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

KEY TERMS

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

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.

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

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 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.

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

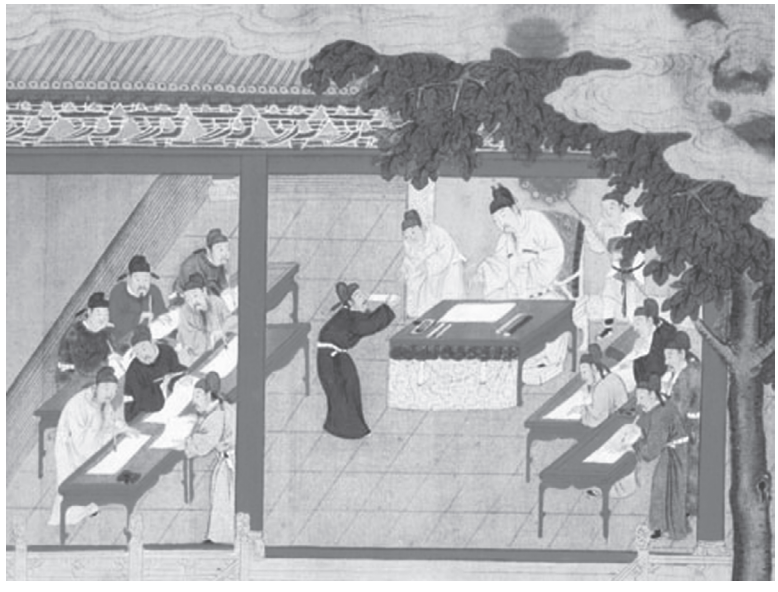
The origins of psychological testing can be found in the public service examinations used by Chinese dynasties to select those who would work for them. These were large-scale exercises involving many applicants and several days of testing, which from the era of the Han dynasty (206 BCE–220 CE) involved 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

Figure 1.1 Imperial examination in China



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.

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

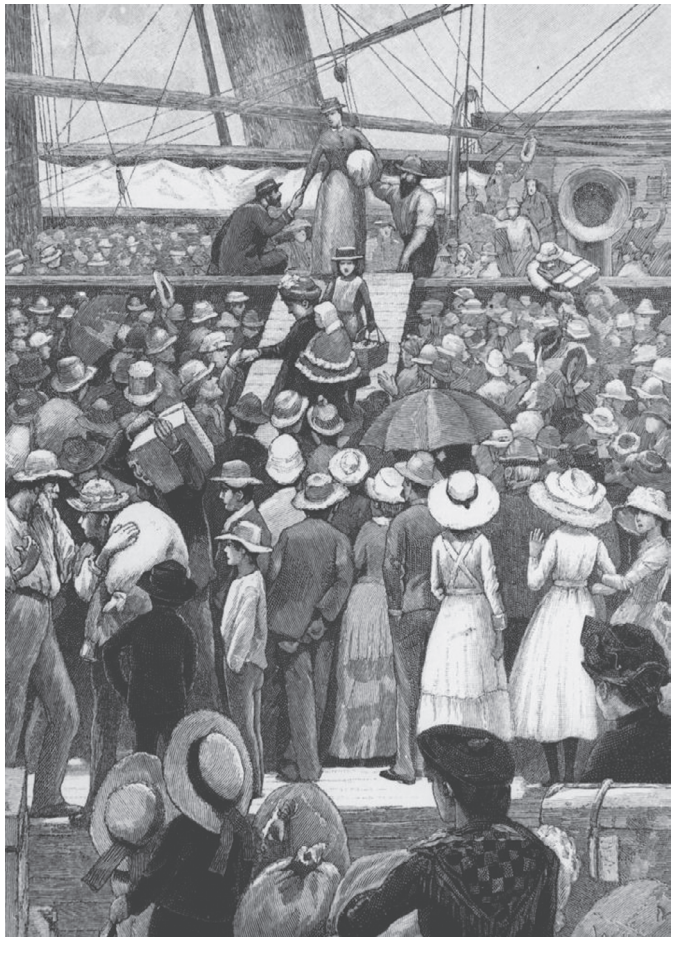
a social context in which outcomes are valued differently by different observers. The most dramatic demonstration of this was the use of testing to enforce immigration policies that most of us today would recognise as manifestly unfair and unjust (see Box 1.2 for an example in the Australian context). 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.

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

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.

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.

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

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

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.

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.

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