero. (Errors are random and therefore cannot relate systematically to any other variable.)

$$\rho Te = 0$$

where ρTe means the correlation between T and e.

4. The correlation between error components on two tests is zero. (Again, errors are random.)

$$\rho ee' = 0$$

where e and e′ are the error components of observed scores on two tests.

5. The correlation between the error component of the observed score on one test and the true score component on another is zero. (Again, errors are random.)

$$\rho eT' = 0$$

A number of propositions can be derived from these assumptions for use in developing and evaluating psychological tests. Several of these are simply stated here, without the corresponding derivation from the basic assumptions.

There is a reciprocal relation between true score and error score variances such that reducing one increases the other.

The square of the correlation between true and observed scores is the proportion of true to observed score variance. The correlation of a test with itself ($r_{XX'}$) provides such an estimate.

The standard deviation of error scores—the standard error of measurement—can be estimated from knowledge of the standard deviation of observed scores and the correlation of a test with itself:

$$\sigma_e = \sigma_o \sqrt{1 - r_{xx'}}$$

The correlation of a test with itself is related to the number of items that make up the test and the average intercorrelation of the items:

$$r_{xx'} = \frac{k\bar{r}_{ij}}{(1 + -1)\bar{r}_{ij}}$$

A test can be made more reliable by increasing its length:

$$r_{kk'} = \frac{kr_{xx'}}{1 + (k - 1)r_{xx'}}$$

where k = the factor by which the test has to be lengthened to increase the reliability form $r_{xx'}$ to $r_{kk'}$.

The correlation between observed scores is always less than the correlation between true scores. Error serves to *attenuate* the correlation. This means that the reliabilities of two tests place an upper limit on their intercorrelation.

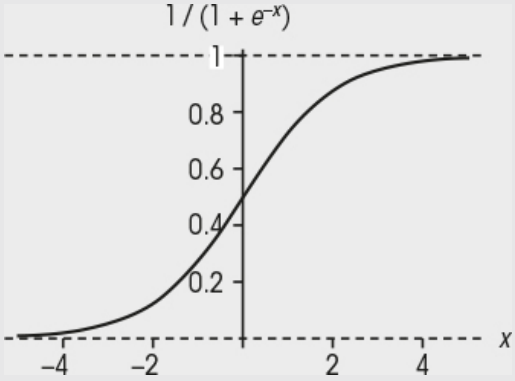
$$r_{xy} = \sqrt{r_{xx'}}\sqrt{r_{yy'}}$$

An essential concept in this theoretical framework and one fundamental to test reliability is the idea of a test correlating with itself. This gave rise to the idea of a *parallel test* or one that is so like the first that it does not matter which one is used, as Gulliksen put it (Osterlind, 2005). A very strict set of conditions was first used to operationalise this likeness, but these proved difficult to meet in practice (e.g. how can the equivalence of error distributions for two tests be assessed?). Nowadays, those using CTT rely on tests being *essentially tau equivalent* or *congeneric* rather than parallel. Essential tau equivalence does not require two tests to have equal means and equal error variances. It does require the tests to assess the same construct and assess it equally well and, further, that the tests be independent of each other (i.e. a response to an item on one test has no influence on response to an item on the other test). Congeneric tests require a uniform linear relationship, as with essentially tau equivalent tests, and that error variances are random and symmetric for the population. One implication of all this is that reliability indices that are developed from CTT and used in psychological testing can vary with circumstances and cannot be considered fixed or absolute.

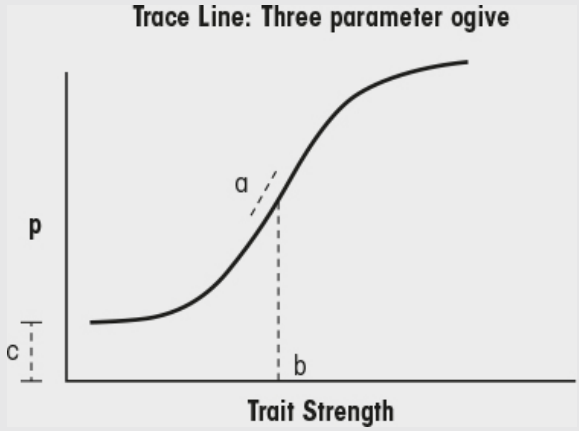
Item response theory (IRT)

IRT focuses on the response to individual items in a way CTT does not. As long as items conform approximately to a normal ogive, CTT will accommodate them. Some items may be found in an item analysis of a test to be better than others in terms of their correlation with total score or their frequency of endorsement, but the fact that they may be of different forms is of no particular concern in determining total score. IRT is interested in form. The logistic function is frequently assumed as a useful model for item responding. This function has been adopted in biological research to model growth. Probability of an event rises slowly and is then exponential before slowing and at maturity stops. The standard logistic function is shown below, where the maximum

is 1, the x value corresponding to the midpoint of the curve is zero, and the steepness of the curve is 1.



When used as a model of item responding, three parameters are of interest: where on the y axis it begins (c), the steepness at the point of the inflexion (a), and where the curve is located above the x axis (b).

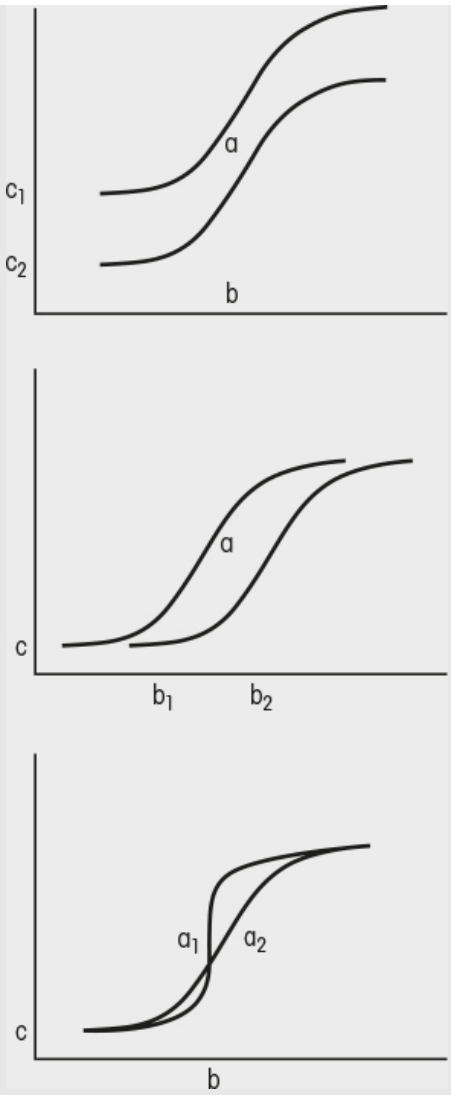


A model that includes all three parameters, a three parameter logistic model (3PL) is:

$$P(\theta) = c + \frac{1 - c}{1 + e^{-Da(\theta - b)}}$$

where θ is the position of the person on the latent trait to be estimated
a, b and c are the discrimination, difficulty and guessing parameters respectively
D is a constant (1.7) that makes the shape of the function similar to the cumulative normal distribution function;
e (Euler's number) is the base of the natural logarithm, a constant (2.71828 approximately).

Examples of curves with different parameters are shown below.



The top figure shows the ICCs for two items differing in the guessing parameter (a and b are the same); guessing is more marked for the item with the higher ICC on the Y axis. The middle figure shows two ICCS differing in b (a and c are the same). The curve on the left is for an easier item than the curve on the right. The bottom figure shows the ICCs for two ICCs differing in a (c and b are constant). The item with the steeper ICC is more discriminating. Although 3PL (all three parameters included) and 2PL (omitting the guessing parameter) models are studied, the 1PL model is the one most used in practical test development. The parameter of interest here is b (difficulty), and the other parameters are assumed not to influence the outcome (c) or are held constant for all items (a). A particular version of the 1PL model, in which the discrimination parameter (a) is fixed at 1, is widely used and is known as the Rasch model after its originator.

Rasch model

The Rasch model is a model of the probability of responding to an item conditional on the difficulty of the item and the position of the person on the latent trait. Formally, the Rasch model proposes that the probability of response to an item given a person's standing on the latent trait:

$$P(\theta) = \frac{1}{1 + e^{-D(\theta - b)}}$$

The model is similar to the 3PL model above but simpler, with the *a* and *c* parameters being removed. The expression can be written, ignoring the scaling factor (D), for the simplest case where there are only two alternative responses to the item (e.g. agree/disagree, yes/no or correct/incorrect)

$$P(x = 1|\theta) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}}$$

which is read as the probability that x = 1 (say, getting the item right) dependent on the person's position on the latent trait is equal to e raised to (θ – b), the position on the latent trait minus the difficulty of the item, divided by 1 plus e raised to (θ – b). That is, it is an expression for the probability of getting the item right in terms of the difference of the person's position on the attribute underlying response to the item and the difficulty of the item. The probability of getting the item wrong (1 minus the probability of getting it right) is then

$$P(x = 0|\theta) = \frac{1}{1 + e^{(\theta - b)}}$$

Now the odds (as distinct from the probability)[4] of getting the item right is P/1 – P; that is, the probability of getting an item right divided by 1 minus that probability (i.e. the probability of getting the item wrong). We have those two probabilities in the preceding two equations, and so forming the ratio of the two probabilities and simplifying we have:

$$\begin{aligned}
&\text{Odds}(x = 1) = e^{(\theta - \beta)} \\
&\text{and} \\
&\log \textit{to base e odds}(x = 1) = (\theta - \beta) \\
&\textit{or} \\
&\text{In } \textit{odds} = (\theta - \beta)
\end{aligned}$$

With θ constant, the log odds of getting the item right get shorter as b increases and get longer as *b* decreases. With b constant, the log odds of getting the item right increase as θ increases and

get shorter as θ decreases. For an item of known difficulty, the difference in log odds represents the difference in θ. The comparison in terms of θ holds for different items, as long as item difficulty is held constant.

ln(odds) is the logit (the inverse of the logistic function), which provides a metric for position on the latent trait and item difficulty. Logits can take values between plus and minus infinity but in practice ± 3 to 4 logits cover the range of test scores. A logit of 0 corresponds to an item which has a .5 probability of a correct answer for a person whose odds of getting the item right is 50/50. Positive logits indicate increased difficulty and higher position on the latent trait and conversely negative logits indicate decreased difficulty and lower position on the latent trait.

Tests using the logit scale can be thought of as expressing the respondent's score on the test as

$$Test\ Score = \mathrm{In}\left(\frac{Per\ cent\ Correct}{1 - Per\ cent\ Correct}\right) + Average\ Difficulty$$

That is, test scores increase as item difficulty increases, when percentage correct is held constant, but when item difficulty is held constant, test score is not a linear function of percentage correct. It is the sigmoid function shown at the beginning of the IRT section of this appendix. The test score based on the Rasch model will be monotonically related to the simple summation of items using CTT, but will not be the same. It is a non-linear transformation.

The Rasch model allows for the fact that a test may not be equally reliable at all points on the latent trait. According to CTT the standard error of measurement (SEM) is the same for all individuals irrespective of their standing on the underlying attribute. Empirically this is known not to be the case, and with IRT this can be examined. The amount of psychometric 'information' about the latent trait that is provided by each item at each level of the trait can be computed as well as for the test as a whole. The Item Information Function (Ii) is simply the product of the probability of getting the item right (Pi) and 1 minus that probability: $I_{i(\theta)} = P_{i(\theta)}Q_{i(\theta)}$. It is at a maximum when 50 per cent of respondents get the item wrong and 50 per cent get it right. The test information function is the sum of the item information functions. The SEM is inversely related to the information function: $SEM(\theta) = 1/\sqrt{I(\theta)}$.

The application of the Rasch model to a set of test score data is shown in Thorndike (1982) and in Furr and Bachrach (2014), but specialist software is needed in practice, such as Winsteps (Linacre, 2001) or RUMM (Rasch Analyst, 1996).

The Rasch model requires that a number of assumptions be made about the data. These include the assumptions that the test is unidimensional (i.e. that only one construct or latent trait is being assessed). A further assumption is that guessing is not involved in responding to the items, and that the items are equally discriminating (i.e. they do not vary in the slope of the function), which critics have suggested are unrealistic. It depends as well on the assumption of **local independence**: the only factors to enter into determining response to an item are the respondent's standing on the underlying trait and the difficulty of the item. Using information from an earlier item to answer a later item is a clear violation of local independence.

local independence
the situation where the only factors influencing response to a psychological test item are the item's difficulty and the respondent's position on the underlying trait; for example, exposure to

other items of the test does not increase or decrease the probability of responding in a particular way

Intraclass correlation

Consider a situation in which we have made d repeated measurements on n participants on a continuous variable Y. The repeated measurements might arise from testing and then retesting the participants on a psychological test, or they might be ratings by a number of judges on a psychological dimension. In the simplest case, we have two repeated measurements (d = 2); that is, occasion 1 and occasion 2 of testing or judge 1 and judge 2.

Participants	Measurement 1	Measurement 2
1	Y11	Y12
2	Y21	Y22
3	Y31	Y32
4	Y41	Y42
5	Y51	Y52

Total variance in Y can then be partialled into an effect due to participants (individuals differ), an effect due to difference in measurements (different scores on different occasions or different ratings by different judges), and to error of measurement (e.g. momentary lapses in attention). The latter is not a systematic effect (the particular lapse of attention is unlikely to be repeated) in the way that the effect due to measurements may be. For example, having completed the test once, participants achieve a better score on the second occasion (a practice effect) or one judge is generally more conservative (or lenient) than another in their ratings.

$$\sigma^2 Y = \sigma^2 P + \sigma^2 M + \sigma^2 e$$

Estimates for these variances are calculated from a two-way analysis of variance. McGraw and Wong (1996) indicate how this is done, or one can use SPSS.

Two estimates of reliability (ρ) can be derived in this situation

$$\rho_i = \frac{\sigma^2 P}{\sigma^2 P + \sigma^2 M + \sigma^2 e}$$

That is, the variance due to participants is expressed as a proportion of total variance, which will increase as the variance due to measurements and measurement error is reduced.

$$\rho_{pm} = \frac{\sigma^2 P}{\sigma^2 P + \sigma^2 e}$$

Here, variance due to participants is expressed as a proportion of total variance less variance due to measurements.

These are both intraclass correlation coefficients (ICCs), but the second is the usual Pearson product moment correlation, which is typically applied where the repeated measurements are occasions of testing (i.e. the test-retest estimate of reliability). It differs from the first ICC in that the systematic effect due measurements (the practice effect or a fatigue effect, if measurements decease on the second occasion) is not included. Some argue that it should be—that the first of the two ICCs should be used in assessing test-retest reliability—and some argue that it should not. All would agree that the first ICC should be used when measurements are across judges; that is, when inter-rater reliability is being assessed. In this case one is interested in the degree of agreement between judges and not simply in whether they rank participants in the same order.

More on the decision theoretic approach to validity

In the clinical literature, the valid positive rate is only one of index of test validity . The valid positive rate is the number of valid positives divided by the total number of persons tested. The hit rate, on the other hand, is the sum of number of valid positives and the number of valid negatives divided by the total number of persons tested. In terms of the 2 × 2 table below, it is (B + C)/N. Rather than considering all who are tested, we might narrow our consideration to just those who have the characteristic of interest and ask how well does the test do in identifying them. That is, we divide the number of valid positives by the base rate (expressed as the number of cases in the sample and not as the proportion in the population). This index is termed the test's **sensitivity**. In terms of the 2 × 2 table, it is B/(B + A). Alternatively, we can ask how good is the test in ruling out those who do not have the characteristic. In the example given in the text where a test was being used to screen for a malignancy (see Chapter 5), the point is made that the false negatives are of great importance. If we express the number of valid negatives as a ratio of the number of those who do not have the characteristic (as a proportion of 1 − BR), we have the index we want. This is termed the test's **specificity**. In terms of the 2 × 2, it is C/(D + C). A further index is the positive predictive value of the test. That is the proportion of valid positives among those identified by the test as having the characteristic (the selection ratio). In terms of the 2 × 2 table it is B/(B + D).

sensitivity
the proportion of those who have the behaviour of interest who are so predicted by the test or assessment device

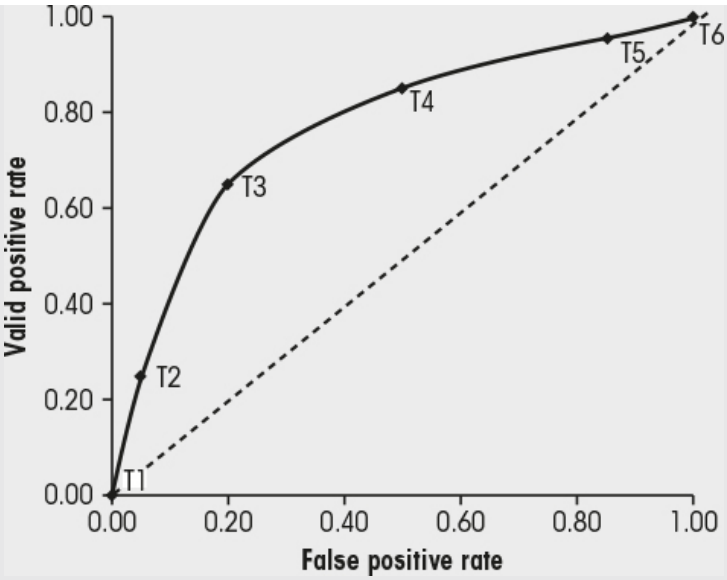
specificity
the proportion of those who do not show the behaviour of interest who are so predicted by the test or assessment device

A	B	BR
C	D	1 – BR
1 – SR	SR	N

The cutting score is often set on the basis of previous research, but it can be varied; when it is varied, the values of the various indices just described vary as well. For example, the sensitivity of the test varies with the cutting score adopted. If total maximum score is the cutting score, all valid positives will be identified—but then the test serves no purpose. To find the sensitivity of the test independent of cutting score the signal detection paradigm is employed, and a **receiver operating characteristic (ROC) curve** is formed by plotting valid positives against false positives at each of a number of cutting scores. The figure below is based on the data provided in Hsiao, Bartko and Potter (1989). The **area under the curve** (AUC) of the ROC provides an index of sensitivity independent of the cutting score, which can be calculated using a number of programs including SPSS (Stephan et al., 2003). The theoretical maximum AUC is 1.0, but no test is perfect and values less than 1 are to be expected. The method provides a basis for comparing the sensitivity of different tests under a given criterion condition.

receiver operating characteristic (ROC) curve
the curve of sensitivity against 1 minus specificity

area under the curve
the area under the receiver operating characteristic curve that is an index of sensitivity of a test or other assessment device that is independent of the particular cutting score on a test used to allocate test takers to the category showing the behaviour of interest



Exploratory factor analysis

Exploratory factor analysis has as its aim the reduction of a matrix of correlations to a set of components or factors fewer in number than the original correlations, in a way that allows recovery of the original correlation matrix from the factors. The correlations may be between scores for items in a test or among scores on different tests or among tests and non-test measures.

exploratory factor analysis
the use of factor analysis inductively to identify the factor structure of a set of variables

A **factor** is a linear combination of the elements of a data matrix (Nunnally, 1967). It is a variable on which members of the sample will have different values, just as they have different values (scores) on each of the tests or items in the analysis.

factor
a linear combination of test scores that attempts to summarise the intercorrelation of scores on tests or test items; it is often given meaning in terms of theory or hypothesis about psychological processes that underlie the intercorrelation such as latent traits

Suppose five tests have been administered to six people, with the results shown in the following table. (The small sample is for purposes of illustration and a much larger sample, 100 to 200, would normally be used in practice.)

Persons	1	2	3	4	5
1	3	5	7	8	2
2	4	6	6	8	3
3	2	3	4	5	1

Persons	1	2	3	4	5
4	5	4	6	6	3
5	5	3	5	5	4
6	6	4	6	6	5

The correlation matrix for this data set is shown.

Tests	1	2	3	4	5
1		0.019	0.307	–0.123	0.961
2			0.718	0.960	0.000
3				0.803	0.274
4					–0.104
5					

Inspection of the matrix indicates that some of the tests relate highly (e.g. 1 and 5, 2 and 4) and that some show little if any relation (e.g. 1 and 2, 2 and 5).

A factor (F_1) can be computed as a simple linear combination of the elements of the matrix.

$$F_1 = (1 \times T1\ score) + (1 \times T2\ score) + (1 \times T3\ score) + (1 \times T4\ score) + (1 \times T5\ score)$$

which simply requires that we sum the scores across tests. The weights can vary (Thurstone required that they be +1 or −1) and they can be fractions. The choice of weights is somewhat arbitrary, although some choices will lead to better solutions than others (i.e. the original correlations will be better recovered or reproduced from the results).

If the simplest equation for a factor is applied to the data matrix, for Person 1 we have:

$$F_1 = 3 + 5 + 7 + 8 + 2 = 25$$

and the **factor scores** for all persons are as follows.

factor score
the score that a person has on a factor and that is often interpreted to reflect their standing on a latent trait

Tests						Factor score
Persons	1	2	3	4	5	
1	3	5	7	8	2	25
2	4	6	6	8	3	27
3	2	3	4	5	1	15
4	5	4	6	6	3	24
5	5	3	5	5	4	22
6	6	4	6	6	5	27

Because the factor is a variable, the factor scores can be correlated with scores on each test in turn to yield correlation coefficients that are termed **factor loadings**. If the six scores under Test 1 are correlated with the six scores under 'Factor score' in the above table, a value of 0.66 is obtained. The factor loadings on all five tests are shown in the next table.

factor loading
the correlation of scores on a test or test item and a factor score, and that can be used in identifying the nature of the factor

Tests						Factor loading
Tests	1	2	3	4	5	
1		0.019	0.307	–0.123	0.961	0.66
2			0.718	0.960	0.000	0.71
3				0.803	0.274	0.85
4					–0.104	0.66
5						0.66

The loading can be interpreted in the same way as a correlation coefficient, with the square of the coefficient indicating the amount of variance that is common to the variables that are correlated. That is, the factor loading of 0.66 for Factor F_1 on Test 1 means that 43 per cent of the variance in scores on Test 1 can be accounted for by Factor F_1.

The process of **factor extraction** can now be repeated, with a second linear combination being proposed. Inspection of the original correlation matrix suggested that there was more than a single factor involved and hence it is reasonable to proceed to a second factor. The process could continue until k – 1 factors (where k is the number of variables being analysed) have been extracted, but the purpose of applying the technique would be defeated with so many factors. One of the issues in using the technique is the decision about when to stop factoring.

factor extraction
the process of calculating the factor or factors that can summarise a matrix of correlations among scores on tests or test items

If a second factor is to be extracted and this is done on the original matrix, then Factor F_1 and the new factor (Factor F_2) will be correlated. To ensure the second factor is independent, the variance in the tests that it accounts for is first removed. This can be done by regression; that is, using the factor to predict scores on each of the tests and then subtracting from each person's score the part that is predictable from the factor. This leaves a set of scores that do not share variance with the first factor, and this set of scores is used in the next phase of the analysis.

To proceed to a second factor, SPSS is used and the method of Principal Components selected. This provides a more precise solution than attempting to fit a second factor by eye. When this is done, the following table results. It presents the factor loadings on the five tests for the first and second factors. Note the loadings for the first factor differ from those shown above because the simple linear combination used there does not provide an optimal solution. The computer program allows for a number of options to be tried so that the best solution is obtained.

Component		
Test	F_1	F_2
1	0.201	0.971
2	0.929	0.226
3	0.919	0.133
4	0.942	0.325
5	0.207	0.962

Inspection of the loading matrix indicates that F_1 loads highly on Tests 2, 3 and 4 with low or weak loadings on Tests 1 and 5, whereas the reverse is the case for F_2. A weak loading is often described as one that involves less than 10 per cent of the variance (i.e. a loading of 0.316 or smaller). A cut off of 0.3 to 0.4 is used in practice. A strong loading, on the other hand, is one where there is at least more shared than non-shared variance between the variable and factor (0.71 or greater). A cut off of 0.7 is often used in practice.

In the example analysis, the **variance accounted for** by F_1 and F_2 can be summed to indicate the total variance in each test accounted for by the two factors. Alternatively, the variance accounted for in each test can be summed and averaged (over the number of tests) to indicate the variance accounted for by each factor.

variance accounted for
the percentage of variance of the total or reliable variance of a set of variables that each factor has in common with the variables as a set; it is calculated from the matrix of factor loadings

Component				
	Test	F_1	F_2	h^2
	1	0.201	0.971	0.9832
	2	0.929	0.226	0.9141
	3	0.919	0.133	0.8623
	4	0.942	0.325	0.9930
	5	0.207	0.962	0.9683
Sum of squared loadings		2.6782	2.0427	
	V	0.54	0.41	

In the table above, the first row shows a h^2 value (or **communality**) of 0.9832. That is, 98 per cent of the variance in Test 1 is accounted for by F_1 and F_2 in combination. This is calculated by squaring the loading of F_1 on Test 1 (0.201^2 = 0.0404) and adding it to the square of the loading of F_2 on Test 1 (0.971^2 = 0.9428). The remaining entries in the column are calculated in the same way. The first entry in the row titled 'Sum of squared loadings' is the sum of each of the loadings under F_1 squared (i.e. 0.0404 + 0.8630 + … + 0.0428). Similarly, the squared loadings under F_2 are summed. The row labelled V (variance accounted for) is the average of the sum of squared loadings from the rows above. Component F_1 accounts for 54 per cent of the total variance in the tests and F_2 accounts for 41 per cent. Together they account for 96 per cent (a figure seldom if ever attained with real data).

communality
the amount of variance in a given variable that is shared with the factors constituting a particular factor matrix

How satisfactory is this solution? It accounts for a large percentage of the variance (95 per cent of the variables as a set), but the objective is to provide a solution from which the original correlation matrix can be reproduced. In the case of a two-factor solution, the original correlations should be given by the product of the loadings of each of the components on the tests that are correlated. For example:

$$rT1T2 = (rT1F_1 \times rT2F_1) + (rT1F_2 \times rT2F_2)$$

which reads that the correlation between T1 and T2 is equal to the product of the loadings of Test 1 and Test 2 on Factor F_1 plus the product of the loadings of Test 1 and Test 2 on F_2. From

the matrix of factor loadings:

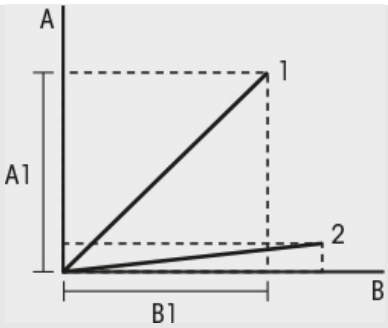
$$\text{rT1T2 should equal } (0.201 \times 0.929) + (0.971 \times -0.226) \text{ or } 0.033$$

In fact, the correlation from the correlation matrix is 0.019, which means an error of about 0.1. Correlations for other pairs of tests likewise show small discrepancies and so the objective of perfect recovery of the correlation matrix has not been achieved. The residuals (the difference between actual and predicted) are, however, reasonably small. The ten original correlations can be described in terms of loadings on just two factors, with a tolerable margin of error.

There is often a further stage in the analysis in which the factors are 'rotated' to increase their interpretability. The factors are labelled in terms of the variables on which they have the highest loadings. If Tests 2, 3 and 4 were tests that required a reasonable level of schooling, Factor F_1 could be labelled a verbal/educational factor. If Tests 1 and 5 were tests of spatial relations then that could be the label for Factor F_2. Labelling is arbitrary and, as with the choice of weights, different choices can be made. Labelling can sometimes be assisted if the factor loadings are changed to increase or reduce their influence on some of the variables in the analysis, a process termed '**rotation**'. The total variance explained is not altered in the process, just the way it is apportioned.

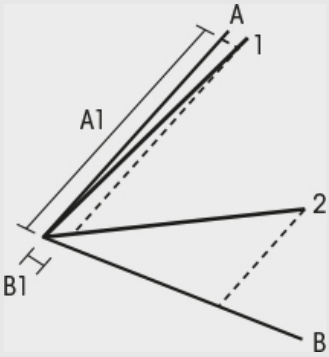
rotation
a method of varying the loadings of a factor on each of a set of variables, originally performed geometrically, with a view to producing a more psychologically meaningful factor structure

Rotation derives from the fact that a geometrical solution of the factor problem can be provided. In the early days of factor analysis the calculations and the geometry were done by hand. The starting point is the idea that a correlation can be conceptualised as the angle between two lines. An angle of 90 degrees means that the tests are at right angles to each other or independent (r = 0). As the angle decreases, there is a closer relationship between the two lines until they overlap (r = 1). The relationship is given by a transformation (the cosine) of the angle. A set of reference axes for the angle can be drawn and the angle expressed in terms of distances along the reference axes (assuming some facility with coordinate geometry). The reference axes are the factors and the distances are the loadings. In the following figure, the correlation between Tests 1 and 2 is represented by the angle between the lines 1 and 2. Two reference axes for the angle, A and B, are shown. Perpendiculars from line 1 to axis A and axis B and from line 2 to the axes are shown. The distances along the reference axes are the loadings of the factors on the reference axes, so that the correlation between the lines can be described in terms of the distances on the reference axes. (The loadings of Factor A on the two tests are shown, but the loadings for Factor B have not been drawn in to ensure the figure is not too confusing.)



For two variables the use of two reference axes achieves nothing, but for a larger number of variables there is an economy if they can be expressed in terms of a smaller number of reference axes. In the example above, ten correlations were reduced to two factors (axes).

Where the axes are placed is arbitrary in the sense that the angle can be expressed in terms of distances with a different placement of the axes. The following figure attempts to illustrate. The angle is the same but the position of the reference axes has changed and so have the distances. A now has a larger loading on Test 1 and B has a smaller loading. Maximising the loadings of one factor on a set of tests and minimising its loadings on a different set of tests makes for a clearer result. Thurstone (1954) advocated this outcome, in what he called simple structure.



The factors have now been rotated. Note that the right angle between the axes has been preserved in the shift of the reference axes but that is not necessary and some (e.g. Cattell & Muerle, 1960) have argued for oblique rather than orthogonal axes—that is, for angles other than 90 degrees between the factors—on the grounds that is unlikely the **latent variables** researchers seek to identify are really orthogonal to each other. The matrix of factor loadings that results after rotation is referred to as the **factor structure**.

latent variable
a variable that is not directly observable but is hypothesised to exist on the basis of psychological theory; when it is a variable giving rise to individual differences it is referred to as a latent trait

factor structure
the matrix of factor loadings for a set of variables, usually after factor rotation

Factor analysis as outlined to date involves a good measure of judgment: what variables to include, how large a sample is necessary, which method to employ, how many factors to extract, how the factors are to be labelled, and how any rotation of the factors is to proceed—see Henson and Roberts (2006) for a brief review. Bad decisions at any of these points can leave the results of

the analysis open to serious criticism. This is true, however, with much professional work, and the limitations of factor analysis in this regard should not be overstated.

Item analysis

The CTT model yields somewhat different item statistics to that for IRT models.

CTT

Item difficulty (p) is probability of endorsement of the correct answer in the sample:

p = Number of respondents correctly answering the item/Number of respondents

It is in fact an index of 'easiness' rather than difficulty, because the higher the p the less difficult the item. In general, items of intermediate difficulty, in the region of 0.5 (say, 0.5 ± 0.2), are generally sought. Sometimes a few easy items (large p) are included at the beginning of a test for motivational purposes. These would not be included in determining total score on the test.

If discrimination is desired at some point (to identify, for example, the top 25 per cent in ability level for special treatment) then items would be selected in terms of item difficulty that approximates that cutting point (e.g. p = 0.25). If guessing is likely then the item difficulty can be manipulated to allow for this. For example, on a true/false test with a high likelihood of guessing influencing the answer, an item difficulty level at approximately 0.85 is recommended (Reynolds & Livingston, 2014). However, it usually makes more sense to address the issue of guessing at the time of item writing rather than after the fact.

Item discrimination is now usually determined by the correlation of an item with total score on the test. For a test with dichotomously scored items this is usually computed with the point biserial correlation (rpb), which is the ordinary product moment correlation applied to variables, one of which is continuous and the other dichotomous. Larger coefficients indicate greater discrimination. Negative correlations are possible but unlikely and usually arise when the item has been scored in the wrong direction (i.e. to indicate less of the trait in question).

Point biserial correlations can only have a maximum value of 1 when p = 0.5. It is often the case that the coefficients are not interpreted in terms of their absolute values (how close they are to 1.0) but in a relative sense (which items have better discrimination indexes), given of course that they are not zero or very close to it (e.g. < 0.1).

The biserial correlation (r_b)is sometimes used in test construction. It is not part of the product moment family of correlations in the way the point biserial is, but can be used to estimate the point biserial correlation. It is used, as with the point biserial correlation, when one of the variables is dichotomous and the other is continuous, but when it is used there is the underlying assumption that the dichotomy is artificial. An artificial dichotomy is one in which the dichotomous variable is in fact continuous and an arbitrary cut point has been chosen. For example, those passing a class test and those failing do not constitute a true dichotomy as membership could shift if a different cut point were used. For many psychological variables dichotomies are artificial and in that sense the biserial correlation is more realistic. It can however take values greater than 1, which can be confusing.

$$r_{\mathrm{b}} = r_{pb}\left(\sqrt{\frac{pq}{Y_i}}\right)$$

In calculating item-total correlations it is usual to 'correct' total score by removing the contribution of the item to the total before calculating the correlation coefficient for the item. Thus the total used is the total minus the item score. If the item is part of the total there is necessarily some correlation between item and total. Computer software makes this easy to do.

One other item statistic that it is possible to compute easily with available computer software is the alpha-if-item-deleted index, the alpha coefficient if the item is removed from the test. As alpha estimates the average inter-item correlation, if removal of an item increases alpha then the item is detracting from the internal consistency of the test and is a candidate for deletion from the test. However, alpha depends on the length of the test and use of this item statistic alone could have an adverse effect on test reliability.

The statistics are used in combination in deciding the best items to retain. First items with p values outside the chosen range of item difficulty are eliminated and then items with higher discrimination indexes chosen. The nature of the test needs to be borne in mind when using these statistics. For example, speeded tests that need to be completed in a set time period may lead to low p values and low item-total correlations for items late in the test because only a small proportion of the sample have been able to complete them. For power tests, where there is no time limit, the item statistics may be more interpretable throughout the range.

There are two other statistics that are sometimes used. One is the difference between the upper and lower scoring groups on the test. If the sample is divided into the top and bottom 27 per cent on the basis of total score on the test, it becomes possible to compare each of the items for these groups. This is a method favoured before the advent of high speed computers and is in effect a way of correlating each item with total score, although in a way that produces a biased estimate of the correlation (it omits the middle scoring 46 per cent of the sample). It does have some use, however, with multiple choice tests because it allows examination of how well the 'distracters'—the incorrect options provided in addition to the correct answer—are working. Distracters that have high endorsement are working too well and distracters that have high endorsement for one group (e.g. the low scoring 27 per cent) are particularly problematic.

The other statistic is sometimes referred to as item validity and refers to the correlation between an item and score on an external criterion being used to validate the test. By selecting items with high item validity, the correlation between the score on the final version of the test and the external criterion should be maximised. This was a common strategy when validity was thought of only in terms of criterion validity and external keying was the method of choice for test construction, neither of which is true anymore. If a test is being developed for one highly specific use, then attention to item validity may make sense, but to the extent that one seeks to maximise the correlation with one criterion measure, one may be reducing the correlation of the test with another criterion that is only moderately correlated with it, and thereby reducing the value of the test.

IRT

The statistics evaluated in an IRT item analysis will depend on the model being used. In a 3PL model item difficulty (b), item discrimination (a) and guessing (c) are evaluated. For the Rasch model only b is available. Programs for IRT analyses provide a number of item statistics to do with how well an item or the response of a member of the sample fits the model being applied, as well as the degree of fit in each case. Where responses of a number of cases do not fit the model these would be deleted and the analysis recalculated (although some would argue against doing this). The fit of items to the model are then evaluated in the light of overall fit and the fit statistics. Items for which the fit was poor would be candidates for removal. In the case of a Rasch analysis, the

spread of item difficulty would then be assessed in the light of the intended use of the test. Other statistics provided would be the item information and total test information functions.

Hulin, Drasgow and Parsons (1983) give examples of comparisons of traditional and IRT-based item analysis, and an extract from one of these is reproduced in the table below. The **item analysis indexes** included under the traditional approach are item discrimination (r_b and r_{pb}) and the commonly used measure of item difficulty (probability of getting the item right). Under the IRT indices are discrimination (*a*), difficulty (*b*) and the guessing parameter (*c*), which is not calculated in the traditional method but which is sometimes accounted for by a correction to total score.

item analysis indexes
the statistics arising in the process of item analysis used to evaluate each item in terms of its likely contribution to the psychological test being developed; in classical test theory they include item difficulty and item discrimination, and in item response theory they include the parameters of the ICC

	Traditional item statistics			IRT item statistics		
Item	r_b	r_{pb}	p	*a*	*b*	*c*
35	0.761	0.578	0.687	1.539	–0.288	0.160
2	0.753	0.356	0.951	0.824	–2.652	0.160
4	0.750	0.359	0.943	0.744	–2.795	0.160
36	0.725	0.476	0.874	0.911	–1.151	0.160
20	0.706	0.516	0.760	1.35	–0.556	0.160

Adapted from Hulin, Drasgow and Parsons (1983, p. 87).

Inspection of the above table indicates that Item 2 has the second-highest rb index (the items were ordered in the complete table in terms of magnitude on this index), but three items have higher *a* values. In fact, there are twenty-seven items (not shown in this extracted table) for which the corresponding *a* value derived on the basis of the IRT analysis was higher than that for Item 2. This means that Item 2 is likely to be retained in a CTT item analysis because of its high rb, but may not be retained in an IRT analysis because there is a large number of items with higher *a* values. The difficulty indices for the five items line up fairly well across the traditional and IRT methods to judge from the rank order for these five items. Note that with the traditional method, difficulty is really easiness (the proportion passing the item), whereas the IRT estimate is of difficulty, with smaller estimates (more negative numbers) indicating a less difficult item. Thus, model selection and model fit can have an impact on the decisions made in the course of item analysis.

One other item statistic for IRT analysis is **item information**. The information in an item is the extent to which knowledge of the score on the test reduces our uncertainty about the level of the trait (theta). It is given by the product of the probability of getting the item correct and the probability of getting the item wrong, given theta. Where those probabilities are 0.5 the item conveys maximum information, and as they depart from 0.5 information decreases.

item information
the term used in item response theory to describe the value of an item in identifying a respondent's position on the underlying trait of interest

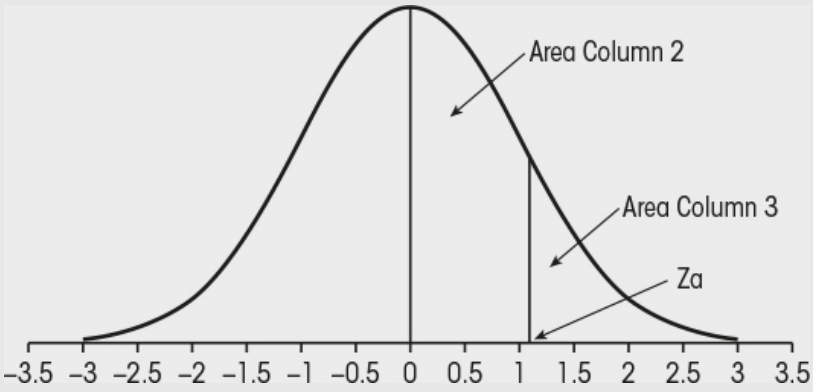
The standard error of estimate (SEE) is the parallel statistic in the Rasch model to the SEM, but it is not based on the assumption that errors of measurement are equal at all levels of the underlying trait (e.g. that estimating high scores has SEM as estimating scores in the middle of the range.

For such a test, SEE is calculated for a given level of theta as an inverse function of the information value of the items in the test.

$$SEE = \frac{1}{\sqrt{I_\theta}}$$

where I_θ is the sum of the item information values for a particular level on the latent trait.

Table A1 Table of the standard normal curve



z_a	Area from 0 to z_a	Area beyond z_a
0.0000	0.0000	0.5000
0.0100	0.0040	0.4960
0.0200	0.0080	0.4920
0.0300	0.0120	0.4880
0.0400	0.0160	0.4840
0.0500	0.0199	0.4801
0.0600	0.0239	0.4761
0.0700	0.0279	0.4721
0.0800	0.0319	0.4681

z_a	Area from 0 to z_a	Area beyond z_a
0.0900	0.0359	0.4641
0.1000	0.0398	0.4602
0.1100	0.0438	0.4562
0.1200	0.0478	0.4522
0.1300	0.0517	0.4483
0.1400	0.0557	0.4443
0.1500	0.0596	0.4404
0.1600	0.0636	0.4364
0.1700	0.0675	0.4325
0.1800	0.0714	0.4286
0.1900	0.0753	0.4247
0.2000	0.0793	0.4207
0.2100	0.0832	0.4168
0.2200	0.0871	0.4129
0.2300	0.0910	0.4090
0.2400	0.0948	0.4052
0.2500	0.0987	0.4013
0.2600	0.1026	0.3974
0.2700	0.1064	0.3936
0.2800	0.1103	0.3897
0.2900	0.1141	0.3859
0.3000	0.1179	0.3821
0.3100	0.1217	0.3783
0.3200	0.1255	0.3745
0.3300	0.1293	0.3707
0.3400	0.1331	0.3669
0.3500	0.1368	0.3632
0.3600	0.1406	0.3594
0.3700	0.1443	0.3557
0.3800	0.1480	0.3520

z_a	Area from 0 to z_a	Area beyond z_a
0.3900	0.1517	0.3483
0.4000	0.1554	0.3446
0.4100	0.1591	0.3409
0.4200	0.1628	0.3372
0.4300	0.1664	0.3336
0.4400	0.1700	0.3300
0.4500	0.1736	0.3264
0.4600	0.1772	0.3228
0.4700	0.1808	0.3192
0.4800	0.1844	0.3156
0.4900	0.1879	0.3121
0.5000	0.1915	0.3085
0.5100	0.1950	0.3050
0.5200	0.1985	0.3015
0.5300	0.2019	0.2981
0.5400	0.2054	0.2946
0.5500	0.2088	0.2912
0.5600	0.2123	0.2877
0.5700	0.2157	0.2843
0.5800	0.2190	0.2810
0.5900	0.2224	0.2776
0.6000	0.2257	0.2743
0.6100	0.2291	0.2709
0.6200	0.2324	0.2676
0.6300	0.2357	0.2643
0.6400	0.2389	0.2611
0.6500	0.2422	0.2578
0.6600	0.2454	0.2546
0.6700	0.2486	0.2514
0.6800	0.2517	0.2483

z_a	Area from 0 to z_a	Area beyond z_a
0.6900	0.2549	0.2451
0.7000	0.2580	0.2420
0.7100	0.2611	0.2389
0.7200	0.2642	0.2358
0.7300	0.2673	0.2327
0.7400	0.2704	0.2296
0.7500	0.2734	0.2266
0.7600	0.2764	0.2236
0.7700	0.2794	0.2206
0.7800	0.2823	0.2177
0.7900	0.2852	0.2148
0.8000	0.2881	0.2119
0.8100	0.2910	0.2090
0.8200	0.2939	0.2061
0.8300	0.2967	0.2033
0.8400	0.2995	0.2005
0.8500	0.3023	0.1977
0.8600	0.3051	0.1949
0.8700	0.3078	0.1922
0.8800	0.3106	0.1894
0.8900	0.3133	0.1867
0.9000	0.3159	0.1841
0.9100	0.3186	0.1814
0.9200	0.3212	0.1788
0.9300	0.3238	0.1762
0.9400	0.3264	0.1736
0.9500	0.3289	0.1711
0.9600	0.3315	0.1685
0.9700	0.3340	0.1660
0.9800	0.3365	0.1635

z_a	Area from 0 to z_a	Area beyond z_a
0.9900	0.3389	0.1611
1.0000	0.3413	0.1587
1.0100	0.3438	0.1562
1.0200	0.3461	0.1539
1.0300	0.3485	0.1515
1.0400	0.3508	0.1492
1.0500	0.3531	0.1469
1.0600	0.3554	0.1446
1.0700	0.3577	0.1423
1.0800	0.3599	0.1401
1.0900	0.3621	0.1379
1.1000	0.3643	0.1357
1.1100	0.3665	0.1335
1.1200	0.3686	0.1314
1.1300	0.3708	0.1292
1.1400	0.3729	0.1271
1.1500	0.3749	0.1251
1.1600	0.3770	0.1230
1.1700	0.3790	0.1210
1.1800	0.3810	0.1190
1.1900	0.3830	0.1170
1.2000	0.3849	0.1151
1.2100	0.3869	0.1131
1.2200	0.3888	0.1112
1.2300	0.3907	0.1093
1.2400	0.3925	0.1075
1.2500	0.3944	0.1056
1.2600	0.3962	0.1038
1.2700	0.3980	0.1020
1.2800	0.3997	0.1003

z_a	Area from 0 to z_a	Area beyond z_a
1.2900	0.4015	0.0985
1.3000	0.4032	0.0968
1.3100	0.4049	0.0951
1.3200	0.4066	0.0934
1.3300	0.4082	0.0918
1.3400	0.4099	0.0901
1.3500	0.4115	0.0885
1.3600	0.4131	0.0869
1.3700	0.4147	0.0853
1.3800	0.4162	0.0838
1.3900	0.4177	0.0823
1.4000	0.4192	0.0808
1.4100	0.4207	0.0793
1.4200	0.4222	0.0778
1.4300	0.4236	0.0764
1.4400	0.4251	0.0749
1.4500	0.4265	0.0735
1.4600	0.4279	0.0721
1.4700	0.4292	0.0708
1.4800	0.4306	0.0694
1.4900	0.4319	0.0681
1.5000	0.4332	0.0668
1.5100	0.4345	0.0655
1.5200	0.4357	0.0643
1.5300	0.4370	0.0630
1.5400	0.4382	0.0618
1.5500	0.4394	0.0606
1.5600	0.4406	0.0594
1.5700	0.4418	0.0582
1.5800	0.4429	0.0571

z_a	Area from 0 to z_a	Area beyond z_a
1.5900	0.4441	0.0559
1.6000	0.4452	0.0548
1.6100	0.4463	0.0537
1.6200	0.4474	0.0526
1.6300	0.4484	0.0516
1.6400	0.4495	0.0505
1.6500	0.4505	0.0495
1.6600	0.4515	0.0485
1.6700	0.4525	0.0475
1.6800	0.4535	0.0465
1.6900	0.4545	0.0455
1.7000	0.4554	0.0446
1.7100	0.4564	0.0436
1.7200	0.4573	0.0427
1.7300	0.4582	0.0418
1.7400	0.4591	0.0409
1.7500	0.4599	0.0401
1.7600	0.4608	0.0392
1.7700	0.4616	0.0384
1.7800	0.4625	0.0375
1.7900	0.4633	0.0367
1.8000	0.4641	0.0359
1.8100	0.4649	0.0351
1.8200	0.4656	0.0344
1.8300	0.4664	0.0336
1.8400	0.4671	0.0329
1.8500	0.4678	0.0322
1.8600	0.4686	0.0314
1.8700	0.4693	0.0307
1.8800	0.4699	0.0301

z_a	Area from 0 to z_a	Area beyond z_a
1.8900	0.4706	0.0294
1.9000	0.4713	0.0287
1.9100	0.4719	0.0281
1.9200	0.4726	0.0274
1.9300	0.4732	0.0268
1.9400	0.4738	0.0262
1.9500	0.4744	0.0256
1.9600	0.4750	0.0250
1.9700	0.4756	0.0244
1.9800	0.4761	0.0239
1.9900	0.4767	0.0233
2.0000	0.4772	0.0228
2.0100	0.4778	0.0222
2.0200	0.4783	0.0217
2.0300	0.4788	0.0212
2.0400	0.4793	0.0207
2.0500	0.4798	0.0202
2.0600	0.4803	0.0197
2.0700	0.4808	0.0192
2.0800	0.4812	0.0188
2.0900	0.4817	0.0183
2.1000	0.4821	0.0179
2.1100	0.4826	0.0174
2.1200	0.4830	0.0170
2.1300	0.4834	0.0166
2.1400	0.4838	0.0162
2.1500	0.4842	0.0158
2.1600	0.4846	0.0154
2.1700	0.4850	0.0150
2.1800	0.4854	0.0146

z_a	Area from 0 to z_a	Area beyond z_a
2.1900	0.4857	0.0143
2.2000	0.4861	0.0139
2.2100	0.4864	0.0136
2.2200	0.4868	0.0132
2.2300	0.4871	0.0129
2.2400	0.4875	0.0125
2.2500	0.4878	0.0122
2.2600	0.4881	0.0119
2.2700	0.4884	0.0116
2.2800	0.4887	0.0113
2.2900	0.4890	0.0110
2.3000	0.4893	0.0107
2.3100	0.4896	0.0104
2.3200	0.4898	0.0102
2.3300	0.4901	0.0099
2.3400	0.4904	0.0096
2.3500	0.4906	0.0094
2.3600	0.4909	0.0091
2.3700	0.4911	0.0089
2.3800	0.4913	0.0087
2.3900	0.4916	0.0084
2.4000	0.4918	0.0082
2.4100	0.4920	0.0080
2.4200	0.4922	0.0078
2.4300	0.4925	0.0075
2.4400	0.4927	0.0073
2.4500	0.4929	0.0071
2.4600	0.4931	0.0069
2.4700	0.4932	0.0068
2.4800	0.4934	0.0066

z_a	Area from 0 to z_a	Area beyond z_a
2.4900	0.4936	0.0064
2.5000	0.4938	0.0062
2.5100	0.4940	0.0060
2.5200	0.4941	0.0059
2.5300	0.4943	0.0057
2.5400	0.4945	0.0055
2.5500	0.4946	0.0054
2.5600	0.4948	0.0052
2.5700	0.4949	0.0051
2.5800	0.4951	0.0049
2.5900	0.4952	0.0048
2.6000	0.4953	0.0047
2.6100	0.4955	0.0045
2.6200	0.4956	0.0044
2.6300	0.4957	0.0043
2.6400	0.4959	0.0041
2.6500	0.4960	0.0040
2.6600	0.4961	0.0039
2.6700	0.4962	0.0038
2.6800	0.4963	0.0037
2.6900	0.4964	0.0036
2.7000	0.4965	0.0035
2.7100	0.4966	0.0034
2.7200	0.4967	0.0033
2.7300	0.4968	0.0032
2.7400	0.4969	0.0031
2.7500	0.4970	0.0030
2.7600	0.4971	0.0029
2.7700	0.4972	0.0028
2.7800	0.4973	0.0027

z_a	Area from 0 to z_a	Area beyond z_a
2.7900	0.4974	0.0026
2.8000	0.4974	0.0026
2.8100	0.4975	0.0025
2.8200	0.4976	0.0024
2.8300	0.4977	0.0023
2.8400	0.4977	0.0023
2.8500	0.4978	0.0022
2.8600	0.4979	0.0021
2.8700	0.4979	0.0021
2.8800	0.4980	0.0020
2.8900	0.4981	0.0019
2.9000	0.4981	0.0019
2.9100	0.4982	0.0018
2.9200	0.4982	0.0018
2.9300	0.4983	0.0017
2.9400	0.4984	0.0016
2.9500	0.4984	0.0016
2.9600	0.4985	0.0015
2.9700	0.4985	0.0015
2.9800	0.4986	0.0014
2.9900	0.4986	0.0014
3.0000	0.4987	0.0013
3.0100	0.4987	0.0013
3.0200	0.4987	0.0013
3.0300	0.4988	0.0012
3.0400	0.4988	0.0012
3.0500	0.4989	0.0011
3.0600	0.4989	0.0011
3.0700	0.4989	0.0011
3.0800	0.4990	0.0010

z_a	Area from 0 to z_a	Area beyond z_a
3.0900	0.4990	0.0010
3.1000	0.4990	0.0010
3.1100	0.4991	0.0009
3.1200	0.4991	0.0009
3.1300	0.4991	0.0009
3.1400	0.4992	0.0008
3.1500	0.4992	0.0008
3.1600	0.4992	0.0008
3.1700	0.4992	0.0008
3.1800	0.4993	0.0007
3.1900	0.4993	0.0007
3.2000	0.4993	0.0007
3.2100	0.4993	0.0007
3.2200	0.4994	0.0006
3.2300	0.4994	0.0006
3.2400	0.4994	0.0006
3.2500	0.4994	0.0006
3.3000	0.4995	0.0005
3.3500	0.4996	0.0004
3.4000	0.4997	0.0003
3.4500	0.4997	0.0003
3.5000	0.4998	0.0002
3.6000	0.4998	0.0002
3.7000	0.4999	0.0001
3.8000	0.4999	0.0001
3.9000	0.5000	0.0001
4.0000	0.5000	0.0000

[1] Older texts on psychometrics (Nunnally, 1976; Allen & Yen, 1979) use the formula with N in the denominator, whereas more recent texts (Raykov & Marcoulides, 2011; Revelle, in

preparation) use N – 1. The argument for N – 1 is that one always wants to generalise to the population and hence N – 1 is to be preferred because it provides a less biased estimator of the population standard deviation. When N is large (say >100), which is usually the case in real-world application of psychometrics, the difference in results with the two formulas is negligible. SPSS uses N – 1 in calculations.

[2] The term 'trace line' is generally used in the older classical test theory whereas item characteristic curve is generally used in Item response curve theory. The terms have essentially the same meaning.

[3] The term 'attribute' is used here interchangeably with terms such as construct and trait or latent trait (the term used in IRT). Osterlind (2005) defines a construct as a theoretical conception of a psychological process (p. 63). Thorndike (1982) defines a latent trait as a 'hypothesised but unobservable characteristic that accounts for a particular set of consistencies in behaviour and differences among persons' (p. 5).

[4] Probability (P) of event i is the frequency of event i divided by the total number of events, Pi = fi/N. Odds of the event i is the frequency of event i divided by the number of events that are not i: Odds = fi/(N – fi). Odds = P/1 –P P = odds/(1 + odds)

Glossary

achievement test
a test to assess past learning

aptitude test
a test to assess future learning potential

area under the curve
the area under the receiver operating characteristic curve that is an index of sensitivity of a test or other assessment device that is independent of the particular cutting score on a test used to allocate test takers to the category showing the behaviour of interest

artificial intelligence
(AI) a technology-based intelligence that attempts to mimic human intelligence; recent expressions are sophisticated chess and GO playing programs, and self-driving vehicles

assessment centre
a comprehensive testing procedure applied to groups that includes a diverse range of testing tools and techniques

attention
the ability to focus on or select one stimulus or process while ignoring another; it has at least three components (i.e. attention span, focused attention and selective attention)

attribute
(or characteristic) the consistent set of behaviours, thoughts or feelings that is the target of a psychological test

base rate
the proportion of individuals in the population who show the behaviour of interest in a given psychological testing or assessment situation

behavioural observation scale
(BOS) questions used in a rating scale that are based on actual behaviours;

they are rated for their frequency of occurrence (e.g. from '1 = almost never displayed' to '5 = almost always displayed')

behaviourally anchored rating scale
(BARS) a rating scale that includes actual behaviours to indicate the response

biographical data
(biodata) measures of past activities, effort and interests that reflect motivation, personality, values and interest, which assume that past behaviours will be consistent with future behaviours

CHC theory of intelligence
the Cattell-Horn-Carroll model; a merging of the Cattell and Horn's *Gf-Gc* theory and Carroll's three stratum theory, which proposes three levels or strata of abilities: narrow, broad and general (or 'g')

classical test theory
the set of ideas, expressed mathematically and statistically, that grew out of attempts in the first half of the twentieth century to measure psychological variables; and that turns on the central idea of a score on a psychological test comprising both true and error score components

clinical interview
a technique for collecting information about a client; it may take many forms, for example, a psychoanalytic perspective includes detailed exploration of the personal and family history of the client, particularly with respect to psychosocial development, conflict, and defence, self and interpersonal processes

clinical neuropsychologist
a psychologist who specialises in understanding, assessing and treating individuals' cognitive and behavioural impairments resulting from brain injury

clinical neuropsychology
a sub-branch of neuropsychology that is applied in nature and concerned with the assessment and treatment of cognitive impairments resulting from brain injury

clinical psychologist
a psychologist who specialises in the diagnosis, assessment, treatment and prevention of psychological and mental health problems

communality
the amount of variance in a given variable that is shared with the factors constituting a particular factor matrix

competency to stand trial
an assessment of whether a defendant is able to stand trial because his/her mental state was affected at the time of the offence or at the time of the trial

computerised adaptive testing
(CAT) programs that rapidly identify a test taker's ability level from a small number of items by (a) administering an initial item, (b) administering a more difficult or easier item depending on whether the initial item was correct or incorrect, (c) again administering a more difficult or easier item depending on the response to the second item, and (d) so on

concurrent validity
a form of predictive validity in which the index of social behaviour is obtained close in time to score on the psychological test (or other assessment device)

construct
a specific idea or concept about a psychological process or underlying trait that is hypothesised on the basis of a psychological theory

construct validity
the meaning of a test score made possible by knowledge of the pattern of relationships it enters into with other variables and the theoretical interpretation of those relationships

constructed response test
(CRT) a test that requires the test taker to construct the answer in response to the question; no options are provided (as are in multiple choice tests)

content analysis
the process of analysing textual information, either written or oral by, for example, searching for themes, examining frequencies of key words or constructs, and identifying repeating relationships; the procedure can be carried out manually or with computer-based software

content validity
the meaning that can be attached to a score on a psychological test (or other assessment device) on the basis of inspection of the material that constitutes the test

contextual performance
discretionary social behaviours directed at successful performance of the work group or organisation; sometimes referred to as 'citizenship behaviours'

convergent and discriminant validity
the subjection of a multitrait–multimethod matrix to a set of criteria that specify which correlations should be large and which small in terms of a psychological theory of the constructs

counter-productive behaviours
behaviours that are largely under the control of the individual or reflect problematic employee characteristics, and which impede the progress and success of the organisation

criterion-referenced test
a psychological test that uses a predetermined empirical standard as an objective reference point for evaluating the performance of a test taker

criterion referencing
a way of giving meaning to a test score by specifying the standard that needs to be reached in relation to a limited set of behaviours

critical incident
an example of extreme levels of behaviour or performance (both poor and exemplary behaviours), which are usually key determinants of subsequent outcomes

Cronbach's alpha
an estimate of reliability that is based on the average intercorrelation of the

items in a test

crystallised intelligence (Gc)
the accumulated knowledge and skills resulting from educational and life experiences

culture fair test
a test devised to measure intelligence while relying as little as possible on culture-specific knowledge (e.g. language); tests are devised to be suitable across different peoples, with the goal to measure fluid rather than crystallised intelligence

custody evaluation
an evaluation conducted to determine in cases of divorce, abuse or neglect or guardianship which parent should have custody of a child

cutting point
(or cutting score) the test score or point on a scale, in the case of another assessment device, that is used to split those being tested or assessed into two groups predicted to show or not show some behaviour of interest

Depression Anxiety and Stress Scale (DASS)
a 42-item self-report scale that aims to measure the state of depression, anxiety and stress in adults over the previous week

deviation IQ
a method that allows an individual's score to be compared with same-age peers; the score is reported as distance from the mean in standard deviation units

Diagnostic and Statistical Manual of Mental Disorders (DSM)
a standard classification system of mental disorders published by the American Psychiatric Association for professionals to use to diagnose mental disorders

differential item functioning
the possibility that a psychological test item may behave differently for different groups of respondents

domain-sampling model
a way of thinking about the composition of a psychological test that sees

the test as a representative sample of the larger domain of possible items that could be included in the test

educational and developmental psychologist
a psychologist who specialises in assessing and treating children and adults with learning and developmental needs

emotional intelligence
(EQ) a controversial construct (considered by many to not be an 'intelligence') that refers to the person's capacity to monitor and manage their own emotions and to understand the emotions of others, and to use these insights to function better interpersonally

empirical approach
a way of constructing psychological tests that relies on collecting and evaluating data about how each of the items from a pool of items discriminate between groups of respondents who are thought to show or not show the attribute the test is to measure; also an approach to personality that relates the reports that people make about their characteristic behaviours to their social functioning and thereby provide tools for personality prediction

equivalent forms reliability
the estimate of reliability of a test obtained by comparing two forms of a test constructed to measure the same construct

ethics
a set of principles for guiding behaviour; in the case of psychological testing and assessment, for guiding professional behaviour

executive functions
higher-level functions considered to be mediated by the prefrontal lobes; responsible for goal-directed behaviours, these functions usually include components such as working memory, concept formation, problem solving and planning

expectancy table
a table that presents the probability of an outcome on a criterion of interest in terms of score on score range on a test

expert witness
someone who can or is required to provide factual information as well as an opinion, based on their background and training in a court of law

explicit theories of intelligence
theories of intelligence devised by psychologists and other scientists; the theories grow out of and are validated using scientific methods, although they can be informed by implicit theories

exploratory factor analysis
the use of factor analysis inductively to identify the factor structure of a set of variables

factor
a linear combination of test scores that attempts to summarise the intercorrelation of scores on tests or test items; it is often given meaning in terms of theory or hypothesis about psychological processes that underlie the intercorrelation such as latent traits

factor analysis
a mathematical method of summarising a matrix of values (such as the inter-correlation of test scores) in terms of a smaller number of values (factors) from which the original matrix can be reproduced

factor extraction
the process of calculating the factor or factors that can summarise a matrix of correlations among scores on tests or test items

factor loading
the correlation of scores on a test or test item and a factor score, and that can be used in identifying the nature of the factor

factor score
the score that a person has on a factor and that is often interpreted to reflect their standing on a latent trait

factor structure
the matrix of factor loadings for a set of variables, usually after factor rotation

false negative decision
a decision that incorrectly allocates a test taker or person being assessed to

the category of those predicted not to show some behaviour of interest on the basis of their score on a test or other assessment device

false positive decision
a decision that incorrectly allocates a test taker or person being assessed to the category of those predicted to show some behaviour of interest on the basis of their score on a test or other assessment device

fluid intelligence (Gf)
the more pure, inherited aspects of intelligence used to solve novel problems and deal with new situations

Flynn effect
refers to a steady increase in scores on IQ tests since about the 1930s; first drawn to the public's attention by James Flynn

forensic psychological testing and assessment
the collection of relevant and useful data and information using psychological tests and other assessment techniques to assist professionals in the legal and criminal justice systems to make decisions about offenders or those suspected of an offence

forensic psychologist
a psychologist who specialises in the provision of psychological services relating to the legal and criminal justice areas

forensic psychology
a branch of psychology that specialises in the application of psychological knowledge and skills to the working of the legal and criminal justice systems

formative assessment
an assessment aimed at facilitating learning as well as evaluating it

'g' (general mental ability)
the common variance when the results of different tests of mental ability are correlated (sometimes referred to as 'psychometric g', 'Spearman's g' or the 'general factor')

general mental ability
(GMA) global intellectual ability

generalisability theory
a set of ideas and procedures that follow from the proposal that the consistency or precision of the output of a psychological assessment device depends on specifying the desired range of conditions over which this is to hold

global intelligence
the overall or summary ability of an individual, which might be represented as the Full Scale IQ in modern intelligence tests; in hierarchical models of intelligence, global intelligence (or 'g') sits at the top of the intelligence hierarchy

graphic rating scale
a simple rating device used to elicit human judgment, typically completed by marking a point on a line or by circling a number (say from 1 to 10) to indicate the strength of agreement with the item

hierarchical models of intelligence
psychometric models that represent intelligence hierarchically, with many narrow abilities (first-order factors) at the first level, which define a smaller number of broader abilities (second-order factors), and the broader abilities are then represented by a general or 'g' factor at the top

high-stakes test
a test where the results have important consequences for the test taker

Holland's hexagon
a model that indicates the relationships among Holland's personality types and environments, with similar types placed closer to one another and dissimilar types placed farther away

implicit theories of intelligence
models or schema of the construct of intelligence generated by individuals and based largely on their observations of how the world works

incremental validity
the extent to which knowledge of score on a test (or other assessment device) adds to that obtained by another, pre-existing test score or psychological characteristic

industrial and organisational (I-O) psychology
the study of job performance and worker health issues to assist individuals, groups and organisations

integrity test
either a specific type of personality test or a direct measure to assess a job applicant's honesty, trustworthiness and reliability

intelligence
cognitive abilities such as problem solving and learning, although some definitions include other aspects of the individual such as personality and creativity

interpersonal approach
an approach to personality that proposes that personality exists only in the interaction between people and that the study of interpersonal processes is therefore central to personality assessment

inter-rater reliability
the extent to which different raters agree in their assessments of the same sample of ratees

interval scale
a scale that orders objects in terms of the attribute in such a way that the distances on the scale represent distances between objects

item
the various forms the content of a psychological test can take

item analysis
the process of studying the behaviour of items when administered to a group of respondents, usually with a view to the selection of some of the items to form a psychological test

item analysis indexes
the statistics arising in the process of item analysis used to evaluate each item in terms of its likely contribution to the psychological test being developed; in classical test theory they include item difficulty and item discrimination, and in item response theory they include the parameters of the ICC

item characteristic curve
the term for a trace line in item response theory

item-generation technology
new computer programs that focus on generating an item model or template, from which many individual items can be generated

item information
the term used in item response theory to describe the value of an item in identifying a respondent's position on the underlying trait of interest

item response theory
(IRT) a family of theories that specifies the functional relationship between a response to a single test item and the strength of the underlying latent trait

item score
the score for each item on the test

item validity
the extent to which the score on an item correlates with an external criterion relevant to the attribute or construct that is the subject of test construction

IQ (intelligence quotient)
the overall intelligence score obtained from one of the many current intelligence tests; the IQ score is a raw score conversion drawn from the normative sample, which has an arbitrary set mean of 100 and an arbitrary set standard deviation of 15 for each age group

job analysis
the process of gathering detailed information about the main tasks and contextual responsibilities for a particular job

job knowledge test
a test designed to assess knowledge, such as specific technical or professional knowledge, required for a job

job tryout
hiring someone for a short period of time to determine how well they fit in and perform on the job; a probationary period has a similar purpose

KSAOs
the knowledge, skills, abilities and other characteristics of an employee or prospective employee needed to be able to undertake their job satisfactorily

language
for most right-handers, the function of the left cerebral hemisphere; it includes the ability to understand and produce speech

latent factor-centred design
the use of underlying, latent constructs to represent both multiple measures (e.g. scores for reading, arithmetic and geography) and single tests (e.g. self-regulation); latent constructs reflect more 'pure' and efficient representations of a group of tests or a group of items

latent trait
the hypothesised continuously and normally distributed dimension of individual differences that is the sole source of a consistent set of observable behaviours, thoughts and feelings, which is the target of a psychological test

latent variable
a variable that is not directly observable but is hypothesised to exist on the basis of psychological theory; when it is a variable giving rise to individual differences it is referred to as a latent trait

Likert scale
a graphical scale originally with five points used by a respondent to represent the strength of an underlying attitude or emotion

linear transformation
a transformation that preserves the order and equivalence of distance of the original set of scores

local independence
the situation where the only factors influencing response to a psychological test item are the item's difficulty and the respondent's position on the underlying trait; for example, exposure to other items of the test does not increase or decrease the probability of responding in a particular way

local norms
norms developed for specific population groups or geographical regions

logit scale
an equal interval scale that locates the person's standing on the underlying trait of interest in terms of the percentage of items they get correct on the test and the average difficulty level of the items

malingering
responding or behaving in such a way to present oneself in a negative or positive manner during psychological testing

measurement
the assignment of numbers to objects according to a set of rules for the purpose of quantifying an attribute

memory
the ability to encode, store and retrieve past information

mental status examination
a comprehensive set of questions and observations used by psychologists to gauge the mental state of a client, which usually covers areas such as appearance, behaviour, orientation, memory, sensorium, affect, mood, thought content and thought process, intellectual resources, insight and judgment

method variance
the variability among scores on a psychological test or other assessment device that arises because of the form as distinct from the content of the test

Minnesota Multiphasic Personality Inventory (MMPI)
a test developed to assess major patterns of personality and emotional disorders using the empirical-keying approach; the latest version, MMPI–2 was published in 1989 and it requires a test taker to respond to 567 items and takes 60 to 90 minutes to complete

model of measurement
the formal statement of observations of objects mapped to numbers that represent relationships among the objects

motor functions
abilities such as lateral dominance, strength, fine motor skills (speed and dexterity), sensorimotor integration and praxis

multidimensional adaptive technology
(MAT) programs that allow assessment of multiple dimensions of a construct of interest, which allows for a better fit between the theorised, multidimensional construct or model and the obtained data than assessing a single dimension

multiple choice test
(MCT) a test where each question has a number of options, of which only one is correct

multiple intelligences
a theory usually associated with Howard Gardner, who proposed that intelligence comprises multiple, discrete modalities that are not aggregated to 'g'

multitrait–multimethod matrix
the pattern of correlations resulting from testing all possible relationships among two or more methods of assessing two or more constructs

multivariate (trait) approach
the oldest approach to personality that in its modern form proposes that there are a number of dimensions of individual difference that people have in common and that serve to specify the individual's personality

neuropsychological assessment
the application of neuropsychological tests and other data-collection techniques to answer referral questions or solve problems for individuals with a known or suspected brain injury

neuropsychology
a branch of psychology that aims to study the relationships between the brain and behaviour

nominal measurement
the lowest form of measurement that assigns numbers to objects to represent their discreteness from each other

nonlinear transformation
a transformation that preserves the order but not the equivalence of distance of the original scores

normal curve
a bell-shaped distribution of scores that conforms to a particular mathematical function that is a good approximation for random variables that cluster around a single mean

normalised standard score
a score in a distribution that has been altered to conform to a normal distribution by calculating the z scores for each percentile equivalent of the original raw score distribution

normative sample
tables of the distribution of scores on a test for specified groups in a population that allow interpretation of any individual's score on the test by comparison to the scores for a relevant group

norm-referenced test
a psychological test that uses the performance of a representative group of people (i.e., the norm) on the test for evaluating the performance of a test taker

norm referencing
a way of giving meaning to a test score by relating it to the performance of an appropriate reference group for the person

norms
tables of the distribution of scores on a test for specified groups in a population that allow interpretation of any individual's score on the test by comparison to the scores for a relevant group

objective procedure
the use of the same standardised materials, administration instructions, time limits and scoring procedures for all test takers

ordinal scale
a scale that has the property of a nominal scale, but also identifies an ordering of objects in terms of the attribute

organisational psychologist
a psychologist who specialises in the area of work, human resource management and organisational training and development

paradigms in personality assessment
approaches to personality assessment that share: assumptions about how personality is best studied; methods for collecting personality data; and criteria for making judgments about what constitute adequate statements about personality

peer rating
a rating of the KSAOs of an internal job applicant by the job applicant's co-worker/s

percentile
an expression of the position of a score in a distribution of scores by dividing the distribution into 100 equal parts; also known as 'centile'

performance appraisal
the assessment of a worker's job performance, typically carried out on a regular basis, such as six-monthly or annually

performance test
a psychological test that requires test takers to respond by answering questions or solving problems; they are usually administered individually

Personality Assessment Inventory(PAI)
a 344-item self-report scale designed to collect information relating to clinical diagnosis, treatment planning and screening for psychopathology in adults

personnel selection
the process of choosing which job applicants should receive an offer of employment

personological approach
an approach to personality that began with the work of Henry Murray who sought to study personality in terms of the (principally) psychogenic needs of the individual and the extent to which the environment promoted or inhibited these needs

person–organisation fit
compatibility between the individual and organisations that occurs when one of the parties can satisfy the needs of the other, or both have their needs satisfied

plan for item writing
a plan of the number and type of items that are required for a test, as indicated in the test specification

positive psychology
a relatively recent approach in psychology that stresses the behaviours, thoughts and feelings that characterise optimal functioning rather than dysfunction

predictive validity
the extent to which a score on a psychological test (or other assessment device) allows a statement about standing on a variable indexing important social behaviour independent of the test

primary mental abilities
seven broad ability factors that were identified by Thurstone: verbal comprehension, reasoning, perceptual speed, numerical ability, word fluency, associative memory and spatial visualisation; initially thought to be independent of one another, they were later shown to be correlated, and thus to also contain a 'g' factor

psychoanalytic approach
an approach to personality that originated in the work of Sigmund Freud on the role of unconscious motivational processes in normal and abnormal personality functioning; it was elaborated on by a number of researchers during the course of the twentieth century

psychological assessment
a broad process of answering referral questions, which includes but is not limited to psychological testing

psychological report
a report to provide a client or a referral agent with the answer(s) to the referral questions based on results of testing and assessment; it is usually provided in a written format that has a commonly agreed structure

psychological test
an objective procedure for sampling and quantifying human behaviour to make inferences about a particular psychological construct or constructs using standardised stimuli and methods of administration and scoring

psychological testing
the process of administering a psychological test, and obtaining and interpreting the test scores

psychometric properties
the criteria that a psychological test has to fulfil in order to be useful; they include how accurate and reproducible the test scores are, and how well the test measures what it intends to measure

psychometric theory
a theory concerned with the measurement of psychological constructs (like intelligence); the two main theories underpinning test development are classical test theory and item response theory; psychometric techniques typically include factor analysis and its variants

random sampling
a procedure in which every member of a population of interest has an equal probability of being selected and the selection of one member does not affect in any way the selection of any other member

Rasch model
a model that relates the probability of response of a particular sort (e.g. right/wrong) to the difference between a person's standing on a latent variable and the difficulty of the item

ratio scale
a scale that has the properties of an interval scale but also has a true zero

rational-empirical approach
a way of constructing psychological tests that relies on both reasoning from what is known about the psychological construct to be measured in the test, and collecting and evaluating data about how the test and the items that comprise it actually behave when administered to a sample of respondents

raw score total
(or raw score) the total score on the test found by summing item scores

receiver operating characteristic (ROC) curve
the curve of sensitivity against 1 minus specificity

reference check
a means of verifying job applicant information provided in a resume and collected in an interview; typically done by contacting past employers and/or individuals who can vouch for the applicant

referral question
a request for psychological testing or assessment is usually raised by a client or other professionals who work with the client; it can be general or specific

reliability
the consistency with which a test measures what it purports to measure in any given set of circumstances

reliability coefficient
an index—often a Pearson product moment correlation coefficient—of the ratio of true score to error score variance in a test as used in a given set of circumstances

RIASEC
John Holland's codes for the six types of individual and workplace 'personalities' that he identified (Realistic, Investigative, Artistic, Social, Enterprising and Conventional)

risk assessment
an assessment conducted to determine how risky or dangerous an inmate is for the purpose of sentencing, parole or classification

rotation
a method of varying the loadings of a factor on each of a set of variables, originally performed geometrically, with a view to producing a more psychologically meaningful factor structure

's' (specific ability)
limited to a single or small number of tasks, as opposed to 'g', which is

reflected in all mental ability tasks; all tasks require the application of 'g' and 's', and individuals differ on levels of both

selection interviews
usually included as part of any selection exercise, interviews generate ratings based on job applicant responses to questions, which are used to predict success on the job

selection on the criterion
in personnel selection, the process of appointing all job applicants for a trial period and then retaining only those who have performed satisfactorily

selection ratio
the proportion of those tested or assessed who can be allocated to the category of showing the behaviour of interest in a given psychological testing or assessment situation

self-report test
a psychological test that requires test takers to report their behaviour or experience; these tests can be administered individually or in a group

sensitivity
the proportion of those who have the behaviour of interest who are so predicted by the test or assessment device

sensory functions
the ability to encode and perceive visual, auditory and somatosensory stimuli reliably and accurately

social-cognitive approach
an approach to personality that examines the relationships between people's behaviour, the situations in which these behaviours occur, and their cognitions about them

social desirability bias
a form of method variance common in the construction of psychological tests of personality that arises when people respond to questions that place them in a favourable or unfavourable light

specific-ability test
an individual test or test battery that is designed to assess specific or

narrow cognitive abilities, rather than generate a measure of broader abilities or 'g'

specificity
the proportion of those who do not show the behaviour of interest who are so predicted by the test or assessment device

split-half reliability
the estimate of reliability obtained by correlating scores on the two halves of a test formed in some systematic way (e.g. odd versus even items)

standard
a fixed level of attainment

standard error of estimate
an index of the amount of error in predicting one variable from another

standard error of measurement
an expression of the precision of an individual test score as an estimate of the trait it purports to measure

standard score
the distance of a score in a normal distribution from the mean expressed as a ratio of the standard deviation of the distribution

standardised score
a score based on a z score but set to a distribution with a particular mean and standard deviation considered convenient for a particular purpose

standardised test
a test administered and scored in a set way

Stanford-Binet Intelligence Scale
Lewis Terman of Stanford University revised the Binet-Simon test for use in the US; released in 1916, the Stanford-Binet has been revised many times and continues to be widely used

stanine
a score on a nine-point scale with the points set in terms of percentiles

sten score
a point on a scale that has 5 units above and 5 units below the mean, which is set at 5.5 with a standard deviation of 2

stratified sampling
a method of sampling in which the sample is drawn from the population in such a way that it matches it with respect to a number of characteristics that are considered important for the purposes of the study

structure-of-intellect (SOI) model
JP Guilford's multifaceted model of intelligence consisting of 150 intellectual abilities arranged along three dimensions of operations, content and product

summative assessment
an assessment that has a purely evaluative function

T score
a score standardised to a distribution with a mean of 50 and a standard deviation of 10

task performance
the core technical aspects and basic tasks that comprise a job

test bias
the systematic favouring of one group over another in test outcomes; this can be due to more than one cause

test manual
the document that accompanies a psychological test and that records the way in which the test was developed, how the test is to be administered (including the groups for which it is relevant), information on the reliability and validity of the test when used for use for specific purposes, and norms for test interpretation

test obsolescence
the notion that a psychological test loses its utility because the theory that it was based on has been shown to be wrong, or because the content of its items is no longer appropriate because of social or cultural change

test-retest reliability
the estimate of reliability obtained by correlating scores on the test obtained on two or more occasions of testing

test specification
a written statement of the attribute or construct that the test constructor

is seeking to measure and the conditions under which it will be used

therapeutic assessment
an assessment conducted by psychologists with the purpose of assisting and treating a client

time-parameterised testing
seeks to solve the problem of the trade-off between speed of responding on a test and accuracy of responding

trace line
a graph of the probability of response to an item as a function of the strength of or position on a latent trait

triarchic theory of intelligence
a theory proposed by Robert Sternberg in which intelligence comprises three components: analytical abilities ('componential'), creative abilities ('experiential') and practical abilities ('contextual'); it suggests that individuals high on the three components should experience real-life success

two-factor (Gf-Gc) theory of intelligence
Cattell's original theory, which decomposed 'g' into two component parts: fluid and crystallised intelligence (Gf and Gc)

type of measurement
the scales of measurement proposed by Stevens; that is, nominal, ordinal, interval and ratio

valid negative decision
a decision that correctly allocates a test taker or person being assessed to the category of those predicted not to show some behaviour of interest on the basis of their score on a test or other assessment device

valid positive decision
a decision that correctly allocates a test taker or person being assessed to the category of those predicted to show some behaviour of interest on the basis of their score on a test or other assessment device

validity
the extent to which evidence supports the meaning and use of a psychological test (or other assessment device)

validity generalisation
(VG) the demonstration that validity generalises across job selection exercises for different jobs by conducting meta-analyses of studies reporting validity coefficients

variance accounted for
the percentage of variance of the total or reliable variance of a set of variables that each factor has in common with the variables as a set; it is calculated from the matrix of factor loadings

virtual reality
a computer-based technology that mimics a real (e.g. for pilot training) or constructed environment (e.g. game technologies) for the user, who is placed in, and can interact with, the environment; advanced packages also include touch and smell

visuo-spatial functions
usually considered functions of the right cerebral hemisphere, include the ability to perceive and understand visuo-spatial relationships and undertake three-dimensional constructional tasks

vocational interests
interests with specific relevance to the workplace, which tend to be stable over time, influence motivation and behaviour, and indicate the type of activities and environments the person prefers

Wechsler Adult Intelligence Scale (WAIS)
developed by David Wechsler, and one of the most widely used, individually administered, intellectual assessment batteries; the latest version, WAIS–IV, was published in 2008

Wechsler-Bellevue Intelligence Scale
the forerunner to the popular Wechsler Adult Intelligence Scale, it was created by David Wechsler and released in 1939 as a test of general intellectual ability; revised many times, it remains the most widely used individual test of ability

work sample tests
based on the assumption that current, observed behaviour will predict future behaviour, they require job applicants to carry out tasks that mirror those that will be required on the job

z score
a linear transformation of test scores that expresses the distance of each score from the mean of the distribution of scores in units of the standard deviation of the distribution

References

Abu-Hamour, B, Al Hmouz, H, Mattar, J & Muhaidat, M (2012). The use of Woodcock-Johnson tests for identifying students with special needs—a comprehensive literature review. *Procedia—Social and Behavioral Sciences*, *47*, 665–73.

ACER—see Australian Council for Educational Research.

Achenbach, T M (1978). The Child Behavior Profile: I. Boys aged 6–11. *Journal of Consulting and Clinical Psychology*, 46(3), 478–88.

Achenbach, T M & Rescorla, L A (2001). *Manual for ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth and Families.

Acheson, S K (2005). Review of the Hare Psychopathy Checklist–Revised 2nd edition. In R A Spies & B S Plake (Eds.), *The sixteenth mental measurement yearbook* (pp. 429–31). Lincoln, NE: University of Nebraska Press.

Ackerman, M J (2010). *Essentials of forensic psychological assessment* (2nd ed.). Hoboken, NJ: Wiley.

Ackerman, M J & Kane, A W (1998). *Psychological experts in divorce actions* (3rd ed.). New York, NY: Aspen Law & Business.

Ackerman, M J & Schoendorf, K (1992). *Ackerman Schoendorf Scales of Parent Evaluation of Custody (ASPECT)*. Los Angeles, CA: Western Psychological Services.

Adams, Y, Drew, N & Walker, R (2014). Principles of practice in mental health assessment with Aboriginal Australians. In P Dudgeon, H Milroy & R Walker (Eds.), *Working together: Aboriginal and Torres Strait Islander mental health and welling principles and practice*. Canberra, ACT: Australian Government.

Ægisdóttir, S & Einarsdóttir, S (2012). Cross-cultural adaptation of the Icelandic Beliefs about Psychological Services Scale (I-BAPS). *International Perspectives in Psychology: Research, Practice, Consultation*, *1*(4), 236–51.

Ægisdóttir, S, White, M J, Spengler, P M, Maugherman, A S, Anderson, L A, Cook, R S, Nichols, C N, Lampropoulos, G K, Walker, B S, Cohen, G & Rush, J D (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341–82.

Aiken, L R (1996). *Assessment of intellectual functioning*. New York, NY: Plenum Press.

Alfonso, V C, Flanagan, D P & Radwan, S (2005). The impact of Cattell-Horn-Carroll theory on test development and the interpretation of cognitive abilities. In D P Flanagan & P L Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (2nd ed.; pp. 185–202*)*. New York, NY: Guilford Press.

Allen, M J & Yen, W M (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Allport, G W & Odbert, H S (1936). Trait names: A psycho-lexical study. *Psychological Monographs*, *47*(1, whole no. 211), 1–171.

Altman, D G & Bland, J M (1995). Absence of evidence is not evidence of absence. *British Medical Journal*, *311*, 485.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders (DSM–IV–TR)* (4th ed.; rev.). Washington, DC: Author.

American Psychological Association (2009). *Guidelines for child custody evaluations in family law proceedings*. Washington, DC: Author.

Anderson, J, Kearney, G E & Everett, A V (1968). An evaluation of Rasch's structural model for test items. *British Journal of Mathematical and Statistical Psychology*, *21*(2), 231–8.

Anderson, L W & Krathwohl, D R, et al. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Boston, MA: Allyn & Bacon.

Anderson, N, Schlueter, J E & Geisinger, K F (Eds.). (2016). *Tests in Print IX: An index to tests, test reviews, and the literature on specific tests by Buros Center*. Lincoln, NE: Buros Centre for Testing.

Andrews, G & Slade, T (2001). Interpreting scores on the Kessler Psychological Distress Scale (K10). *Australian and New Zealand Journal of Public Health*, *25*, 494–7.

Ansell, E B, Kurtz, J E, DeMoor, R M & Markey, P M (2011). Validity of the PAI Interpersonal Scales for measuring the dimensions of the interpersonal circumplex. *Journal of Personality Assessment*, *93*(1), 33–9.

Anstadt, S P, Bradley, S & Burnette, A (2013). Virtual worlds: Relationship between real life and experiences in Second Life. *International Review of Research in Open and Distance Learning, 14*, 160–90.

Antony, M M, Bieling, P J, Cox, B J, Enns, M W & Swinson, R P (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment, 10*, 176–81.

Archer, R P, Buffington-Vollum, J K, Stredny, R V & Handel, R W (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84–94.

Armstrong, P I, Allison, W & Rounds, J (2008). Development and initial validation of brief public domain RIASEC marker scales. *Journal of Vocational Behavior*, *73*, 287–99.

Armstrong, P I, Smith, T J, Donnay, D A C & Rounds, J (2004). The strong ring: A basic interest model of occupational structure. *Journal of Counseling Psychology*, *51*, 299–313.

Atkinson, J W, Bongort, K & Price, L H (1977). Explorations using computer simulation to comprehend thematic apperceptive measurement of motivation. *Motivation and Emotion*, *1*, 1–27.

Australian Bureau of Statistics (2009). Stroke. *Profiles of Disability, Australia, 2009*. Cat. no. 4429.0. Retrieved from www.abs.gov.au/ausstats/abs@.nsf/Lookup/4429.0main+features100262009.

Australian Bureau of Statistics (2009–10). *Year Book Australia, 2009–2010*. Retrieved from www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1301.02009%E2%80%9310?OpenDocument.

Australian Council for Educational Research (2003). *ACER select manual*. Camberwell, Vic: Author.

Australian Curriculum Assessment and Reporting Authority (ACARA) (2010). *National Assessment Program. Literacy and Numeracy. Frequently asked questions*. Retrieved from www.naplan.edu.au/faqs/napfaq.html. Cited in Brady, L (2013). NAPLAN: Critiquing the criticisms. *Curriculum and Teaching, 28*(1), 47–55.

Australian Primary Principals Association (2009). *Australian Primary Principals Association position paper on the publication of nationally comparable school performance information*. Retrieved from www.appa.asn.au/wp-content/uploads/2015/08/School-performance-information.pdf

Australian Primary Principals Association (2010). *The reporting and use of NAPLAN*. Kaleen, ACT: Author. Available at www.appa.asn.au.

Australian Psychological Society (2007). *Code of ethics*. Melbourne, Vic: Author.

Australian Psychological Society (2015). *Practice guide for psychological testing with people with disability*. Melbourne, Australia: Author.

Australian Psychology Accreditation Council (2010). *Rules for accreditation and accreditation standards for psychology courses*. Victoria: Author.

Bader, L A (1998). *Bader Reading and Language Inventory* (3rd ed.). Upper Saddle River, MN: Prentice Hall.

Bagby, R M, Wild, N & Turner, A (2003). Psychological assessment in adult mental health settings. In J R Graham & J A Naglieri (Eds.), *Handbook of Psychology: Vol. 10. Assessment psychology* (pp. 213–34). New York, NY: Wiley.

Baldwin, J M (1901). *Dictionary of philosophy and psychology*. New York, NY: Macmillan.

Bandura, A (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37, 122–47.

Bandura, A (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Barak, A & English, N (2002). Prospects and limitations of psychological testing on the internet. *Journal of Technology in Human Services, 19*, 65–89.

Baral, B D & Das, J P (2004). What is indigenous to India and what is shared? In R J Sternberg (Ed.), *International Handbook of Intelligence* (pp. 270–301). New York, NY: Cambridge University Press.

Barbazette, J (2006). *Training needs assessment: Methods, tools, and techniques*. New York, NY: John Wiley & Sons.

Bartle, R A (2010). From MUDs to MMORPGs: The history of virtual worlds. In J Hunsinger, L Klastrub & M Allen (Eds.), *International handbook of internet research* (pp. 23–39). New York, NY: Springer.

Bartram, D (2000). Internet recruitment and selection: Kissing frogs to find princes. *International Journal of Selection and Assessment, 8*, 261–74.

Bartram, D (2004). Assessment in organisations. *Applied Psychology: An International Review, 53*, 237–59.

Bartram, D (2005). The changing face of testing. *The Psychologist, 18*, 666–8.

Bartram, D & Brown, A (2004). Online testing: Mode of administration and the stability of OPQ32i scores. *International Journal of Selection and Assessment, 12*, 278–84.

Bartram, D & Hambleton, R K (2006). *Computer-based testing and the internet: Issues and advances.* Chichester, UK: John Wiley and Sons.

Bartram, D & Lindley, P A (1994). *Scaling, norms and standardization. BPS Opening Learning Programme on Psychological Testing.* London: British Psychological Society.

Baumeister, R F (1987). How the self became a problem: A psychological review of historical research. *Journal of Personality and Social Psychology, 52*(1), 163–76.

Beatty, W W, Ryder, K A, Gontkovsky, S T, Scott, J G, McSwan, K L & Bharucha, K J (2003). Analyzing the subcortical dementia syndrome of Parkinson's disease using the RBANS. *Archives of Clinical Neuropsychology, 18*, 509–20.

Beatty, W W, Mold, J W & Gontkovsky, S T (2003). RBANS Performance: Influences of sex and education. *Journal of Clinical and Experimental Neuropsychology, 25*, 1065–9.

Beck, A T & Steer, R A (1987). *Manual for the Beck Anxiety Inventory*. San Antonio, TX: The Psychological Corporation.

Beck, A T, Steer, R A & Brown, G K (1996). *Manual for the Beck Depression Inventory—Second Edition (BDI–II)*. San Antonio, TX: The Psychological Corporation.

Bennett, R E (2011). Formative assessment: A critical review. *Assessment in Education, Principles, Policy and Practice*, *18*(1), 5–25.

Benson, N, Hulac, D M & Kranzler, J H (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*, *22*, 121–30.

Bigler, E D, Rosa, L, Schultz, F, Hall, S & Harris, J (1989). Rey Auditory-Verbal Learning and Rey-Osterreith Complex Figure Design Test performance in Alzheimer's disease and closed head injury. *Journal of Clinical Psychology*, *45*, 277–80.

Binet, A & Henri, V (1895). La psychologie individuelle. *L'annee Psychologique*, *2*, 411–65.

Binet, A & Simon, T (1905). New methods for the diagnoses of the intellectual level of subnormals. *L'annee Psychologique, 11*, 191–244.

Bjelland, l, Dahl, A A, Haug, T T & Neckelmann, D (2002). The validity of the Hospital Anxiety and Depression Scale: An Updated literature review. *Journal of Psychosomatic Research, 52*, 69–77.

Black, P & Dylan, W (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Block, J (1995). A contrarian view of the five-factor approach to personality. *Psychological Bulletin*, 117, 187–215.

Bloom, B S (1969). Some theoretical issues relating to educational evaluation. In R W Tyler (Ed.), *Educational evaluation: New roles, new means. The 63rd yearbook of the National Society for the Study of Education, Part 2 (Vol. 69)* (pp. 26–50). Chicago, IL: University of Chicago Press.

Board, B J & Fritzon, K (2005). Disordered personalities at work. *Psychology, Crime & Law*, *11*(1), 17–32.

Bonnie, R (1992). The competence of criminal defendants: A theoretical reformulation. *Behavioural Science and the Law*, *10*, 291–316.

Bonnie, R (1993). The competence of criminal defendants: Beyond Dusky and Drope. *Miami Law Review*, *47*, 539–601.

Bonta, J (1996). Risk-needs assessment and treatment. In A T Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp. 18–32). Thousand Oaks, CA: Sage Publications.

Boring, E G (1923). Intelligence as the tests test it. Cited in T B Rogers (1995), *The psychological testing enterprise: An introduction*. Belmont, CA: Wadsworth.

Borman, W C & Motowidlo, S J (1993). Expanding the criterion domain to include elements of contextual performance. In N Schmitt & W Borman (Eds.), *Personnel selection in organizations*. San Francisco, CA: Jossey-Bass.

Bowman, M L (1989). Testing individual differences in Ancient China. *American Psychologist*, *44*, 576–8.

Brady, L (2013). NAPLAN: Critiquing the criticisms. *Curriculum and Teaching*, *28*(1), 47–55.

Bray, D W & Grant, D L (1966). The assessment center in the measurement of potential for business management. *Psychological Monographs*, 80 (whole no. 625).

Brenner, C (1974). *An elementary textbook of psychoanalysis* (revised ed.). New York, NY: Random House.

Brenner, E (2003). Consumer-focused psychological assessment. *Professional Psychology*, *34*, 240–7.

Brody, N (1972). *Personality: Research and theory*. New York, NY: Academic Press.

Brody, N & Ehrlichman, H (1998). *Personality psychology: the science of individuality.* Upper Saddle River. NJ: Prentice Hall.

Brown, T A, Chorpita, B F, Korotitsch, W & Barlow, D H (1997). Psychometric properties of the Depression Anxiety Stress Scales (DASS) in clinical samples. *Behaviour Research and Therapy*, *35*, 79–89.

Burger, J M (2000). *Personality* (5th ed.). Belmont. CA: Wadsworth.

Butcher, J N (2012). Computerised personality assessment. In *Handbook of Psychology* (2nd ed.). New York, NY: Basic Books.

Butcher, J N, Dahlstrom, W G, Graham, J R, Tellegen, A & Kaemmer, B (1989). *Minnesota Multiphasic Personality Inventory–Second Edition (MMPI–2): Manual for administration and scoring.* Minneapolis, MN: University of Minnesota Press.

Camara, W J, Nathan, J S & Puente, A E (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, *31*, 141–54.

Campbell, D P & Holland, J L (1972). A merger in vocational interest research: Applying Holland's theory to Strong's data. *Journal of Vocational Behavior*, *2*, 353–76.

Campbell, J P, McCloy, R A, Oppler, S H & Sager, C E (1993). A theory of performance. In N Schmitt & W Borman (Eds.), *Personnel selection in organizations*. San Francisco, CA: Jossey-Bass.

Canivez, G L & Watkins, M W (2016). Review of the Wechsler Intelligence Scale for Children–Fifth Edition: Critique, commentary, and independent analyses. In A S Kaufman, S E Raiford & D L Coalson (Eds.), *Intelligent testing with the WISC–V* (pp. 683–702). Hoboken, NJ: Wiley.

Carlson, J F, Geisinger, K F & Jonson, J L (Eds.). (2014). *The nineteenth mental measurements yearbook*. Lincoln, NE: Buros Center for Testing.

Carpenter, P A, Just, M A & Shell, P (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404–31.

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

Carstairs, J & Myors, B (2009). Internet testing: A natural experiment reveals test score inflation on a high-stakes, unproctored cognitive test. *Computers in Human Behavior*, *25*, 738–42.

Carstairs, J R & Shores, E A (2000). The Macquarie University Neuropsychological Normative Study (MUNNS): Rationale and methodology. *Australian Psychologist*, *35*, 36–40.

Carstairs, J R, Myors, B, Shores, E A & Fogarty, G (2006). Influence of language background on tests of cognitive abilities: Australian data. *Australian Psychologist*, *41*, 48–54.

Cattell, R B (1940). A culture-free intelligence test. *Journal of Educational Psychology*, *31*, 161–79.

Cattell, R B (1946). *Description and measurement of personality*. Yonkers-on-Hudson, NY: World Book.

Cattell, R B (1957). *Personality and motivation: Structure and measurement*. Yonkers, NY: World Book Company.

Cattell, R B (1979). Are culture fair tests possible and necessary? *Journal of Research and Development in Education*, *12*(2), 2–13.

Cattell, R B (1987). *Intelligence: Its structure, growth, and action*. New York, NY: Elsevier.

Cattell, R B, Cattell, A K & Cattell, H E P (1993). *16PF Fifth Edition Questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.

Cattell, H E P & Mead, A D (2008). The Sixteen Personality Factor Questionnaire. In G J Boyle, G Matthews & D H Saklofske (Eds.), *The Sage Handbook of Personality Theory and Testing: Vol. 2,*

Personality Measurement and Testing. Los Angeles, CA: Sage Publications.

Cattell, R B & Muerle, J L (1960). The 'maxplane' program for factor rotation to oblique simple structure. *Educational and Psychological Measurement*, *20*, 569–90.

Cervone, D (2005). Personality architecture: Within-person structures and processes. *Annual Review of Psychology*, *56*, 423–52.

Cervone, D, Shadel, W G & Jencius, S (2001). Social-cognitive theory of personality assessment. *Personality and Social Psychology Review*, *5*, 33–51.

Chan, A S, Shum, D & Cheung, R W Y (2003). Recent development of cognitive and neuropsychological assessment in Asian countries. *Psychological Assessment*, *15*(3), 257–67.

Chan, R C K, Shum, D, Toulopoulou, T & Chen, E Y H (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology*, *23*, 201–16.

Charsky, D (2010). From edutainment to serious games: A change in the use of game characteristics. *Games and Culture, 5*, 177–98.

Charter, R A, Walden, D K & Padilla, S (2000). Too many simple clerical scoring errors: The Rey Figure as an example. *Journal of Clinical Psychology*, *56*, 571–4.

Chen, C (2007). Intelligent web-based learning system with personalized learning path guidance. *Computers & Education*, *51*, 787–814.

Chen, J Q (2004). The Project Spectrum approach to early education. In J Johnson & J Roopnarine (Eds.), *Approaches to early childhood education* (4th ed.; pp. 251–9). Upper Saddle River, NJ: Pearson.

Chen, J Q & Gardner, H (2012). Assessment of intellectual profile: A perspective from multiple-intelligences theory. In D P Flanagan & P L Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.; pp. 145–55). New York, NY: Guilford.

Ciechalski, J C (2009). Review of the Self-Directed Search. In E A Whitfield, R W Feller & C Wood (Eds.), *A Counselor's Guide to Career Assessment Instruments* (5th ed.). Broken Arrow, OK: National Career Development Association.

Cizek, G J (2001). More unintended consequences of highstakes testing. *Educational Measurement: Issues and Practice*, Winter, 19–27.

Cizek, G J (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31–43.

Claes, L, Van Mechelen, I & Vertommen, H (2004). Assessment of situation-behaviour profiles and their guiding cognitive and affective processes: A case study from the domain of aggressive behaviors. *European Journal of Psychological Assessment*, *20*(4), 216–26.

Clark, B (1988). *Growing up gifted: Developing the potential of children at home and at school* (3rd ed.). Columbus, OH: Merrill.

Clark, L A & Watson, D (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–19.

Cleary, T A, Humphreys, L G, Kendrick, S A & Wesman, A (1975). Educational uses of tests with disadvantaged students. *American Psychologist*, 15–41.

Cliff, N (1982). What is and isn't measurement. In G Keren (Ed.), *Statistical and methodological issues in psychology and social sciences research* (pp. 3–38). Hillsdale, NJ: Erlbaum.

Cole, M (1999). Culture-free versus culture-based measures of cognition. In R J Sternberg (Ed.), *The nature of cognition* (pp. 645–64). Cambridge, MA: The MIT Press.

Coltheart, M & Caramazza, A (2006). *Cognitive neuropsychology twenty years on*. Hove, UK: Psychology Press.

Commonwealth of Australia (2000). *Immigration Restriction Act 1901*. Available at www.foundingdocs .gov.au/places/cth/cth4ii.htm.

Costa, P T & McCrae, R R (1992a). Four ways five factors are basic. *Personality and Individual Differences, 13*, 653–65.

Costa, P T & McCrae, R R (1992b). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P T & McCrae, R R (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G J Boyle, G Matthews & D H Sakloske (Eds.), *The Sage Handbook of Personality Theory and Assessment, Vol 2. Personality Measurement and Testing* (pp. 179–98). Thousand Oaks, CA: Sage.

Crawford, J R & Henry, J D (2003). The Depression Anxiety Stress Scales (DASS): Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology, 42*, 111–31.

Crawford, J, Cayley, C, Lovibond, P F, Wilson, P H & Hartley, C (2011). Percentile norms and accompanying interval estimates from an Australian general adult population sample for self-report mood scales (BAI, BDI, CRSD, CES-D, DASS, DASS-21, STAI-X, STAI-Y, SRDS, and SRAS). *Australian Psychologist, 46*, 3–14.

Cronbach, L J (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cronbach, L J (1970). *Essentials of psychological testing* (3rd ed.). New York, NY: Harper & Row.

Cronbach, L J (1990). *Essentials of psychological testing* (5th ed.). New York, NY: Harper Collins.

Cronbach, L J & Gleser, G C (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Chicago Press.

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

Cronbach, L J, Gleser, G C, Nanda, H & Rajaratnam, N (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Crowne, D P & Marlowe, D (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349–54.

Dahlstrom, W G (1993). Tests: Small samples, large consequences. *American Psychologist, 48*, 393–9.

Dahlstrom, W G & Welsh, G S (1960). *An MMPI handbook*. London: Oxford University Press.

Darby, D & Walsh, K (2005). *Walsh's neuropsychology: A clinical approach*. Edinburgh, UK: Elsevier Churchill Livingstone.

Davidson, G (1995). Cognitive assessment of indigenous Australians: Towards a multiaxial model. *Australian Psychologist, 30*, 30–4.

Davidson, G (1996). Fairness in a multicultural society! Reply to Dyck. *Australian Psychologist, 31*, 70–2.

Davis, A G (1993). *A survey of adult aphasia* (2nd ed.). Engelwood Cliffs, NJ: Prentice Hall.Dawes, R M (1976). Shallow Psychology. In J S Carrol & J W Payne (Eds.), Cognition and social behaviour (pp. 3–11). Potomac, M D: Lawrence Erlbaum.

de Lemos, M M (1969). The development of conservation in Aboriginal children. *International Journal of Psychology, 4*, 255–69.

Delis, D, Kaplan, E & Kramer, J (2001). *Delis-Kaplan Executive Function Scale*. San Antonio, TX: Psychological Corporation.

Department of Education and Training (2014). *Selected higher education statistics—2014 student data*. Retrieved from https://education.gov.au/selected-higher-education-statistics-2014-student-data.

Dickson, D H & Kelly, I W (1985). The 'Barnum effect' in personality assessment: A review of the literature. *Psychological Reports, 57*,

367–82.

Dingwall, K M & Cairney, S (2010). Psychological and cognitive assessment of Indigenous Australians. *Australian and New Zealand Journal of Psychiatry*, *44*, 20–30.

Doll, E A (1935). A genetic scale of social maturity. *American Journal of Orthopsychiatry*, *5*, 180–8.

DuBois, P E (1970). *A history of psychological testing*. Boston, MA: Allyn & Bacon.

Dudgeon, P, Milroy, H & Walker, R (2014). *Working together: Aboriginal and Torres Strait Islander mental health and welling principles and practice*. Canberra, Australia: Australian Government.

Duff, K, Pattern, D, Schoenberg, M R, Mold, J, Scott, J G & Adams, R L (2003). Age- and education-corrected independent normative data for the RBANS in a community dwelling elderly sample. *Clinical Neuropsychologist*, *17*, 351–66.

Dulfer, N, Polesel, J & Rice, S (2012). *The experience of education: The impacts of high stakes testing on school students and their families: An educator's perspective*. The Whitlam Institute, University of Western Sydney. Retrieved 2 August 2016 from www.whitlam.org/__data/assets/pdf_file/0010/409735/High_Stakes_Testing_An_Educators_Perspective.pdf.

Dyck, M (1996). Cognitive assessment in a multicultural society: Comment on Davidson. *Australian Psychologist*, *31*(1), 66–9.

Dziegielewski, S F (2015). *DSM-5 in action* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Ellenberger, H (1970). *The discovery of the unconscious*. New York, NY: Basic Books.

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

Entwisle, D R (1972). To dispel fantasies about fantasy-based measures of achievement motivation. *Psychological Bulletin, 77,* 377–91.

Eysenck, H J & Eysenck, M W (1975a). *Personality and individual differences: A natural science approach*. New York, NY: Plenum.

Eysenck, H J & Eysenck, S B G (1975b). *Manual of the Eysenck Personality Questionnaire*. London: Hodder and Stoughton.

Family Court of Australia, Federal Circuit Court of Australia & Family Court of Western Australia (2015). *Australian Standards of Practice for Family Assessments and Reporting—February 2015*. Retrieved from www.familycourt.gov.au/wps/wcm/connect/fcoaweb/about/policies-and-procedures/asp-family-assessments-reporting.

Faust, D & Ziskin, J (1988). The expert witness in psychology and psychiatry. *Science, 241*, 31–5.

Faust, D, Ziskin, J & Hiers, J B (1991). *Brain damage claims: Coping with neuropsychological evidence*. Los Angeles, CA: Law and Psychology Press.

Faust, K, Nelson, B D, Sarapas, C & Pliskin, N H (2016): Depression and performance on the repeatable battery for the assessment of neuropsychological status. *Applied Neuropsychology: Adult, 7*, 1–7

First, M B, Williams, J B W, Karg, R S & Spitzer, R L (2016). *Structured Clinical Interview for DSM-5 Disorders Clinical Version (SCID-5-CV)*. Washington, DC: American Psychiatric Press.

Fiske, D W (1966). Some hypotheses concerning test adequacy. *Educational and Psychological Measurement, 2*, 69–88.

Fiske, D W (1971). *Measuring the concepts of personality*. Chicago, IL: Aldine.

Flanagan, D P, Alfonso, V C & Ortiz, S O (2012). The cross-battery assessment approach. In D P Flanagan & P L Harrison (Eds.),

Contemporary intellectual assessment: Theories, tests, and issues (3rd ed.; pp. 459–83). New York, NY: Guilford.

Flanagan, D P, Ortiz, S O & Alfonso, V C (2013). *Essentials of cross battery assessment* (3rd ed.). Hoboken, NJ: Wiley.

Flanagan, D P, Ortiz, S O, Alfonso, V C & Mascolo, J T (2002). *The achievement test desk reference (ADTR): Comprehensive assessment and learning disabilities*. Boston, MA: Allyn & Bacon.

Flanders, L R & Utterback, D (1985). Management Excellence Inventory: A tool for management development. *Public Administration Review*, *45*, 403–10.

Fleenor, J W & Prince, J M (1997). *Using 360-degree feedback in organizations*. Greensboro, NC: Center for Creative Leadership. Retrieved from www.ccl.org/leadership/pdf/research/using360feedback.pdf.

Fleishman, E A & Reilly, M E (1992). *Fleishmann Job Analysis Survey*. Santa Clare, CA: Consulting Psychologists Press.

Fletcher, J M, Coulter, W A, Reschly, D J & Vaughn, S (2004). Alternative approaches to the definition and identification of learning disabilities: Some questions and answers. *Annals of Dyslexia*, *54*(2), 304–31.

Flynn, J R (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 171–91.

Forer, B R (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology*, *44*(1), 118–23.

Fowler, R D & Matarazzo, J D (1988). Psychologists and psychiatrists as expert witnesses. *Science*, *241*, 1143.

Fowler, R D, Seligman, M E P & Koocher, G (1999). The APA 1998 Annual Report. *American Psychologist*, 537–68.

Frances, A J & Widiger, T (2012). Psychiatric diagnosis: Lessons from the DSM-IV past and cautions for the DSM-5 future. *Annual Review of Clinical Psychology*, *8*(1), 109–30. doi:10.1146/annurev-clinpsy-032511-143102.

Francis, R D (1999). *Ethics for psychologists: A handbook*. Melbourne, Vic: ACER Press.

Frank, L (1939). Projective methods for the study of personality. *Journal of Psychology*, *8*, 389–413.

Furnham, A, Eysenck, S B G & Saklofske, D (2008). The Eysenck personality measures: Fifty years of scale development. In G J Boyle, G Matthews & D H Sakloske (Eds.), *The Sage Handbook of Personality Theory and Assessment, Vol 2. Personality Measurement and Testing* (pp. 199–218). Thousand Oaks, CA: Sage.

Furr, R M & Bachrach, V R (2014). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: Sage.

Gardner, H (1983). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic.

Gardner, H (1999). *Intelligence reframed: Multiple intelligences for the 21st century*. New York, NY: Basic Books.

Gardner, H (2006). *Five minds for the future*. Boston, MA: Harvard Business School Press.

Gazzaniga, M S & Heatherton, T F (2003). *Psychological science: Mind, brain, and behavior*. New York, NY: W W Norton.

Geffen, G M, Butterworth, P & Geffen, L B (1994). Test-retest reliability of a new form of the Auditory Verbal Learning Test (AVLT). *Archives of Clinical Neuropsychology*, *9*, 303–16.

Geffen, G M, Moar, K S, O'Hanlon, A P, Clark, C R & Geffen, M B (1990). Performance measures of 16 to 86 year old males and females on the Auditory Verbal Learning Test. *Clinical Neuropsychologist*, *4*, 45–63.

Geiger, M A, Boyle, E J & Pinto, J K (1993). An examination of ipsative and normative versions of Kolb's Revised Learning Style Inventory. *Educational and Psychological Measurement, 53*, 717–26.

George, J M & Jones, G R (1997). Experiencing work: Values, attitudes and moods. *Human Relations, 50*, 393–416.

Gilliland, S W (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology, 79*, 691–701.

Giordano, P J (1997). Establishing rapport and developing interview skills. In J R Matthews & C E Walker (Eds.), *Basic skills and professional issues in clinical psychology* (pp. 59–82). Boston, MA: Allyn & Bacon.

Glennon, J R, Albright, LE & Owens, WA (1965). *A catalogue of life history items*. NC: Richardson Foundation.

Gold, J M, Iannone, V & Buchanan, R W (1999). Repeatable battery for the assessment of neuropsychological status as a screening test in schizophrenia: Sensitivity, reliability, and validity. *American Journal of Psychiatry, 156*, 1944–50.

Goldberg, L R (1999). A broad-bandwidth, public-domain personality inventory measuring the lower level facets of several Five-Factor models. In I Mervielde, I J Deary, F De Fruyt & F Ostendorf, *Personality psychology in Europe, Vol. 7* (pp. 7–28). Tilburg: Tilburg University Press.

Golden, C J (1978). *Stroop Colour and Word Test*. Chicago: Stoelting Company.

Golden, C J & Freshwater, S M (2002). *Stroop Colour and Word Test*. Chicago: Stoelting Company.

Golden, C J, Espe-Pfeifer, P & Wachsler-Felder, J (2000). *Neuropsychological interpretation of objective psychological tests*. New York, NY: Kluwer Academic.

Goldstein, G (1992). Historical perspectives. In A E Puente & R J McCaffrey (Eds.), *Handbook of neuropsychological assessment: A biopsychosocial perspective* (pp. 1–9). New York, NY: Plenum.

Goldstein, G & Hersen, M (2000). *Handbook of psychological assessment* (3rd ed.). New York, NY: Pergamon.

Gontkovsky, S T, Beatty, W W & Mold, J W (2004). Repeatable Battery for the Assessment of Neuropsychological Status in a normal, geriatric sample. *Clinical Gerontologist, 27*, 79–86.

Goodenough, F L (1949). *Mental testing: Its history, principles, and applications*. New York, NY: Rinehart.

Goodglass, H, Kaplan, E & Barresi, B (2000). *Boston Diagnostic Aphasia Examination* (3rd ed.). Philadelphia, PA: Lippincott Williams and Wilkins.

Goodnow, J (1976). Some sources of cultural difference in performance. In G E Kearney & D W McElwain (Eds.), *Aboriginal cognition: Retrospect and prospect*. Canberra, Australia: Australian Institute of Aboriginal Studies.

Gorsuch, R L (2006). *The 16PF Express Edition: A supplemental chapter to the 16PF Fifth Edition Administrator's Manual*. Champaign, IL: Institute for Personality and Ability Testing.

Gottfredson, G D, Holland, J L & Ogawa, D K (1982). *Dictionary of Holland Occupational Codes*. California: Consulting Psychologists Press.

Gottfredson, L S (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*, 13–23.

Gottfredson, L S (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 31*, 343–97.

Gottfredson, L S (2005). Using Gottfredson's theory of circumscription and compromise in career guidance and counseling. In S D Brown &

R W Lent, *Career development and counseling: Putting theory and research to work* (pp. 71–100). New York, NY: Wiley.

Gottfredson, L S (2009). Logical fallacies used to dismiss the evidence on intelligence testing. In R P Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 11–65). Washington, DC: American Psychological Association.

Gottfredson, S D & Moriarty, L J (2006). Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, *52*(1), 178–200.

Gough, H (1987). *California Psychological Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Gough, H G & Heilbrun, A B Jr (1983). *The Adjective Check List manual*. Palo Alto, CA: Consulting Psychologists Press.

Gould, S J (1981). *The mismeasure of man*. New York, NY: WW Norton & Company.

Graham, J R (1993). *MMPI–2: Assessing personality and psychopathology*. New York, NY: Oxford.

Greathouse, D & Shaughnessy, M F (2016). An interview with Amy Gabel about the WISC-V, *Journal of Psychoeducational Assessment*, *34*(8), 800–10.

Green, A, Garrick, T, Sheedy, D, Blake, H, Shores, A & Harper, C (2008) Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Preliminary Australian normative data. *Australian Journal of Psychology*, *60*(2), 72–9.

Greenberg, J (1987). A taxonomy of organizational justice theories. *Academy of Management Review*, *12*, 9–22.

Greenberg, S & Shuman, D (1997). Irreconcilable conflict between therapeutic and forensic roles. *Professional Psychology: Research and Practice*, *28*, 50–7.

Gregg, N (1989). Review of the learning style inventory. In J C Conoley & J J Kramer (Eds.), *The Tenth Mental Measurements Yearbook* (pp. 441–2). Lincoln, NE: The Buros Institute of Mental Measurements.

Gregory, R (1999). *Foundations of intellectual assessment*. Boston, MA: Allyn & Bacon.

Grilo, C M, Fehon, D C, Walker, M & Martino, S (1996). A comparison of adolescent inpatients with and without substance abuse using the Millon Adolescent Clinical Inventory. *Journal of Youth and Adolescence*, *25*(3), 379–88.

Grilo, C M, Sanislow, C A, Fehon, D C, Martino, S & McGlashan, T H (1999). Psychological and behavioral functioning in adolescent psychiatric inpatients who report histories of childhood abuse. *American Journal of Psychiatry*, *156*(4), 538–43.

Groth-Marnat, G (2000a). Introduction to neuropsychological assessment. In G Groth-Marnat (Ed.), *Neuropsychological assessment in clinical practice: A guide to test interpretation and integration* (pp. 3–25). New York, NY: Wiley.

Groth-Marnat, G (2000b). Visions of clinical assessment: Then, now, and a brief history of the future. *Journal of Clinical Psychology*, *56*, 349–85.

Groth-Marnat, G (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Groth-Marnat, G & Wright, A J (2016). *Handbook of psychological assessment* (6th ed.). Hoboken, NJ: Wiley.

Grove, W M & Meehl, P (1996). Comparative efficiency of informal and formal prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy and Law*, *2*, 293–323.

Grove, W M, Zald, D H, Lebow, B S, Snitz, B E & Nelson, C (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30.

Gudjonsson, G H & Haward, L R C (1998). *Forensic psychology: A guide to practice*. London, UK: Routledge.

Guilford, J P (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Guilford, J P (1985). The structure-of-intellect model. In B B Wolman (Ed.), *Handbook of intelligence* (pp. 225–66). New York, NY: Wiley.

Guilford, J P (1988). Some changes in the structure of intellect model. *Educational and Psychological Measurement, 48*, 1–4.

Gulliksen, H (1950). *Theory of mental tests*. New York, NY: Wiley.

Gustafsson, J E (1989). Broad and narrow abilities in research on learning and instruction. In R Kanfer, P L Ackerman & R Cudek (Eds.), *Abilities, motivation and methodology: The Minnesota Symposium on Learning and Individual Differences* (pp. 203–32). Hillsdale, NJ: Erlbaum.

Guttman, L (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–50.

Hall, C S & Lindzey, G (1978). *Theories of personality* (3rd ed.). New York, NY: Wiley.

Handy, C (1994). *The age of paradox*. Boston, MA: Harvard Business School Press.

Hanes, K R, Andrewes, D G, Smith, D J & Pantelis, C (1996). A brief assessment of executive central dysfunction: Discriminant validity and homogeneity of planning, set shift and fluency measures. *Archives of Clinical Neuropsychology, 18*, 185–91.

Hanson, R K (2005). Twenty years of progress in violence risk assessment. *Journal of Interpersonal Violence, 20*(2), 212–17.

Hanson, R K & Morton-Bourgon, K E (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*(1), 1–21.

Hanson, R K & Thornton, D (1999). *Static 99: Improving actuarial risk assessments for sex offenders, 1999–02*. Ottawa, Canada: Department of the Solicitor General of Canada.

Harcourt Educational Measurement (2003). *OLSAT: Otis-Lennon School Ability Test technical manual*. San Antonio, TX: Author.

Hare, R D (1998). The Hare PCL–R: Some issues concerning its use and misuse. *Legal and Criminological Psychology, 3*, 101–22.

Hare, R D (2003). *Manual for the Hare Psychopathy Checklist–Revised Second Edition*. Toronto, Ontario: Multi-Health Systems.

Hassed, C S (2000). Depression: Dispirited or spiritually deprived. *Medical Journal of Australia, 173*, 545–7.

Hathaway, S R & McKinley, J C (1943). *The Minnesota Multiphasic Personality Inventory manual*. Minneapolis, MN: Minnesota Press.

Hathaway, S R & McKinley, J C (1951). *The Minnesota Multiphasic Personality Inventory Manual (Revised)*. New York, NY: Psychological Corporation.

Hattie, J (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–64.

Hebben, N & Milberg, W (2009). *Essentials of neuropsychological assessment* (2nd ed.). New York, NY: Wiley.

Heckman, J J & Kautz, T (2012). Hard evidence on soft skills. *Labour Economics, 19*(4), 451–64.

Heilbrun, K (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16*, 257–72.

Heilbrun, K (2001). *Principles of forensic mental health assessment*. New York, NY: Kluwer Academic.

Heilbrun, K, Bank, S, Follingstad, D & Frederick, R (2000). *Petition for forensic psychology as an APA specialization*. Presented to the Committee for the Recognition of Specialties and Proficiencies in

Professional Psychology. Washington, DC: American Psychological Association.

Heilbrun, K, Roger, R & Otto, R K (2002). Forensic assessment: Current status and future directions. In J R P Ogloff (Ed.), *Psychology and law: Reviewing the discipline* (pp. 119–46). New York, NY: Kluwer Academic.

Heilman, K M & Valenstein, E (2012). *Clinical neuropsychology* (5th ed.). New York, NY: Oxford University Press.

Henson, R K & Roberts, J K (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*, 393–416.

Heritage, M & Heritage, J (2013). Teacher questioning: The epicenter of instruction and assessment. *Applied Measurement in Education*, *26*, 176–90.

Herriot, P (2002). Selection and self: Selection as a social process. *European Journal of Work and Organizational Psychology*, *11*, 385–402.

Herrmann, C (1997). International experience with the Hospital Anxiety and Depression Scale: A review of validation data and clinical results. *Journal of Psychosomatic Research*, *42*, 17–41.

Herrnstein, R J & Murray, C (1994). *The bell curve*. New York, NY: The Free Press.

Hersen, M (Ed.) (2004). *Comprehensive handbook of psychological assessment (Vols 1–4),* New York, NY: Wiley.

Hersen, M & Thomas, J C (2007). *Handbook of clinical interviewing with adults*. Thousand Oaks, CA: Sage Publications.

Hesketh, B & Neal, A (1999). Technology and performance. In D R Ilgen & E D Pulakos (Eds.), *The changing nature of performance:*

Implications for staffing, motivation and development (pp. 21–55). San Francisco, CA: Jossey-Bass.

Hiatt, M D & Cornell, D G (1999). Concurrent validity of the Millon Adolescent Clinical Inventory as a measure of depression in hospitalized adolescents. *Journal of Personality Assessment, 73*(1), 64–79.

Hibbard, S (2003). A critique of Lilienfeld et al's (2000). 'The scientific status of projective techniques'. *Journal of Personality Assessment, 80*(3), 260–71.

Hoge, R D (2005). Youth level of service/case management inventory. In T Grisso, G Vincent & D Seagrave (Eds.), *Mental health screening and assessment in juvenile justice* (pp. 283–94). New York, NY: Guilford Press.

Holdnack, J A, Lissner, D, Bowden, S C & McCarthy, K A L (2004). Utilising the WAIS-III /WMS-III in clinical practice: Update of research and issues relevant to Australian normative research. *Australian Psychologist, 39*, 220–7.

Holland, J L (1958). A personality inventory employing occupational titles. *Journal of Applied Psychology, 42*, 336–42.

Holland, J L (1992). *Making vocational choices: A theory of vocational personalities and work environments* (2nd ed.). Odessa, FL: Psychological Assessment Resources.

Holland, J L & Rayman, J R (1986). The Self-Directed Search. In W B Walsh & S H Osipow (Eds.), *Advances in vocational psychology: The assessment of interests*. Hillsdale, NJ: Erlbaum.

Holland, J L, Shears, M & Harvey-Beavis, A (2012). *Self-directed search—2012 update edition: Form R, Second Australian Edition*. Melbourne, Vic: ACER.

Holt, R R (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology, 56*, 1–12.

Hooper, H E (1948). A study in the construction and preliminary standardization of a visual organization test for use in the measurement of organic deterioration. Unpublished master's thesis, University of Southern California.

Hooper, H E (1958). *The Hooper Visual Organization Test manual*. Los Angeles: Western Psychological Services.

Hooper, H E (1983). *The Hooper Visual Organization Test manual*. Los Angeles, CA: Western Psychological Services.

Horn, J L (1998). A basis for research on age differences in cognitive capabilities. In J J McArdle & R Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 57–91). Chicago, IL: Riverside Press.

Horowitz, L M (2004). *Interpersonal foundations of psychopathology*. Washington, DC: American Psychological Association.

Howe, M A (1975). General aptitude test battery—An Australian empirical study. *Australian Psychologist, 10,* 32–44.

Howell, D C (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Hsiao, J K, Bartko, J J & Potter, W Z (1989). Diagnosing diagnoses. Receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry, 46,* 664–7.

Hulin, C L, Drasgow, F & Parsons, C K (1983). *Item response theory*. Homewood, IL: Dow Jones-Irwin.

Hunsley, J & Meyer, G J (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446–55.

International Test Commission (2005). *International guidelines on computer-based and internet delivered testing*. Belgium: International Test Commission.

Irvine, S H & Kyllonen, P C (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

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

Ivanova, M Y, Achenbach, T M, Rescorla, et al. (2007). The generalizability of teacher's report form syndromes in 20 cultures. *School Psychology Review*, *36*, 468–83.

Jackson, D N (1970). A sequential system of personality scale development. In C Speilberger (Ed.), *Current topics in clinical and community psychology* (Vol. 2). New York, NY: Academic Press.

Jackson, D N (1971). The dynamics of structured personality tests. *Psychological Review*, *78*, 229–48.

Jackson, D N (1984). *Personality research form manual*. Port Huron, MI: Research Psychologists Press.

Jacobs, K, Watt, D & Roodenburg, J (2013). Why can't Jonny read? Bringing theory into cognitive assessment. *Inpsych*, December. Downloaded 28 January 2016 from www.psychology.org.au/inpsych/2013/december/jacobs/.

Jeanneret, R & Silzer, R (1998). *Individual psychological assessment: Predicting behavior in organizational settings*. San Francisco, CA: Jossey-Bass.

Jensen, A R (1980). *Bias in mental testing*. New York, NY: Free Press.

Jensen, A R (2004). Obituary-John Bissell Carroll. *Intelligence*, *32*, 1–5.

John, O P & Srivastava, S (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L A Pervin & P P John (Eds.), *Handbook of personality: Theory and research* (2nd ed.; pp. 102–38). New York, NY: Guilford Press.

Joiner, R, Gavin, J, Brosnan, M, Cromby, J, Gregory, H, Guiller, J & Moon, A (2013). Comparing first and second generation digital natives' internet use, internet anxiety, and internet identification. *Cyberpsychology, Behavior, and Social Networking*, *16*, 549–52.

Jonason, P K & Webster, D G (2010). The Dirty Dozen: A concise measure of the Dark Triad. *Psychological Assessment*, *22*(2), 420–32.

Jones, L V & Thissen, D (2007). A history and overview of psychometrics. In C R Rao & S Sinharay (Eds.), *Psychometrics* (pp. 1–28). Amsterdam, Netherlands: Elsevier.

Jöreskog, K G (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.

Joseph, D L & Newman, D A (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology*, *95*, 54–78.

Judge, T A, Boudreau, J W & Bretz, R D (1994). Job and life attitudes of male executives. *Journal of Applied Psychology*, 79, 767–82.

Kagan, J (1994). *Galen's prophecy: Temperament in human affairs*. New York, NY: Basic Books.

Kamphaus, R W, Petoskey, M D & Rowe, E W (2000). Current trends in psychological testing of children. *Professional Psychology: Research and Practice*, *31*, 155–64.

Kanaya, T, Scullin, M H & Ceci, S J (2003). The Flynn effect and US policies: The impact of rising IQ scores on American society via mental retardation diagnosis. *American Psychologist*, *58*, 778–90.

Kaufman, A S & Kaufman, N L (1983). *Kaufman Assessment Battery for Children*. Circle Pines, MN: American Guidance Service.

Kaufman, A S & Kaufman, N L (2004). *Kaufman Assessment Battery for Children* (2nd Ed.). Circle Pines, MN: American Guidance Service.

Kaufman, A S & Lichtenberger, E O (1999). *Essentials of WAIS-III assessment*. New York, NY: Wiley.

Kearins, J M (1981). Visual spatial memory in Australian Aboriginal children of desert regions. *Cognitive Psychology*, *13*, 434–60.

Kearney, G E, de Lacey, P R & Davidson, G R (Eds.). (1973). *The psychology of Aboriginal Australians: A book of readings*. Sydney, NSW: Wiley.

Keats, D M & Keats, J A (1988). Human assessment in Australia. In S H Irvine & J W Berry (Eds.), *Human abilities in cultural context*. Cambridge, UK: Cambridge University Press.

Keith, T Z & Reynolds, M R (2010). Cattell-Horn-Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, *47*, 635–50.

Kelly, G A (1955). *The psychology of personal constructs*. New York, NY: Norton.

Kendall, I, Jenkinson, J, de Lemos, M & Clancy, D (1997). *Supplement to guidelines for the use of psychological tests*. Melbourne, Vic: Australian Psychological Society.

Kertesz, A (2007). *Western Aphasia Battery–Revised*. San Antonio, TX: Pearson.

Kessler, R C, Andrews, G, Colpe, L J, et al. (2002). Short screening scales to monitor population, prevalences, and trends in non-specific psychological distress. *Psychological Medicine*, *32*, 959–76.

Kiesler, B J (1996). *Contemporary interpersonal theory and research: Personality, psychopathology, and psychotherapy*. New York, NU: Wiley.

Klassen, R M, Neufeld, P & Munro, F (2005). When IQ is irrelevant to the definition of learning disabilities: Australian school psychologists' beliefs and practice. *School Psychology International*, *26*(3), 297–316.

Klenowski, V & Wyatt-Smith, C (2011). The impact of high stakes testing: the Australian story. *Assessment in Education : Principles, Policy*

and Practice.

Klerman, G L & Weissman, M M (1993). *New applications of personal therapy*. Washington, DC: American Psychiatric Association.

Kliegel, M, Jager, T, Altgassen, M & Shum, D (2008). Clinical neuropsychology of prospective memory. In M Kligel, M A McDaniel & G O Einstein (Eds.), *Prospective memory: Cognitive, neuroscience, developmental, and applied perspective* (pp. 283–308). Mahwah, NJ: Erlbaum.

Kliegel, M, McDaniel, M A & Einstein G O (Eds.). (2008). *Prospective memory: Cognitive, neuroscience, developmental, and applied perspective*. Mahwah, NJ: Erlbaum.

Kline, P (1993). *The handbook of psychological testing*. London: Routledge.

Klich, L Z (1988). Aboriginal cognitive and psychological nescience. In S H Irvine & J W Berry (Eds.), *Abilities in cultural context* (pp. 427–52). Cambridge, NY: Cambridge University Press.

Knight, R G & Godfrey, H P D (1984). Tests recommended by New Zealand hospital psychologists. *New Zealand Journal of Psychology*, *13*, 32–6.

Kohlberg, L (1981). *The philosophy of moral development*. New York, NY: Harper & Row.

Kolb, B & Whishaw, I Q (2009). *Fundamentals of human neuropsychology* (6th ed.). New York, NY: Worth.

Kolb, B & Whishaw, I Q (2015). *Fundamentals of human neuropsychology* (7th ed.). New York, NY: Worth Publishers Inc.

Kristof-Brown, A L, Zimmerman, R D & Johnson, E C (2005). Consequences of individuals' fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. *Personnel Psychology*, *58*, 281–342.

Kristoff, A. (1996). Person–organization fit: An integrative review of its conceptualizations, measurement and implications. *Personnel*

Psychology, *49*, 1–49.

Kuder, G F & Richardson, M W (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151–60.

Kyllonen, P C (1997). Smart testing. In R F Dillon (Ed.), *Handbook on testing*. Westport, CT: Greenwood Press.

Landy, F J & Conte, J M (2007). *Work in the 21st century*. Malden, MA: Blackwell.

Landy, F J & Farr, J L (1980). Performance appraisal. *Psychological Bulletin*, *87*, 72–107.

Larson, E B, Kirschner, K, Bode, R, Heinemann, A & Goodman, R (2005). Construct and predictive validity of the Repeatable Battery for the Assessment of Neuropsychological Status in the evaluation of stroke patients. *Journal of Clinical and Experimental Neuropsychology*, *27*, 16–32.

Latham, G & Wexley, K (1977). Behavioral observation scales. *Journal of Applied Psychology*, *30*, 255–68.

Leivens, F & Harris, M M (2003). Research on internet recruiting and testing: Current status and future directions. In C L Cooper & I T Robinson (Eds.), *International review of industrial and organizational psychology (Vol. 18)*. West Sussex, UK: Wiley.

Lezak, M D (1995). *Neuropsychological assessment* (3rd ed.). New York, NY: Oxford University Press.

Lezak, M D, Howieson, D B, Bigler, E D & Tranel, D (2012). *Neuropsychological assessment* (5th ed.). New York, NY: Oxford University Press.

Lilienfeld, S O, Wood, J M & Garb, H N (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest*, *1*, 27–66.

Linacre, J M (2001). *A user's guide to WINSTEPS/MINISTEPS*. Chicago, IL: Winsteps.com,

Lipsitt, P D, Lelos, D & McGarry, L (1971). Competency for trial: A screening instrument. *American Journal of Psychiatry, 128*, 104–9.

Loo, R (1999). Confirmatory Factor analyses of Kolb's Learning Style Inventory (LSI-1985). *British Journal of Educational Psychology, 69*, 213–19.

Lord, F M & Novick, M R (1968). *Statistical theories of mental test scores*. Reading MA: Addison Wesley.

Lovibond, S H & Lovibond, P F (1995a). *Manual for the Depression Anxiety Stress Scales*. Sydney, NSW: Psychology Foundation.

Lovibond, S H & Lovibond, P F (1995b). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behavior Research and Therapy, 33*, 335–43.

Lowenkamp, C T, Latessa, E J & Holsinger, A M (2006). The risk principle in action: What have we learned from 13,676 offenders and 97 correctional programs? *Crime & Delinquency, 52*(1), 77–93.

Lubinski, D (2000). Scientific and social significance of assessing individual differences: 'Sinking shafts at a few critical points'. *Annual Review of Psychology, 51*, 405–44.

Lucas, R E & Diener, E (2008). Subjective emotional well-being. In M Lewis, J M Haviland-Jones & L F Barrett (Eds.), *Handbook of emotions* (3rd ed.; pp. 471–84). New York, NY: Guilford.

Luce, R D & Tukey, J W (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.

Luria, A R (1966). *Human brain and psychological proc*esses. New York, NY: Harper & Row.

Luria, A R (1973). *The working brain: An introduction to neuropsychology*. New York, NY: Basic Books.

Lyon, G R (1996). Learning disabilities. *Special Education for Students with Disabilities*, *6*(1), 54–76.

MacCorquodale, K & Meehl, P E (1954). Edward C Tolman. In W K Estes, S Koch, K MacCorquodale, P E Meehl, C G Mueller Jr, W H Schoenfeld & W S Verplanck (Eds.), *Modern learning theory: A crucial analysis of five examples*. New York, NY: Appleton-Century-Crofts.

Maddox, T (Ed.) (2008). *Tests: A comprehensive reference for assessment in psychology, education, and business*. Austin, TX: ProEd.

Makransky, G, Bonde, M T, Wulff, J S, Wandall, J, Hood, M, Creed, P A, Bache, I, Silahtaroglu, A & Nørremølle, A (2016). Simulation based virtual learning environment in medical genetics counseling: An example of bridging the gap between theory and practice in medical education. *BMC Medical Education*, *16*, 1.

Maloney, M P & Ward, M P (1976). *Psychological assessment: A conceptual approach*. New York, NY: Oxford University Press.

Mapou, R L (1995). A cognitive framework for neuropsychological assessment. In R L Mapou & J Spector (Eds.), *Clinical neuropsychological assessment: A cognitive approach* (pp. 295–337). New York, NY: Plenum.

Markon, K, E, Krueger, R F & Watson, D (2005). Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. *Journal of Personality and Social Psychology*, *88*(1), 139–57.

Markowitsch, H J & Piefke, M (2010). The functional neuroanatomy of learning and memory. In J Gurd, U Kischka & J Marshall (Eds.), *The handbook of clinical neuropsychology*. New York, NY: Oxford University Press.

Marks, G (2007). *Completing university: Characteristics and outcomes of completing and non-completing students*. Longitudinal Surveys of Australian Youth (LSAY). Australian Council for Educational Research: Melbourne.

Martin, M, Allan, A & Allan, M M (2001). The use of psychological tests by Australian psychologists who do assessments for the courts. *Australian Journal of Psychology*, *53*, 77–82.

Maslow, A H & Mittelmann, B (1941). *Principles of abnormal psychology: The dynamics of psychic illness*. New York, NY: Harper.

Matarazzo, J D (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Baltimore, MD: Williams & Wilkins.

Matarazzo, J D (1980). *Wechsler's measurement and appraisal of adult intelligence*. New York, NY: Oxford University Press.

Matarazzo, J D (1986). Computerized clinical psychological test interpretations: Unvalidated plus all mean and no sigma. *American Psychologist*, *41*, 14–24.

Matarazzo, J D (1990). Psychological assessment versus psychological testing: Validation from Binet to the school, clinic and courtroom. *American Psychologist*, *45*, 999–1017.

Matthews, G, Zeidner, M & Roberts, R (2002). *Emotional intelligence: Science and myth*. Cambridge, MA: MIT Press.

McAdams, D P (2008). Personal narratives and the life story. In O John, R Robins & L Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed.; pp. 241–61). New York, NY: Guilford Press.

McAdams, D P & Olson, B D (2010). Personality development: Continuity and change over the life course. *Annual Review of Psychology*, *61*, 517–42.

McAdams, D P & Pals, J L (2006). A new Big Five: Fundamental principles for an integrative science of personality. *American Psychologist*, *61*(3), 204–17.

McCann, J T (1999). *Assessing adolescents with the MACI: Using the Millon Adolescent Clinical Inventory*. New York, NY: Wiley.

McCarthy, R A & Warrington, E K (1990). *Cognitive neuropsychology: A clinical introduction*. San Diego, CA: Academic Press.

McClelland, D & Burnham, D (2003). Power is the great motivator. *Harvard Business Review*, January. Retrieved from https://hbr.org/2003/01/power-is-the-great-motivator.

McCormick, E J, Mecham, R C & Jeanneret, P R (1977). *Technical manual for the Position Analysis Questionnaire*. Washington, DC: PAQ Services.

McCrae, R R (2009). The Five-Factor Model of personality traits: Consensus and controversy. In P J Corr & G Matthews (Eds.), *The Cambridge handbook of personality psychology* (pp. 148–61). Cambridge, UK: Cambridge University Press.

McCrae, R R & Costa, P T (1999). A five-factor theory of personality. In L A Pervin & O P John (Eds.), *Handbook of personality: Theory and research* (2nd ed.; pp. 139–53). New York, NY: Guilford.

McCrae, R R & Costa, P T (2008). The five-factor theory of personality. In P P John, R W Robins & L A Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed.; pp. 159–81). New York, NY: Guilford.

McCullough, M E, Emmons, R A & Tsang, J (2002). The grateful disposition: A conceptual and empirical topography. *Journal of Personality and Social Psychology*, *82*(1), 112–27.

McDonald, R (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McElwain, D W & Kearney, G E (1970). *Queensland Test handbook*. Melbourne, Vic: ACER.

McElwain, D W & Kearney, G E (1973). Intellectual development. In G E Keaney, P R De Lacey & G R Davidson (Eds.), *The psychology of Aboriginal Australians* (pp. 43–56). Sydney, NSW: Wiley.

McGrath, R E (2011). *Quantitative models in psychology*. Washington, DC: American Psychological Association.

McGraw, K O & Wong, S P (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

McGrew, K S (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf–Gc framework. In D P Flanagan, J L Genshaft & P L Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151–79). New York, NY: Guilford.

McGrew, K S (2008). The Australian standardization of the Woodcock Johnson III Cognitive and Achievement Battery. Paper presented at the Conference of the Australian Psychological Society.

McGrew, K S & Flanagan, D P (1995). A 'cross-battery' approach to intelligence test interpretation. *Communique, 24*, 28–9.

McGrew, K S & Flanagan, D P (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment*. Needham Heights, MA: Allyn & Bacon.

McKay, C, Wertheimer, J C, Fichtenberg, N L & Casey, JE (2008). The Repeatable Battery for The Assessment of Neuropsychological Status (RBANS): Clinical Utility in a Traumatic Brain Injury Sample. *Clinical Neuropsychologist, 22*(2), 228–41. doi: 10.1080/13854040701260370.

McKenzie, A (1980). Are ability tests up to standard? *Australian Psychologist, 15*, 335–50.

Mead, A D & Drasgow, F (1993). Equivalence of computerised and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–58.

Meadows, M & Billington, L (2005). *A review of the literature on marking reliability*. Manchester, UK: AQA. Retrieved from https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_MM_01052005.pdf.

Medical Research Council and Department of Scientific Research (1920). *First annual report of the Industrial Fatigue Research Board*. Retrieved from http://scans.library.utoronto.ca/pdf/7/35/report1919grea/report1919grea.pdf.

Meehl, P E (1954*). Clinical versus statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Meehl, P (1956). Wanted—A good cookbook. *American Psychologist*, 266.

Meeks, L, Kemp, C & Stephenson, J (2014). Standards in literacy and numeracy: Contributing factors. *Australian Journal of Teacher Education, 39*(7), 106–39.

Meier, M J (1997). The establishment of clinical neuropsychology as a psychological specialty. In M E Maruish & J A Moses, Jr (Eds.), *Clinical neuropsychology: Theoretical foundations for practitioners* (pp. 1–31). Mahwah, NJ: Lawrence Erlbaum.

Melton, G B, Petrila, J, Poythress, N & Slobogin, C (1997). *Psychological evaluations for the courts: A handbook for attorneys and the mental health professionals* (2nd ed.). New York, NY: Guilford.

Meteyard, J D & Gilmore, L (2015). Psycho-educational assessment of specific learning disabilities: Views and practices of Australian psychologists and guidance counsellors. *Journal of Psychologists and Counsellors in Schools*, *25*, 1–12.

Meyer, G J, Finn, S E, Eyd, L D, Kay, G G, Moreland, K L, Dies, R R, et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–65.

Meyer, J & Allen, N (1997). *Commitment in the workplace*. Thousand Oaks, CA: Sage.

Michell, J (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Michell, J (2004). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Michell, J (2009). The psychometrians' fallacy: Too clever by half. *British Journal of Mathematical and Statistical Psychology, 62*, 41–55.

Mihura, J L, Meyer, G J, Dumitrascu, N & Bombel, G (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the Comprehensive System. *Psychological Bulletin, 139*, 548–605.

Millon, T (1993). *Manual for the Millon Adolescent Clinical Inventory (MACI)*. Minneapolis, MN: National Computer Systems Assessments.

Milner, A D (1998). *Comparative neuropsychology*. New York, NY: Oxford University Press.

Mirsky, A F, Anthony, B J, Duncan, C C, Ahearn, M B & Kellam, S G (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review, 2*, 109–45.

Mirsky, A F, Fantie, B D & Tatman, J E (1995). Assessment of attention across the lifespan. In R L Mapou & J Spector (Eds.), *Clinical neuropsychological assessment: A cognitive approach. Critical issues in neuropsychology* (pp. 17–48). New York, NY: Plenum Press.

Mischel, M (1968). *Personality and assessment*. New York, NY: Wiley.

Mischel, W (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*, 252–83.

Mischel, W & Shoda, Y (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246–68.

Mislevy, R (2012). Some thoughts on terminology. In *Work in progress, Issue No 2. The Gordon Commission into the Future of Assessment*

in Education. Downloaded from www.gordoncommission.org

Mitrushina, M, Boone, K B, Razani, J & D'Elia, L F (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York, NY: Oxford University Press.

Money, J (1976). *A standardized road-map test of direction sense: Manual*. San Rafael, CA: Academic Therapy Publications.

Monte, C F & Sollod, R N (2003). *Beneath the mask* (7th ed.). New York, NY: Wiley.

Morey, L C (2003). *Essentials of PAI assessment*. New York, NY: Wiley.

Morey, L C (2007). *Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Mowday, R T, Porter, L W & Steers, R M (1982). *Employee-organization linkages: The psychology of commitment, absenteeism, and turnover*. New York, NY: Academic Press.

MRCDSC – see Medical Research Council and Department of Scientific Research

Munsterberg, H (1913). *Psychology and industrial efficiency*. New York, NY: Houghton Mifflin.

Murray, G, Judd, F, Jackson, H, Fraser, C, Komiti, C, Pattison, P & Robins, G (2009). Personality for free: Psychometric properties of a public domain Australian measure of the five-factor model. *Australian Journal of Psychology*, *61*(3), 167–74.

Naglieri, J A & Das, J P (1997). *Das-Naglieri cognitive assessment system*. Itasca, IL: Riverside Publishing.

Naglieri, J A, Das, J P & Goldstein, S (2012). Planning, attention, 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.). New York, NY: Guilford.

Naglieri, J A, Drasgow, F, Schmitt, M, Handler, L, Prifitera, A, Margolis, A & Valasquez, R (2004). Psychological testing on the internet: New problems, old issues. *American Psychologist, 59*, 150–62.

Nathan, P E & Langenbucher, J (2003). Diagnosis and classification. In G Stricker & T A Widiger (Eds.), *Handbook of Psychology: Vol. 8: Clinical psychology*. New York, NY: Wiley.

Nauta, M M (2010). The development, evolution, and status of Holland's theory of vocational personalities: Reflections and future directions for counseling psychology. *Journal of Counseling Psychology, 57*, 11–22.

Neal, A & Griffin, M A (1999). Developing a model of individual performance for human resource management. *Asia Pacific Journal of Human Resources, 37*, 44–59.

Neisser, U, Boodoo, G, Bouchard Jr, T J, Boykin, A W, Brody, N, Ceci, S J & Urbina, S (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Nelson, H E & Willison, J (1991). *The National Adult Reading Test (NART): Test manual* (2nd ed.). Windsor, UK: NFER, Nelson.

Nunnally, J C (1967). *Psychometric theory*. New York, NY: McGraw Hill.

Oakland, T, Douglas, S & Kane, H (2016). Top ten standardized tests used internationally with children and youth by school psychologists in 64 countries: A 24-year follow-up study. *Journal of Psychoeducational Assessment, 34*(2), 166–76.

O'Boyle, E H Jr & McDaniel, M A (2009). Criticisms of employment testing: a commentary. In R P Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 181–97). Washington, DC: American Psychological Association.

O'Connor, G (2002). *ACER Short Clerical Test: Manual*. Melbourne, Vic: ACER Press.

Office of Strategic Services Assessment Staff (1948). *Assessment of men: Selection of personnel for the Office of Strategic Service*. New York, NY: Rinehart.

Ogloff, J R & Douglas, K S (2003). Psychological assessment in forensic settings. In J R Graham & J A Naglieri (Eds.), *Handbook of psychology* (pp. 345–63). New York, NY: John Wiley & Sons.

Ogloff, J R P & Davis, M R (2004). Advances in offender assessment and rehabilitation: Contributions of the risk-needs-responsivity approach. *Psychology, Crime and Law*, *10*(3), 229–42.

O'Gorman, J (2007). *Psychology as a profession in Australia*. Bowen Hills: Australian Academic Press.

O'Gorman, J G & Shum, D H K (2012).The Australian Remote Memory Battery, 10 years on. *Australian Psychologist*, *47*, 14–19.

Oliveira-Souza, R, Moll, J & Eslinger, P J (2004). Neuropsychological assessment. In M Rizzo & P J Eslinger (Eds.), *Principles and practice of behavioural neurology and neuropsychology* (pp. 47–64). Philadelphia, PA: WB Saunders.

O'Neil, W M (1987). *A century of psychology in Australia*. Sydney, NSW: Sydney University Press.

Ord, I G (1977). Australian psychology and Australia's neighbours. In M Nixon & R Taft (Eds.), *Psychology in Australia: Achievements and prospects*. Sydney, NSW: Pergamon.

O'Reilly, C III & Chatman, J (1986). Organizational commitment and psychological attachment: The effects of compliance, identification, and internalization on pro-social behavior. *Journal of Applied Psychology*, *71*, 492–9.

Osterlind, S (2005). *Modern psychometrics: Development and application of modern mental measures*. New York, NY: Prentice Hall.

Otis, A S & Lennon, R T (2003). *Otis–Lennon School Ability Test* (8th ed.). San Antonio, TX: Harcourt Assessment.

Ownby, R L (1997). *Psychological reports: A guide to report writing in professional psychology* (3rd ed.). New York, NY: Wiley.

Ozer, D J (1985). Correlation and the coefficient of determination. *Psychological Bulletin*, *97*, 307–15.

Pardo, J V, Pardo, P J, Janer, K W & Raichle, M E (1990). The anterior cingulated cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences*, *87*, 256–9.

Parsons, M W & Hammeke, T A (Eds.) (2014). *Clinical neuropsychology: A pocket handbook for assessment* (3rd ed.). Washington, DC: American Psychological Association.Paulhus, D L & Trapnell, P D (2008). Self-presentation: An agency-communion framework. In O P John, R W Robins & L A Pervin (Eds.), *Handbook of personality psychology* (pp. 492–517). New York, NY: Guilford.

Paulhus, D L & Williams, K M (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*, 556–63.

Paunonen, S & Jackson, D N (2000). What is beyond the Big Five? Plenty! *Journal of Personality*, *65*(8), 821–35.

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

Penrose, L S & Raven, J C (1936). A new series of perceptual tests: Preliminary communication. *British Journal of Medical Psychology*, *16*, 97–104.

Phongsavan, P, Chey, T, Bauman, A, Brooks, R & Silove, D (2006). Social capital, socio-economic status and psychological distress among Australian adults. *Social Science & Medicine*, *63*, 2546–61.

Posner, J, Russell, J A & Peterson, B S (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Developmental Psychopathology*, *17*(3), 715–34.

Posner, M I & Petersen, S E (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25–42.

Powell, M & Lancaster, S (2003). Guidelines for interviewing children during child custody evaluations. *Australian Psychologist, 38*, 46–54.

Poythress, N, Monahan, J, Otto, R, Edens, J, Bonnie, R, Monahan, J & Hoge, S K (1999). *MacArthur Competence Assessment Tool—Criminal Adjudication*. Odessa, FL: Psychological Assessment Resources.

Prediger, D J (1976). A world-of-work map for career exploration. *Vocational Guidance Quarterly*, *24*, 198–208.

Prediger, D J & Vansickle, T R (1992). Locating occupations on Holland's hexagon: Beyond RIASEC. *Journal of Vocational Behavior*, *40*, 111–28.

Prentky, R & Righthand, S (2003). *Juvenile Sex Offender Assessment Protocol–II (J-SOAP-II): Manual*. Bridgewater, MA.

Preston, C C & Colman, A M (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1–15.

Psychology Board of Australia (2016a). *General registration*. Retrieved from www.psychologyboard.gov.au/Registration/National-psychology-exam/Content-of-the-examination.aspx.

Psychology Board of Australia (2016b). *National psychology examination curriculum*. Retrieved from www.psychologyboard.gov.au/Registration/General.aspx.

Purdue Research Foundation (1948). *Purdue Pegboard Test*. Lafayette, IN: Lafayette Instrument Company.

Quinsey, V L, Harris, G T, Rice, M E & Courmier, L A (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association Press.

Rabin, L A, Barr, W B & Burton, L A (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, *20*, 33–65.

Ramsden, P (2003). Review of the Delis-Kaplan Executive Function System. In J Impara & B Plake (Eds.), *Mental Measurements Yearbook* (pp. 284–6). Lincoln, NE: Buros Institute of Mental Measurement.

Randolph, C (1998). *Repeatable battery for the assessment of neuropsychological status manual*. San Antonio, TX: The Psychological Corporation.

Randolph, C, Tierney, M C, Mohr, E & Chase, T N (1998). The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Preliminary clinical validity. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 310–19.

Randolph, J J, Hicks, T, Mason, D & Cuneo, D J (1982). The competency screening test: A validation study in Cook County, Illinois. *Criminal Justice and Behavior*, *9*, 495–500.

Rapaport, D, Gill, M & Schafer, R (1946). *Diagnostic psychological testing, Volume II*. Chicago, IL: The Year Book Publishers.

Rasch Analyst (1996). *Welcome to RUMM: A Windows based item analysis program employing Rasch unidimension models*. Author.

Raven, J C (1938). *Progressive matrices: A perceptual test of intelligence*. London, UK: HK Lewis.

Raven, J C (1939). *Progressive Matrices: A perceptual test of intelligence*. London, UK: HK Lewis.

Raykov, T & Marcoulides, G (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Reddy, P, Knowles, A & Reddy, S (1995). Language issues in cross-cultural testing. *Australian Psychologist*, *30*(1), 27–9.

Reitan, R M & Wolfson, D (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychology Press.

Revelle, W (in preparation). *An introduction to psychometric theory with applications in R*. Retrieved from www.personality-project.org/r/book.

Rey, A (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Reynolds, C R & Livingston, R B (2014). *Mastering modern psychological testing: Theory and methods*. Boston, MA: Pearson Education.

Richardson, J T E (2003). Howard Andrew Knox and the origins of performance testing on Ellis Island, 1911–1916. *History of Psychology*, *6*, 143–70.

Rickwood, D, Dudgeon, P & Gridley, H (2010). A history of psychology in Aboriginal and Torres Strait Islander mental health. In N Purdie, P Dudgeon & R Walker (Eds.), *Working together: Aboriginal and Torres Strait Islander mental health and wellbeing principles and practice* (pp. 13–24). Canberra, Australia: Department of Health and Ageing. Available at http://aboriginal.childhealthresearch.org.au/media/54847/working_together_full_book.pdf.

Robertson, I H, Ward, T, Ridgeway, V & Nimmo-Smith, I (1994). *Test of everyday attention*. San Antonio, TX: The Psychological Corporation.

Robinson-Zañartu, C & Carlson, J (2013). Dynamic assessment. In K F Geisinger (Ed.), *APA Handbook of testing and assessment in psychology* (pp. 149–67). Washington, DC: American Psychological Association.

Rodriguez, M C (1997). *The art and science of item writing: A meta-analysis of multiple choice item formats*. Paper presented at the

annual meeting of the American Educational Research Association, Chicago, IL, April.

Rodriguez, M C (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: issues and Practice*, Summer, 3–13.

Rogers, R (Ed.) (1997). *Clinical assessment of malingering and deception* (2nd ed.). New York, NY: Guilford.

Rogers, R, Bagby, R M & Dickens, S E (1992). *Structured Interview of Reported Symptoms (SIRS) and professional manual*. Odessa, FL: Psychological Assessment Resources.

Rogers, R, Sewell, K W, Grandjean, N R & Vitacco, M J (2002). The detection of feigned mental disorders on specified competency measures. *Psychological Assessment, 14*, 177–83.

Rogers, T B (1995). *The psychological testing enterprise: An introduction*. Belmont, CA: Wadsworth.

Roid, G (2003). *Stanford-Binet Intelligence Scale–Fifth Edition*. Itasca, IL: Riverside Publishing.

Ross, M W (1984). Intelligence testing in Australian Aboriginals. *Comparative Education, 20*(3), 371–5.

Roth, B, Becker, N, Romeyke, S, Schäfer, S, Domnick, F & Spinath, F M (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118–137.

Rotter, J B (1954). *Social learning and clinical psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Rotundo, M & Sackett, P R (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology, 87*, 66–80.

Rounds, J B, Davidson, M L & Dawis, R V (1979). The fit between Strong-Campbell Interest Inventory General Occupational Themes and

Holland's hexagonal model. *Journal of Vocational Behavior*, *15*, 303–15.

Russell, M & Karol, D (1994). *16PF Fifth Edition: Administrator's manual*. Champaign, IL: IPAT.

Ruzgis, P (1994). Thurstone, L L (1887–1955). In R J Sternberg (Ed.). *Encyclopedia of human intelligence* (pp. 1081–4). New York, NY: Macmillan.

Sackett, P R & Lievens, F (2008). Personnel selection. *Annual Review of Psychology*, *59*, 419–50.

Sadler, D R (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–44.

Sadler, R (1998). Formative assessment: revisiting the territory. *Assessment in Education: Principles, Policy and Practice*, 5(1), 77–85.

Salekin, R T (2002). Factor-analysis of the Millon Adolescent Clinical Inventory in a juvenile offender population. *Journal of Offender Rehabilitation*, *34*(3), 15–29.

Sawyer, J M (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, *66*, 178–200.

Scheu, I E & Lawrence, T (2013). Considerations of translating psychological tests into digital mediums: A case study. *Journal of Educational Computing Research*, *49*, 133–54.

Schmidt, F L & Hunter, J E (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–74.

Schmidt, F L, Hunter, J E, McKenzie, R C & Muldrow, T W (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, *64*, 609–26.

Schneider, W J & McGrew, K (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, NY: Guilford.

Schrank, F A, Mather, N & McGrew, K S (2014). *Woodcock-Johnson IV Tests of Achievement*. Rolling Meadows, IL: Riverside.

Schriesheim, C A & Hill K D (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41*, 1101–14.

Schrank, F A, McGrew, K S & Mather, N (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities*. Rolling Meadows, IL: Riverside.

Schultheiss, O C (2008). Implicit motives. In O P John, R W Robins & L A Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed.; pp. 603–33). New York, NY: Guilford.

Scott, S G & Einstein, W O (2001). Strategic performance appraisal in team-based organizations: One size does not fit all. *Academy of Management Executive, 15*, 107–16.

Scott, W D (1908). *The psychology of advertising*. New York, NY: Arno.

Scriven, M (1967). The methodology of evaluation. In R W Tyler & R M Gagne (Eds.), *Perspectives of curriculum evaluation*. Chicago, IL: Rand-McNally.

Segall, D O (2009). Principles of multidimensional adaptive testing. In W J Van Der Linden (Ed.). *Elements of adaptive testing* (pp. 57–75). New York, NY: Springer.

Seligman, M E P & Csikszentmihalyi, M (2000). Positive psychology: An introduction. *American Psychologist, 55*, 5–14.

Senate Standing Committee on Education and Employment (2014). *Effectiveness of the National Assessment Program—Literacy and Numeracy: Final Report*. Canberra, ACT: Commonwealth of Australia.

Sharpley, C F & Pain, M D (1988). Psychological test usage in Australia. *Australian Psychologist, 23*, 361–9.

Shears, M & Harvey-Beavis, A (2001). *Self Directed Search, Australian manual, Second Australian edition*. Melbourne, Vic: ACER.

Shellenberger, S (1982). Presentation and interpretation of psychological data in educational settings. In C R Reynolds & T B Gutkin (Eds.), *The handbook of school psychology*. New York, NY: Wiley.

Shepard, L & Baker, E (2016). In Memoriam: Robert L. Linn. *Educational Measurement: Issues and Practice*, 62–3.

Shostrom, E L (1980). *Personal orientation inventory manual*. San Diego, CA: Educational and Industrial Testing Service.

Shum, D (2003). Review of the Learning Style Inventory (Version 3). In J Impara & B Plake (Eds.), *Mental Measurements Yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Shum, D & O'Gorman, J (2001). A test of remote memory for use in Australia. *Australian Journal of Psychology*, *53*, 36–45.

Shum, D, McFarland, K & Bain, J (1990). The construct validity of eight tests of attention. *Clinical Neuropsychologist*, *4*, 151–62.

Shum, D, O'Gorman, J & Alpar, A (2004). Effects of incentive and preparation time on performance and classification accuracy of standard and malingering-specific memory tests. *Archives of Clinical Neuropsychology*, *19*, 817–23.

Shum, D, Sweeper, S & Murray, R (1996). Performance on verbal implicit and explicit memory tasks following traumatic brain injury. *Journal of Head Trauma Rehabilitation*, *11*, 43–53.

Shweder, R A & D'Andrade, R G (1980). The systematic distortion hypothesis. In R A Shweder (Ed.), *Fallible judgement in behavioural research: New direction for methodology of behavioral science* (pp. 37–58). San Francisco, CA: Jossey-Bass.

Simons, R, Goddard, R & Patton, W (2002). Hand-scoring error rates in psychological testing. *Assessment*, *9*, 292–300.

Simpson, D (2005). Phrenology and the neurosciences: Contributions of F J Gall and J G Spurzheim. *Australian and New Zealand Journal of Surgery, 75*, 475–82.

Sireci, S G (2005). The most frequently *unasked* questions about testing. In R P Phelps (Ed.), *Defending standardized testing* (pp. 111–21). Mahwah, NJ: Erlbaum.

Slade, T, Johnston, A, Oakley Browne, M A, Andrews, G & Whiteford, H (2009). 2007 national survey of mental health and wellbeing: Methods and key findings. *Australian and New Zealand Journal of Psychiatry*, *43*, 594–605.

Slattery, J P (1989a). *Report of the Royal Commission into Deep Sleep Therapy (Vol. 1). Introduction*. Sydney, NSW: NSW Government Printing Office.

Slattery, J P (1989b). *Report of the Royal Commission into Deep Sleep Therapy (Vol. 4). The Deaths*. Sydney, NSW: NSW Government Printing Office.

Slattery, J P (1989c). *Report of the Royal Commission into Deep Sleep Therapy (Vol. 8). The Departments*. Sydney, NSW: NSW Government Printing Office.

Slattery, J P (1989d). *Report of the Royal Commission into Deep Sleep Therapy (Vol. 9). Psychometric Testing Appendices*. Sydney, NSW: NSW Government Printing Office.

Smith, A (1982). *Symbol Digit Modality Test (SDMT): Manual (Revised)*. Los Angeles: Western Psychological Services.

Smith, A L, Hays, J R & Solway, K S (1977). Comparison of the WISC-R and culture fair intelligence test in a juvenile delinquent population. *Journal of Psychology: Interdisciplinary and Applied*, *97*, 179–82.

Smith, C P (1992). *Motivation and personality: Handbook of thematic content analysis*. Cambridge, UK: Cambridge University Press.

Snyder, C R (1995). Conceptualizing, measuring, and nurturing hope. *Journal of Counseling and Development*, *73*, 355–60.

Snyder, C R, Ilardi, S S, Cheavens, J, Michael, S T, Yamhure, L & Sympson, S (2000). The role of hope in cognitive-behavior therapies. *Cognitive Therapy and Research*, *24*(6), 747–62.

Snyder, C R, Shenkel, R J & Lowery, C R (1977). Acceptance of personality interpretations: The 'Barnum effect' and beyond. *Journal of Consulting and Clinical Psychology*, *45*(1), 104–14.

Spangler, W D (1992). Validity of questionnaire and TAT measures of need achievement: Two meta-analyses. *Psychological Bulletin*, *112*, 140–54.

Sparrow, E P (2010). *Essentials of Conners Behavior Assessments*. Hoboken, NJ: Wiley.

Sparrow, S S, Cicchetti, D V & Balla, D A (2005). *Vineland Adaptive Behavior Scales: Second edition (Vineland II), survey interview form/caregiver rating form*. Livonia, MN: Pearson Assessments.

Sparrow, S S, Cicchetti, D V & Saulnier, C A (2016). *Vineland Adaptive Behavior Scales, Third Edition (Vineland-3*). San Antonio, TX: Pearson.

Spearman, C (1927). *The abilities of man: Their nature and measurement*. London: Macmillan.

Spector, P (1997). *Job satisfaction*. Thousand Oaks, CA: Sage.

Spielberger, C (1983). *Manual for the State-Trait Anxiety Inventory* (rev. ed.). Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C D & Reheiser, E C (2009). Assessment of emotions: anxiety, anger, depression, and curiosity. *Applied Psychology: Health and Well-being*, *1*(3), 271–302.

Squire, L R (1987). *Memory and brain*. New York, NY: Oxford University Press.

Squire, L R (1992). Declarative and non-declarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience, 4*, 232–43.

Starch, D & Elliot, E (1912). Reliability in grading high school work in English. *School Review, 20*, 442–57.

Stephan, C, Wesseling, S, Schink, T & Jung, K (2003). Comparison of eight computer programs for receiver-operating characteristic analysis. *Clinical Chemistry, 49*(3), 433–9.

Stern, A F (2014). The Hospital Anxiety and Depression Scale. *Occupational Medicine, 64*, 393–4.

Sternberg, R J (1985a). Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology, 49*, 607–27.

Sternberg, R J (1985b). *Beyond IQ: A triarchic theory of human intelligence*. New York, NY: Cambridge University Press.

Sternberg, R J (1986). Intelligence is mental self-government. In R J Sternberg & D K Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definition* (pp. 141–8). Norwood, NJ: Ablex.

Sternberg, R J (1993). The Sternberg Triarchic Abilities Test (Level H). Unpublished test.

Sternberg, R J (1997). *Successful intelligence*. New York, NY: Plume.

Sternberg, R J (2012). The triarchic theory of successful human intelligence. In D P Flanagan & P L Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.; pp. 156–77). New York, NY: Guilford.

Sternberg, R J & Berg, C A (1986). Quantitative integration: Definitions of intelligence—a comparison of the 1921 and 1986 symposia. In R J Sternberg and D K Detterman (Eds.), *What is intelligence? Contemporary viewpoints on its nature and definition* (pp. 155–62). Norwood, NJ: Ablex.

Sternberg, R J & Horvath, J A (Eds.). (1999). *Tacit knowledge in professional practice*. Mahwah, NJ: Erlbaum.

Sternberg, R J, Wagner, R K, Williams, W M & Horvath, J A (1995). Testing common sense. *American Psychologist, 50*, 912–27.

Stevens, S S (1946). On the theory of scales and measurements. *Science, 103*, 677–80.

Stough, C (2002). Testing the nation's IQ: A blend of science and entertainment. *Australian Psychological Society InPsych Magazine*. Available at www.psychology.org.au/publications/inpsych/iq.

Strauss, E, Sherman, E M S & Spreen, O (2006). *Compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York, NY: Oxford.

Strong, E K, Jr (1927). *Vocational Interest Blank*. Stanford, CA: Stanford University Press.

Strong, E K Jr (1959). *Strong Vocational Interest Blank*. Palo Alto, CA: Consulting Psychologists Press.

Stroop, J R (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–62.

Suhr, J A (2015). *Psychological assessment: A problem-solving approach*. New York, NY: Guilford Press.

Sullivan, K A & Bowden, S C (1997). Which tests do neuropsychologists use? *Journal of Clinical Psychology*, 53, 657–61.

Sundberg, N D (1977). *Assessment of persons*. Englewood Cliffs, NJ: Prentice Hall.

Syeda, M M & Climie, E A (2014). Test review: Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition*. San Antonio, TX: The Psychological Corporation. *Journal of Psychoeducational Assessment*, *32*(3), 265–72.

Taouk, M, Lovibond, P F & Laube, R (2001). *Psychometric properties of a Chinese version of the short Depression Anxiety Stress Scales (DASS21)*. Report for New South Wales Transcultural Mental Health Centre, Cumberland Hospital, Sydney. Available at www2.psy.unsw.edu.au/DASS/Chinese/tmhc.htm.

Taylor, H C & Russell, J T (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology, 23*, 565–78.

Tellegen, A & Ben-Porath, Y S (2008). *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2 Restructured Form): Technical manual*. Minneapolis, MN: University of Minnesota Press.

Tellegen, A, Ben-Porath, Y S, McNulty, J L, Arbisi, P A, Graham, J R & Kaemmer, B (2003). *The MMPI-2 Restructured Clinical Scales: Development, validation, and interpretation*. Minneapolis, MN: University of Minnesota Press.

Thomas, M L (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18*, 291–307.

Thompson, A P, LoBello, S G, Atkinson, L, Chisholm, V & Ryan, J J (2004). Brief intelligence testing in Australia, Canada, the United Kingdom, and the United States. *Professional Psychology: Research and Practice, 35*, 286–90.

Thorndike, R L (1982). *Applied psychometrics*. Boston, MA: Houghton Mifflin.

Thurstone, L L (1929). Theory of attitude measurement. *Psychological Review, 36*, 221–41.

Thurstone, L L (1938). Primary mental abilities. *Psychometric Monographs*. 1. Chicago, IL: Chicago University Press. Cited in H H Harman (1960), *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Thurstone, L L (1954). An analytical method for simple structure. *Psychometrika, 19*, 173–94.

Thurstone, L L & Thurstone, T G (1941). *The Chicago Tests of Primary Mental Abilities*. Chicago, IL: Science Research Associates.

Tinsley, H E A (1992). Introduction: Special issue on Holland's theory. *Journal of Vocational Behavior, 40,* 109–10.

Tippins, N T, Beatty, J, Drasgow, F, Gibson, W M, Pearlman, K, Segall, D O et al. (2006). Unproctored internet testing in employment settings. *Personnel Psychology, 59,* 189–225.

Tisi, J, Whitehouse, G, Maughan, S & Burdett, N (2013). *A review on marking reliability research (Report for Ofqual).* Slough, UK: NFER.

Tombaugh, T N (1996). *Test of Memory Malingering*. Los Angeles: Multi-Health Systems.

Tong, T, Chignell, M, Lam, P, Tierney, M C & Lee, J (2014). Designing serious games for cognitive assessment of the elderly. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care, 3*(1), 28–35.

Topping, G D & O'Gorman, J G (1997). Effects of faking set on validity of the NEO-FFI. *Personality and Individual Differences, 23*(1), 117–24.

Tracey, T J & Rounds, J (1995). The arbitrary nature of Holland's RIASEC types: A concentric circles structure. *Journal of Counselling Psychology, 42,* 431–40.

Tracey, T J G & Rounds, J (1996). The spherical representation of vocational interests. *Journal of Vocational Behavior, 48,* 3–41.

Travers, K M, Creed, P A & Morrissey, S (2015). The development and initial validation of a new scale to measure explanatory style. *Personality and Individual Differences, 81,* 1–6.

Tulving, E (1972). Episodic and semantic memory. In E Tulving & W Donaldson (Eds.), *Organization of memory* (pp. 382–403). New York, NY: Pergamon Press.

Tupes, E C & Christal, R E (1961/1992). Recurrent personality factors based on trait ratings. Technical Report ASD-TR-61-97, Lackland

Air Force Base, TX: Personnel Laboratory, Air Force Systems Command. Reprinted in *Journal of Personality*, *60*, 225–51.

Ulrich, R E, Stachnik, R J & Stainton, N R (1963). Student acceptance of generalized personality interpretations. *Psychological Reports,* 1963, *13*, 8314.

Upperton, R A & Thompson, A P (2007). Predicting juvenile offender recidivism: Risk-need assessment and juvenile justice officers. *Psychiatry, Psychology & Law*, *14*(1), 138–46.

Uttl, B & Graf, P (1997). Color Word Stroop test performance across the adult life span. *Journal of Clinical and Experimental Neuropsychology*, *19*, 405–20.

Van Iddekinge, C H, Putka, D J & Campbell, J P (2011). Reconsidering vocational interests for personnel selection: The validity of an interest-based selection test in relation to job knowledge, job performance, and continuance intentions. *Journal of Applied Psychology*, *96*, 13–33.

Van Iddekinge, C H, Roth, P L, Putka, D J & Lanivich, S E (2011). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology*, *96*, 1167–94.

Van Iddekinge, C H, Roth, P L, Raymark, P H & Odle-Dusseau, H N (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology*, *97*, 499–530.

Vandecreek, L & Knapp, S (1997). Record keeping. In J R Matthews & C E Walker (Eds.), *Basic skills and professional issues in clinical psychology* (pp. 155–72). Boston, MA: Allyn & Bacon.

Vanderah, T W & Gould, D J (2016). *Nolte's The human brain: An introduction to its functional anatomy* (7th ed.). Philadelphia, PA: Mosby Elsevier.

Veres, J G, Sims, R R & Locklear, T S (1991). Improving the reliability of Kolb's Learning Style Inventory. *Educational and Psychological*

Measurement, 51, 143–50.

Vernon, P E (1965). Ability factors and environmental influences. *American Psychologist, 20*, 723–33.

Vernon, P E (1979). *Intelligence: Heredity and environment*. San Francisco, CA: W H Freeman.

Villarreal, V (2015). Review: Woodcock-Johnson IV Tests of Achievement. *Journal of Psychoeducational Assessment, 33*(4), 391–8.

Vinchur, A J (2014). A history of psychology applied to employee selection. In L L Koopes (Ed.), *Historical perspectives in industrial and organizational psychology* (pp. 193–218). New York, NY: Psychology Press.

Visser, B A, Ashton, M C & Vernon, P A (2006). Beyond g: Putting multiple intelligences theory to the test. *Intelligence, 34*, 487–502.

Vitelli, R (2001). Review of the Test of Memory Malingering. In B S Plake & J C Impara (Eds.), *The Mental Measurements Yearbook* (pp. 1258–9). Lincoln, NE: The Buros Institute of Mental Measurements.

Vrieze, S I & Grove, W M (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice, 40*(5), 525–31.

Wagner, R K & Sternberg, R J (1991). *Tacit knowledge for managers*. San Antonio, TX: The Psychological Corporation.

Wang, S, Jiao, H, Young, M J, Brooks, T & Olson, J (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*, 5–24.

Wang, K. Shi, H-S, Geng, F-L, Zou, L-Q, Tan, S-P, Wang, Y, Neumann, D L, Shum, D H K, Wang, Y & Chan, R C (2016). Cross-cultural validation of the Depression Anxiety Stress Scale-21 in China. *Psychological Assessment, 28*(5), e88–100.

Warr, P B, Cook, J D & Wall, T D (1979). Scales for the measurement of work attitudes and psychological well-being. *Journal of Occupational and Organizational Psychology, 58*, 229–42.

Waschl, N A, Nettelbeck, T, Jackson, S A & Burns, N A (2016). Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability. *Personality and Individual Differences, 100*, 157–66.

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. 459–83). New York, NY: Guilford.

Waterman, A S (2013). The humanistic-positive psychology divide: Contrasts in philosophical foundations. *American Psychologist, 68*, 124–33.

Waterhouse, L (2006). Multiple intelligences, the Mozart effect, and emotional intelligence: A critical review. *Educational Psychologist, 41*, 207–25.

Watkins, C E Jr, Campbell, V L, Nieberding, R & Hallmark, R (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26*, 54–60.

Way, W D & McClarity, K L (2012). A summary of mode comparability research and considerations. In G J Cizek (Ed.), *Setting performance standards* (pp. 451–63). New York, NY: Routledge.

Ways, W D, Davis, L L, Keng, L & Strain-Seymour, E (2016). From standardization to personalization: The compatibility of scores based on different testing conditions, modes, and devices. In F Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 260–84). New York, NY: Routledge.

Wechsler, D (1939). *The measurement of adult intelligence*. Baltimore: Williams & Wilkins.

Wechsler, D (1945). A standardized memory scale for clinical use. *Journal of Psychology*, *19*, 87–95.

Wechsler, D (1949). *Wechsler Intelligence Scale for Children*. New York, NY: The Psychological Corporation.

Wechsler, D (1955). *Manual for the Wechsler Adult Intelligence Scale*. New York, NY: The Psychological Corporation.

Wechsler, D (1974). *Manual for the Wechsler Intelligence Scale for Children–Revised (WISC–R)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D (1981). *Manual for the Wechsler Adult Intelligence Scale–Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D (1987). *Manual for the Weschler Memory Scale–Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D (1991). *Wechsler Intelligence Scale for Children–Third Edition (WISC–III)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D (1997a). *Manual for the Wechsler Adult Intelligence Scale–Third Edition (WAIS–III)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D (1997b). *Manual for the Wechsler Memory Scale–Third Edition (WMS–III)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D (2002). *Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WPPSI-III)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D (2003). *Wechsler Intelligence Scale for Children–Fourth Edition (WISC–IV)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D (2008). *Wechsler Adult Intelligence Scale: Technical and interpretative manual–Fourth Edition (WAIC–IV)*. San Antonio, TX: Pearson.

Wechsler, D (2009a). *Wechsler Individual Achievement Test* (3rd ed.). San Antonio, TX: Pearson.

Wechsler, D (2009b). *Wechsler Memory Scale–Fourth Edition (WMS-IV)*. San Antonio, TX: Pearson.

Wechsler, D (2012a). *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition (WPPSI-IV): Administration and scoring manual.* San Antonio, TX: The Psychological Corporation

Wechsler, D (2012b). *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition (WPPSI-IV): Technical and interpretative manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D (2012a). *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition (WPPSI–IV): Administration and scoring manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D (2012b). *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition (WPPSI-IV): Technical and interpretative manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D (2014). *Wechsler Intelligence Scale for Children–Fifth Edition (WISC–V): Technical and interpretive manual supplement*. Bloomington, MN: Pearson.

Weis, D, Dawis, R, England, G & Loftquist, L (1967). *Manual for the Minnesota Satisfaction Questionnaire*. Minneapolis, MN: University of Minnesota.

Weiss, D J (Ed.) (1983). *New horizons in testing: Latent trait theory and computerised adaptive testing*. New York, NY: Academic Press.

Welch, H J, Welch, H J & Myers, C S (1932). *Ten years of industrial psychology: An account of the first decade of the National Institute of Industrial Psychology*. London, UK: Pitman.

Westen, D (1995). A clinical-empirical model of personality: Life after the Mischelian ice age and the NEO-lithic era. *Journal of Personality, 63*(3), 496–524.

Westen, D (1998). The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science. *Psychological Bulletin, 124*(3), 333–71.

Westinghouse Electric Corporation Technology Transfer Program (1997). *GETNA—General employee training needs analysis: A paper-and-pencil tool for determining common denominator training needs*. Department of Energy/Carlsbad Area Office. Retrieved from http://www.workinfo.com/free/downloads/getnatna.pdf.

Widiger, T A & Saylor, K I (1998). Personality assessment. In A Bellack & M Hersen (Eds.), *Comprehensive clinical psychology, Vol. 3* (pp. 145–67). New York, NY: Pergamon.

Wiggins, J S (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison Wesley.

Wiggins, J S (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology, 37*, 395–412.

Wiggins, J S (2003). *Paradigms of personality assessment*. New York, NY: Guilford.

Wiggins, J S, Trapnell, P & Phillips, N (1988). Psychometric and geometric characteristics of the revised Interpersonal Adjective Scales (IAS-R). *Multivariate Behavioral Research, 23*, 517–53.

Wilkinson, G S & Robertson, G J (2006). *Wide Range Achievement Test–Fourth Edition: Professional manual*. Lutz, FL: Psychological Assessment Resources.

Wilkinson, D, Zhang, J & Parker, P (2011). Predictive validity of the Undergraduate Medicine and Health Sciences Admission Test for medical students' academic performance. *Medical Journal of Australia*, *194*, 341–4.

Williams, J B (2006). Assertion-reason multiple-choice testing as a tool for deep learning: a qualitative analysis. *Assessment and Evaluation in Higher Education*, *3*1(3), 287-301.

Williams, N (2011). *Psychoanalytic diagnosis: Understanding personality structure in the clinical process* (2nd ed.). New York, NY: Guilford.

Wilmoth, D (2007). Family Court psychological evaluations: How not to be part of the fallout. *In Psych*, December.

Wilson, B, Emslie, H, Foley, J, Shiel, A, Watson, P, Hawkins, K et al. (2005). *Cambridge Prospective Memory Test (CAMPROMPT) manual*. Oxford, UK: Harcourt Assessment.

Winkler, J D, Kanouse, D E & Ware, J E (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, *67*, 555–61.

Winter, D G & Stewart, A J (1977). Power motive reliability as a function of retest instructions. *Journal of Consulting and Clinical Psychology*, *45*, 436–40.

Wise, P S (1989). *The use of assessment techniques by applied psychologists*. Belmont, CA: Wadsworth.

Witt, K J, Oliver, M & McNichols, C (2016). Counseling via avatar: Professional practice in virtual worlds. *International Journal for the Advancement of Counselling*, *38*, 218–36.

Wonderlic, Inc. (2012). *Wonderlic Contemporary Cognitive Ability Test (WPT-R) Administrator's guide.* Vernon Hills, IL.: Wonderlic, Inc.

Wood, J M, Nezworski, M T Lilienfeld, S O & Garb, H N (2003). *What's wrong with the Rorschach?* San Francisco, CA: Jossey Bass.

Wood, J M, Garb, H N, Nezworski, M T, Lilienfeld, S O & Duke, M C (2015). A second look at the validity of widely used Rorschach Indices: Comment on Mihura, Meyer, Dumitrascu, and Bombel. *Psychological Bulletin*, *141*(1), 236–49.

Woodcock, R W (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment*, *8*, 231–58.

Woodcock, R W, McGrew, K & Mather, N (2001). *The Woodcock-Johnson Tests of Achievement: Third edition*. Itasca, IL: Riverside.

World Health Organization (1992–94). *International statistical classification of diseases and related health problems*. Geneva, Switzerland: Author.

Wormith, J S, Olver, M E, Stevenson, H E & Girard, L (2007). The long-term prediction of offender recidivism using diagnostic, personality, and risk/need approaches to offender assessment. *Psychological Services*, *4*(4), 287.

Yang, Y & Green, S B (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, *29*, 377–92.

Yee, N, Bailenson, J N, Urbanek, M, Chang, F & Merget, D (2007). The unbearable likeness of being digital: The persistence of nonverbal social norms in online virtual environments. *Cyberpsychology & Behavior*, *10*, 115–21.

Yuen, E K, Herbert, J D, Forman, E M, Goetter, E M, Comer, R & Bradley, J (2013). Treatment of social anxiety disorder using online virtual environments in Second Life. *Behavior Therapy*, *44*, 51–61.

Yussen, S R & Kane, D T (1985). Children's conceptions of intelligence. In S R Yussen (Ed.), *The growth of reflection in children* (pp. 207–41). New York, NY: Academic Press.

Zenisky, A L (2015). Not a minor concern: Introduction to the themed issue on the assessment of linguistic minorities. *International*

Journal of Testing, 15, 91–3.

Zigmond, A S & Snaith, R P (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavic, 67*, 361–70.

Zimbardo, P G (2004). Does psychology make a significant difference in our lives? *American Psychologist, 59*, 339–51.

Zimbardo, P G (2006). On rethinking the psychology of tyranny: The BBC prison study. *British Journal of Social Psychology, 45*(1), 47–53. doi:10.1348/014466605X81720.

Ziskin, J & Faust, D (1988). *Coping with psychiatric and psychological testimony*. Marina del Rey, CA: Law and Psychology Press.

Zuckerman, E L (2005). *Clinician's thesaurus: The guide to conducting interviews and writing psychological reports*. New York, NY: Guilford Press.

Zuckerman, M (2002). Zuckerman-Khulman personality questionnaire (ZKPQ): An alternative five-factorial model. In B De Raad & M Perugini (Eds.), *Big Five assessment* (pp. 377–96). Gottingen, Germany: Hogrefe & Huber.

Index