Psychological measurements: their uses and misuses

8

'Measure all that can be measured and render measurable all that defies measurement.'
Galileo Galilei

'Not everything that counts can be counted, and not everything that can be counted counts.'
Albert Einstein

The words 'test' and 'measurement', as used in psychology, are misleading because of the implied similarity to scientific measurements and medical tests. Conventional psychological testing is quite different from scientific measurements in natural sciences. What is accomplished by the application of psychological measurement is an estimation of a psychological construct. Psychological tests and measurements of personality, intelligence, attitude and motivation are fundamentally different from quantitative measurements in physical sciences such as height, weight and blood urea. Paul Kline, one of the foremost exponents of psychometric theory clarifies the issue as follows: 'There are no units of [psychological] measurement and no true zeros. Whatever psychological measurement is, it is not scientific measurement as defined in natural sciences ... If we consider what is meant by intelligence or extraversion, just for example, it is by no means clear what units of measurement might be used or what the true zero may mean. This problem applies to the majority of psychological concepts and variables' (Kline, 2000).

Besides, it is often mistakenly believed that psychological tests are 'objective', meaning that their findings and scores reflect an outside existence (as opposed to being subjective) and are real or at least approximate something close to it, as in laboratory tests, for example. The term objectivity has an entirely different meaning when applied to psychological tests or measurements. It refers to the standard ways in which they are administered, recorded and interpreted. This semantic difference has to be kept in mind when using the results of psychological tests. Psychological tests, then, are not tests of mental structures or entities as physical tests, neither are they objective in terms of physical or real existence – they are tests of psychological constructs and are useful to the extent that the underlying theoretical construct and the tests used to measure them are valid (Michell, 1997). It is important to keep these caveats in mind when reading the following account of psychological measurements.

Nevertheless, the term 'psychological measurement' appears to capture the findings of psychological tests better than any other. The terms measurement and test are used interchangeably in this book. Thorndike's oft-quoted statement 'Whatever exists at all, exists in some amount' together

with dictum, 'anything that exists in amount can be measured' may be said to form the basis of the science of measurement of mental or psychological factors, which is known as psychometry.

Despite the difficulties in quantification of psychological characteristics mentioned above, the field of psychometry is a thriving enterprise and, in fact, the ability to 'measure' human psychological attributes is considered to be one of the major advances in psychology.

What are psychological tests?

The definition of a psychological test provided by Anastasi (1982) cannot be bettered:

A psychological test is essentially an objective and standardised measure of a sample of behaviour.

Objectivity and standardisation in this context means that the administration, scoring and interpretation are carried out in a uniform and standard manner. For example, in the use of IQ tests the same test is administered in the same way to all subjects using the same material. It is scored and interpreted according to the strict rules set out in the manual so that the procedure is always consistent. The application of the test and its interpretation do not involve the examiner's subjective judgement. What is measured by tests ultimately is behaviour.

It is to be noted that behaviour, as described here, sometimes involves feelings, motives, interests and other aspects of mental functioning. The tests are designed to capture a representative sample of behaviour that is of interest. The test constructor is faced with the responsibility of devising an array of test items that adequately sample the universe of the domain that is under scrutiny. The information or data obtained by testing is converted to quantitative information or test scores.

Typically psychological tests have been used in three different areas: (1) in occupational settings tests are employed in personnel selection and vocational guidance; (2) in education they are useful for selection through examinations and identification of learning difficulties; and (3) in clinical work psychological tests are used as adjuncts to clinical decision making. This chapter is concerned with clinical applications and what follows is a summary of the various types of tests and their usefulness in clinical settings.

CRITERION-REFERENCED (KEYED) TESTS AND NORM-REFERENCED TESTS

Before discussing the characteristics of good psychological tests, it is important to distinguish between two types of tests: criterion-referenced tests and norm-referenced tests. The method of construction and statistical analysis of the two types are quite different. Tests that compare performance of an individual against agreed outside criteria are called *criterion-referenced tests*. The driving test is a good example. A properly conducted driving test allows a conclusion such as 'the applicant has demonstrated mastery of 75% of the driving skills necessary to be proficient at driving'. The skills that are considered necessary

are selected with the explicit purpose of discriminating between two groups: those who are fit to drive on the road and those who are not. Thus criterion-referenced tests divide the subjects into two groups: those who meet the pre-set criteria and those who do not. These tests are most often used in educational, vocational and clinical settings where the purpose of testing is clear, such as passing an examination, selection for a job or categorisation into diagnostic groups. The contents of the test have no other meaning than their use. Psychiatric diagnostic systems using DSM-IV or ICD-10 (research criteria) are essentially criterion-referenced tests where it is made explicit that a specific number of symptoms and signs be present before making a diagnosis of a particular condition. Those meeting the criteria are deemed 'cases' and those who do not fulfil the criteria are deemed to be 'non-cases'. Thus, criterion-referenced tests, yield all-or-none responses indicating whether the subject has or has not attained the defined criteria. Criterion-referenced tests have been criticised by many for the lack of psychological meaning in their scales and items and the arbitrariness of their cut-off points.

Most tests used in clinical psychological practice are *norm-referenced*, i.e. the subject's test scores are judged against typical test scores from a representative group of individuals (norms). This involves the application of the test to an appropriate group of subjects and the construction of normative data. Such scores are assumed to be normally distributed and the subject's score is interpreted against group norms. Normative data are supplied by the test manufacturer and are available in the test manual. Thus, the conclusions from norm-referenced tests are comparative (there is no set criterion). In the example of the driving test, a norm-referenced test would conclude as follows: 'the examinee has performed better than 65% of all candidates who took the test'. The process of setting up normative data is called test standardisation.

Norms and standardisation

The numerical report of a person's performance on a norm-referenced test or scale ('raw scores'), is of little significance by itself. Interpretation of test scores are meaningful only when the scores are compared with those of a group of individuals of similar age, sex, social class and other important variables. Standardisation of a test involves testing a large sample population, for which the test is designed, in order to obtain a set of norms against which individual scores can be compared.

To be meaningful in interpreting test scores the standardisation procedure should take into consideration two important factors:

1. *Size of the standardisation sample*. The sample should be sufficiently large to reduce standard error of measurement. In studying individual variations such as personality and intelligence, one requires large samples – a factor that makes the procedure tedious and expensive.
2. *Representativeness of the sample*. Adequate comparisons can be made only if similarity exists between the group or individual being tested and the standardisation sample, which should therefore be representative of the

population for whom the test is intended. For example, if a test is designed for those aged 70 years, the normative or standardisation sample should consist of a sample of those aged 70 and above.

Thus, standardisation involves first defining the population for whom the test is designed, and drawing a sample from the population such that it is representative of that population. Inevitably it also involves stratification of the samples into homogeneous groups by such variables as age, sex, social class, level of literacy, and so on. Each subgroup needs a sufficient number of subjects, usually a minimum of 300, to make up an adequate sample. Thus, for a general population study of, say, intelligence, a very large number of subjects are required, in the order of several thousands. The standardisation sample for WAIS-III was 2450 people stratified according to age (thirteen age groups were created), sex, ethnicity, geographical region and educational level.

Expressing the results of norm-referenced tests

Raw scores obtained from tests are converted into standardised scores so that an individual's status with respect to his or her score can be compared with that of the normative group. There are two common methods of reporting performances on psychological tests: percentiles and standard scores. This serves two purposes: it makes scores comparable and, in the case of standard scores – but not in the case of percentiles – it permits the data to be statistically analysed.

Percentiles. Percentile scores provide an index of where one's score stands relative to all others on a scale of 1 to 100. Thus if an individual's IQ is reported to be on the 15th percentile, it means that only 15% of the standardisation sample received scores at or below the score the subject attained. Percentiles have the advantage of being easy to understand and easy to calculate. They are calculated by dividing the number of scores below the index score by the total sample size and then multiplying it by 100. Percentiles are on an ordinal scale (see below) and the difference between percentile units are not equivalent throughout. Hence they do not lend themselves to statistical analysis.

Standard scores. These show how far the individual's raw score deviates from the mean for the relevant normative sample. They are expressed in terms of the mean and the standard deviation of the raw scores for the normative group. The common standard score methods are:

- *T-scores* – a commonly used system with a mean of 50 and a standard deviation of 10.
- *IQ format scores* (also called deviation IQ) – this has a mean of 100 and a standard deviation of 15 (some tests use an *SD* of 16).
- Z-scores – these have a mean of zero and a standard deviation of three.
- *Stantine* (meaning standard as nine) – used less often, they have a mean of 5 and standard deviation of 2.

Standardised scores and their percentile equivalents are shown in Table 8.1. Box 8.1 provides brief definitions of the psychometric terms described so far.

Table 8.1 Percentiles and their equivalents in the standard score systems

Percentile score for normal population	*Z-score (Mean = 0) (SD = 1)*	*T-score (Mean = 50) (SD = 10)*	*Deviation IQ (Mean = 100) (SD = 15)*	*Stantine (Mean = 5) (SD = 2)*
—	– 4 SD	10	—	—
1	– 3 SD	20	55	—
2.5	– 2 SD	30	70	1
16	– 1 SD	40	85	3
50	***0***	***50***	***100***	***5***
84	+ 1 SD	60	115	7
97.5	+ 2 SD	70	130	9
99	+ 3 SD	80	145	—
—	+ 4 SD	90	—	—

CHARACTERISTICS OF GOOD PSYCHOLOGICAL TESTS

In order to be deemed good, a psychological test requires the following properties. It should:

- be reliable;
- be valid;
- possess good norms, i.e. be properly standardised (or fit similar models); and
- be appropriate for the person's age, cultural, linguistic and social background.

Box 8.1 Definitions of some of the common terms used in psychometrics

- *Factor analysis* – a statistical technique used to isolate underlying relationship between sets of variables
- *Normalised scores* – scores obtained by transforming raw scores in such a way that the transformed scores are normally distributed and have a mean of 0 and a standard deviation of 1 (or some linear function of these numbers)
- *Norms* – a list of scores and corresponding percentile ranks, standard scores, or other transformed scores of a group of examinees on whom a test was standardised
- *Standardisation* – administering of a carefully constructed test to a large, representative sample of people under standard conditions for the purpose of determining norms
- *Standard scores* – scores that express an individual's distance from the mean in terms of the standard deviation of the distribution, e.g. *T*-scores, deviation IQs, *Z*-scores and Stantines
- *Raw scores* – an examinee's unconverted score on a test, e.g. number of correct answers
- *Deviation IQ* – intelligence quotient (IQ) obtained by converting raw scores on an intelligence test to a score distribution having a mean of 100 and a known standard deviation, e.g. 15 for Wechsler tests

Scales of measurement

Psychological tests assign numbers to, and yield, numerical scores. But the resulting numbers may mean different things depending on the nature of the scale of measurement. There are four basic scales of measurement, of which only two are applicable to psychological tests:

1. *Nominal*. This classifies subjects on mutually exclusive categories as in male–female, consultants–trainers–administrators. Nominal scales are discrete.
2. *Ordinal*. This represents position in the group and is ranked according to first, second, third, and so on, which gives the order in which individuals are placed but does not tell us how far apart the people in various positions are, e.g. as in ranking by height or weight.
3. *Interval*. This measurement uses equal intervals such as minutes, degrees (temperature), number of words recalled in a memory test or percentage scored in an exam. Intervals on the scale are of equal size, so that 10 to 15 minutes is the same interval as 20 to 25 minutes. For interval scales there is no true zero.
4. *Ratio*. These are interval scales with a true zero point. Most measurements of physical qualities such as height, weight, time and distance are ratio scales.

Interval and ratio levels of measurement provide the greatest information when measuring a variable. For statistical analyses (parametric tests) one needs interval data or ratio data. Most psychological tests provide interval measurements, thus permitting transformation and comparison of scores.

Reliability

Reliability may be defined as the extent to which the outcome of a test remains unaffected by irrelevant variations and procedures of testing. It addresses the extent to which test scores obtained by a person are the same if the person is re-examined by the same test on different occasions. In other words, it is the accuracy with which the test measures a given psychological characteristic. In essence, reliability is the extent to which test scores are free from measurement error. Charles Spearman was the first to point out that all measurements are open to error and that an individual's score on a particular test can vary from time to time and from one situation to another. This theory of true scores, also called the classical theory, states that the observed test score X is the sum of the true score T and error component E. Thus:

$$X = T + E$$

The true score T is the ideal measurement that one strives for and is postulated to be a constant, while the error component E is assumed to be randomly distributed and cannot be eliminated completely. Such errors arise from various sources: degree of distraction, stress in the examinee, variations in the conditions of the test, and so on. But, the important assumption that E is randomly distributed permits the calculation of the reliability coefficient. A

reliability coefficient of 0.7 means that 70% of the variability of test scores can be accounted for by true-score variance and 30% is accounted for by error-score variance. The basic procedure for establishing reliability is to obtain two sets of measurements and compare them, usually using a correlation coefficient. The minimum permissible figure for test reliability is thought to be 0.8. Below this level the test becomes less meaningful. Most psychometric tests in current use yield reliability coefficients of 0.9 or above.

There are several methods of estimating reliability (see Box 8.2):

Internal-consistency reliability

One part of the test needs to be consistent with other parts of the test. That is, different parts of the test should be measuring the same variable. The question that needs to be answered is: 'Is the test consistent with itself?' Thus, in personality test measuring introversion, the items that are designed to measure introversion must be tapping the same domain – introversion. The commonest method used to measure internal consistency reliability is the *split-half reliability*. Here the test is split into two halves so that a random allocation of half the test items or, more commonly, odd and even items, could be scored and compared with the other half and the correlation coefficient calculated. There are several methods of measuring consistency of a test's contents and homogeneity of items (e.g. split-half coefficient, Cronbach's (coefficient) α) and standard statistical software packages are routinely used to compute such measures. In *alternate-form reliability*, a somewhat different approach to the estimation of internal consistency is to use parallel-form reliability, also known as alternate-form reliability. In this approach each subject is given two versions of the same test and the correlation coefficient is calculated. The two tests need to be carefully constructed to resemble each other as much as possible. Many tests of cognitive ability and personality have alternate forms.

Test–retest reliability

The question here is: 'Does the test produce similar results when applied on different occasions?' In order to answer this question the test is administered at least twice to the same group of individuals after an interval, usually of

Box 8.2 Reliability and types of reliability of psychological measurements

The consistency with which a test measures what it is supposed to measure

- *Split-half reliability* – the scores on a randomly chosen half of the test items are correlated with scores on the other half
- *Test–retest reliability* – scores on the test administered to the same persons are correlated with those obtained on separate occasions
- *Alternate-form reliability* – two forms of the test are administered and the correlations calculated
- *Inter-rater reliability* – extent to which two independent raters scoring the test are correlated

about three months (to obviate the effect of memory for the test), and the correlation coefficient computed. Factors such as fatigue, anxiety, boredom, guessing, and poor test instructions are typical sources of measurement error. There may also be 'practice effects' such that more difficult items can be answered correctly on the second occasion. This is a particular problem for abilities tests.

Inter-rater reliability

Whenever the score of a test is dependent on the judgement of the person doing the scoring, it is important to establish the inter-scorer (rater) reliability of the test. Inter-rater reliability is established by comparing the scores of two independent examiners to determine the extent to which they agree in their observations. In examinations, essay questions are notorious for their poor inter-scorer reliability while scoring of multiple-choice questions (MCQs), particularly using computer programs, is accurate and highly reliable. Inter-rater reliability is improved by providing clear and structured guidelines, using test manuals, providing training and practice. Inter-rater reliability can be calculated for the whole scale or for each item. Of the several possible correlation coefficients that can be calculated from the observations of two or more examiners the kappa (κ) provides the best measure of reliability because it takes into consideration the fact that raters can agree by chance.

Validity

Validity refers to what the test purports to measure and how well it measures the characteristic in question. It must be shown that the test actually does measure what it claims to assess.

Validity is, therefore, the meaningfulness of the test. Strictly speaking, validity is not a property of a test but of the use of the test. It can be considered as a measure of the appropriateness and usefulness of the test. In constructing a test, two important issues that have to be taken into consideration are: (1) the construct to be tested (e.g. intelligence) must be theoretically described accurately; and (2) specific test questions should be developed to measure it adequately. Scientific research leading to accumulation of evidence to support the validity of the test is known as test validation. Test manuals are expected to provide the validity of their tests, including the purposes and populations it is valid for. Unlike reliability, there is no single measure of validity or validity coefficients.

For best results it is advisable to use as many different methods as is practically possible to assess the validity of the test. Four types of validity, which are discussed below, are summarised in Box 8.3:

Content validity. The question here is: 'Does the test adequately cover all the important aspects of the domain that is being measured?' For example, in measurement of empathy, construction of the test would have to take into account the cognitive, affective and behavioural dimensions of empathy and not limit the test to one or two of facets of it. It is also important to exclude

Box 8.3 Validity and types of validity of psychological measurements

The extent to which the test measures what it claims to measure

- *Content validity* – the extent to which the test probes the domain adequately
- *Construct validity* – the extent to which the test correlates with other tests measuring similar variables
  - —*Divergent validity* – the extent to which the test successfully discriminates between related constructs
  - —*Convergent validity* – the degree to which the test is positively correlated with other tests measuring the same variables
- *Criterion validity* – the degree of correlation between the test and an external criterion
  - —*Concurrent validity* – the test measure and the criterion are measured at the same time
  - —*Predictive validity* – the criterion is measured after the test
- *Face validity* – the degree to which the test appears to measure the variable

items from the test that could be attributable to other variables. In the case of empathy it may be important to exclude items measuring sympathy. Likewise, in a test of arithmetic ability measuring performance on multiplication only would be considered poor, while at the same time only tests of arithmetic should be included, not algebra. The test designer needs to define carefully and map out the elements in the domain under study including the boundaries of such domains – in short, he or she should be clear about both what goes in and what stays out of a test.

Construct validity. As mentioned at the beginning of this chapter, in psychology many domains that are studied and measured are hypothetical constructs such as honesty, self esteem, marital satisfaction, intelligence and so on. Often, measurements are based on observable behaviour (including self-reports of behaviour, thoughts, and feelings) and it is assumed that from such reports the underlying construct can be inferred. It is often derived by producing as many hypotheses about the construct as possible and testing them out in empirical studies. In constructing intelligence tests, for example, one sets out a series of hypotheses such as how it will correlate with other tests of intelligence, academic performance, occupational status and factor analytic studies, and so on. Inevitably, hypotheses may need revision. Construct validation is thus a gradual, and often protracted process, during which different workers using different methods assess the same construct. Eventually this leads to the accumulation of a considerable quantity of empirical data, which either supports or refutes the construct. Two types of construct validation deserve mention:

Divergent (discriminant) validity. This refers to the lack of correlation among measures that are thought to measure closely related but different constructs.

Donvergent validity. This refers to the agreement among the different measures that claim to measure the same variable.

For example, one would expect different scales measuring empathy to be highly correlated (convergent validity) while at the same time being poorly related to scores on scales that measure sympathy (divergent validity).

Criterion validity. This is the degree of agreement between the test score and a set of external criteria. Unlike other forms of validity it does not involve theoretical constructs or an explanation of the variables – it is the statistical

relationship between measures of test (predictor) and the outside criteria. Generally speaking, the criteria need to be definable, acceptable and easily measurable. For example, in testing blood-alcohol levels for drink driving the results of the test are compared with criteria determined by the legal limit of blood-alcohol concentration (this is most often carried out by the breathalyser test), the assumption being that the alcohol levels in the expired air is a reliable index of blood alcohol levels. While the criterion here is somewhat simple and straight forward – the legal blood-alcohol level – in the case of psychological tests the criteria are often scores obtained by other psychological tests that are considered to be the 'gold standard' at the time of test construction. Any test of intelligence that one may want to devise at the present time would inevitably involve comparison with the WAIS, the industry standard in tests of intelligence

Two commonly used types of criterion validity are concurrent validity and predictive validity. The essential difference between these two forms of validity depends on whether test measure and criterion measure are carried out at the same time or not.

Concurrent validity. Here the test under construction and the test measuring the criterion are administered at the same time. For example, a putative test of intelligence and WAIS are administered at the same time, or close together, and the correlation between the two is computed. In many instances the criterion or 'gold standard' may not exist (if it does there is no need to construct a test in the first place!). In many cases one would expect to improve on already existing tests, refine them, make them shorter or use them for different purposes altogether. For example, one may want to devise a short clinical test of intelligence that could be used as a screening measure in day-to-day usage in a population with learning disabilities. In this case, concurrent validation against the WAIS or WISC would be reasonable and desirable. Difficulties arise when there are no such benchmark tests and one has to be satisfied with using closely related tests that provide moderate degrees of correlation.

Predictive validity. This refers to the effectiveness of the test in predicting an individual's performance on a future criterion measure – thus the criterion is measured later. The question for the test developer is: 'How successful is this test in predicting the criterion at a future date?' In the case of intelligence tests in children the selected criterion may be successful in academic tests. Unfortunately, for most psychological tests defining the criterion is beset with a great many problems. For some characteristics like extroversion it is almost impossible to set up criterion measures. The converse is true as well. One of the major challenges for test developers is to design tests for well-known outcome measures or criteria such as violence (towards person or property) or criminality. Lack of predictive power of most tests of personality to predict future offending behaviour, violence or criminality, is a major failing of all existing psychological tests. This arises partly because of the importance of situational and other factors that contribute to human behaviour and it would be naive to expect a test or a battery of tests to be able to predict such complex behaviours.

Table 8.2 Methods of data collection

Source of data	*Method of collection*	*Example*
Behaviour observation	Direct observation and coding	Strange situation procedure
Test procedures	Tested with test material	WAIS, matrix test
Questionnaires and inventories	Yes–no items	EPQ, MMPI
Rating scales	Likert scales	NEOP-R, attitude tests
Projective tests	Interpretation of pictures or other stimuli	TAT, Rorscha's

Face validity. Face validity is concerned with whether a test 'appears' to measure what it is supposed to measure. It is a subjective evaluation by both those taking the test and those using the test. There are no statistical measures to estimate face validity. But it is important for the people taking the test to have an understanding of what is being measured, to motivate them and to help them cooperate with the examiner. The disadvantage of high face validity is that examinees would be able to guess the purpose of the test and provide desirable answers or purposely distort their responses.

METHODS OF DATA COLLECTION

The information necessary for analysis is collected in various ways depending on the psychological construct under consideration. The most common methods of data collection used in psychological measurements are shown in Table 8.2.

Direct observation. A simple approach to measuring behaviour is, of course, to observe it directly. If, for example, it is decided to measure the wandering behaviour in an in-patient with Alzheimer's disease, close observation and 'event or time sampling' would provide a relatively accurate picture of the extent of the behaviour. Unfortunately, direct observation may not be appropriate or feasible for many psychological constructs.

Test procedures. Here the examiner requests the subject to carry out actions using standard methods and materials. In applying the WAIS the testee is presented with verbal and performance tasks and scored according to set rules.

Self-reports. These are statements about the subject's own account of his or her feelings, thoughts or behaviours. Items are selected on the basis of the theoretical construct under study (e.g. extroversion) using rational and empirical approaches and questionnaires are constructed to cover all aspects of the domain. Tests based on self-reports have been criticised for assuming that people show self-knowledge of the attribute they are questioned about. Another potential problem is that people may feel differently about themselves depending on the time and situations. The different methods of recording responses to the items are:

- *Forced-choice method*. Some questionnaires use a dichotomous True/False or Yes/No forced-choice method. The testee is asked to indicate whether he or she agrees with the statement or not.
- *Likert scale*. One of the most popular scales was developed by Likert in 1938. Subjects are presented with the statement of the item and asked to indicate on a five- or seven-point scale how strongly they agree or disagree. For example in the NEO five factor inventory an item on extraversion is, 'I like to be where the action is', and the examinee is asked to select one of the following responses: strongly disagree, disagree, neutral, agree or strongly agree. A numerical value is attached to each response.

Two other methods of marking items, used exclusively in tests of attitude (see Chapter 9), are:

1. *Thurstone scale*. The subject chooses from a large number of independent statements, each with a numerical value; the intervals between statements are approximately equal. For example, in developing a scale for measuring attitude towards abortion, about 21 statements judged on an eleven-point scale by a panel are selected for the scale and the subject is asked to indicate all the statements that they are in agreement with (Table 8.3). Their attitude is the average of these items.
2. Osgood's semantic differential scale. This entails ratings, on a seven-point scale, of an attitude towards an object, person, or idea on numerous bipolar adjectives scale. For instance, in a scale to measure attitude towards mental illnesses the following bipolar adjectives may be employed:

Curable	----------	Incurable
Predictable	----------	Unpredictable
Innocuous	----------	Dangerous

Tests based on self-reports make several fundamental assumptions. First, people are thought to have self-knowledge of the attributes they are being questioned about and are thought to be reliable observers of their own behaviour. Second, individuals are considered to feel the same way about themselves most of the time, irrespective of situations. Third, it is assumed that the person is motivated to disclose his or her true feelings, behaviours and attitudes.

Table 8.3 Example of a Thurstone scale to measure attitude

How favourable	*Attitude towards abortion. Value on 11-point scale*	*Item*
Least	1.3	Abortion is a punishable offence
	3.6	Abortion is morally wrong
Neutral	5.4	Abortion has both advantages and disadvantages
	7.6	It is up to the individual to decide about abortion
	9.6	Abortion is the only solution to many problems
Most	10.3	Not only should abortion be allowed, but it should be enforced under certain conditions

When administering a psychological test one needs to be aware of the effect on the test of interaction between the examiner and the testee. Because tests are carried out mostly under artificial or contrived situation (in the laboratory or clinic) and not under natural conditions, they are open to distortions. One such problem, the response set, is specific to the use of questionnaires. Two others occur under any experimental condition.

1. Response bias or response set. One possible source of error in using self-rating questionnaires and scales is that of response bias, which is also called response set. This refers to the systematic patterns in which the testees respond, which falsify or distort the validity of the test. Several types of response sets have been identified:

1. Response bias of *acquiescence* is the tendency to agree with the items in the questionnaire regardless of their content. It is most likely to occur when items are couched in general terms rather than being specific. For example, the question 'Are you a friendly person?' is likely to produce more positive answers than asking 'Would those close to you consider you a friendly person?'
2. Response bias of *social desirability* occurs when the testee responds to items in the question in ways that are socially more acceptable. This may be a conscious effort to deceive or an unconscious attempt to create a favourable impression.
3. Response bias of *extreme responding* is the tendency to select responses that are extreme. For example, in a Likert scale the testee responds by marking all the items 0 or 7.
4. Response bias *to the middle* is the likelihood of providing safe or non-committal responses and avoiding extreme responses.

Most questionnaires attempt to overcome response bias by incorporating items that are designed to identify clients who are under-reporting or responding randomly. Addition of the lie items to the scale is one such method. The EPI for example, has several lie items like 'Once in a while do you lose your temper and get angry'? An honest answer to this would be 'yes'. Judicious combination of lie items help in identifying response bias but is almost impossible to eliminate biased reporting altogether.

2. Halo effect. Judgement on the general aspect of the test may colour the answers to most other items and skew the responses in one direction. The first few items of the questionnaire may convince the subject that the rest of the items are similar leading to the same response to all the items. (A similar situation arises in interviews and oral examinations when the examiner's response to a particular aspect of the candidate, for example, being articulate and being well spoken may favourably influence their judgement of the candidate on other abilities as well.) In a broad sense this refers to the tendency to assume that if a person is good at one thing they are also good at others as well.

3. Hawthorne effect. This refers to a positive interaction between the subject and the test procedure that may influence the responses. Initially described in industrial psychology where the very fact of the presence of observers at the industrial plant in Hawthorne increased the activity of the workers and hence production, this effect is often evident when subjects feel that they are being examined or scrutinised.

CLINICAL USES (AND ABUSES) OF PSYCHOLOGICAL TESTS

Psychological tests are useful *adjuncts* to clinical decision making, but they should not be used in place of sound clinical judgement. The main role of the clinician in conducting an assessment is to answer specific questions and make decisions about management. In order to accomplish these the clinician must integrate a wide variety of information from diverse sources such as history, examination, observation, and informants including psychometric data. Tests are only one method of gathering information. The clinician's main task is to situate the test results in a wider perspective to generate hypotheses and aid in decision making. For instance, elevated scores on Scale 6 (Hypomania) in the MMPI in a chief executive may reflect a driven, ambitious and dominant personality rather than hypomania. Thus, the best statement that could be made is that elevated or low scores on a test are *consistent* with a particular condition; they are not *diagnostic*.

Results of psychological tests are liable to be misused in some instances. This is especially so when psychometric testing tends to form the basis of social and educational policies. For example, in the past, there has been a tendency to place children to be classed as having 'special educational needs' solely on the premise of having scored low on IQ tests. The result was that children from low-social-class families and minority groups were placed in 'special schools' where the curriculum was less challenging. Thus, the children were doubly deprived, they were performing poorly and were deprived of a proper education thereby increasing their chances of failing academically. This goes against the basic principles of equality of access to education, equal opportunities and human rights. As a result of legal action taken by parents in some states of the USA the WISC is banned from use in educational settings.

A good knowledge of the tests that are available to measure the psychological variable in question is essential. Information about tests may be obtained from *The Mental Measurements Yearbook* (MMY), which provides an up to date authoritative review of most tests. Tests are revised regularly, usually every five years or so, and need to be regularly updated (the 15th edition of MMY, published in 2003, is available on the web at buros.unl.edu/buros/catalog.html). Information is also available from test publishers in the form of manuals that come with the test. These provide instructions on the use and application of the test together with reliability, validity and normative data. Most important of all is whether the test is appropriate and suitable for the specific clinic situation. Most standard tests such as Wechsler intelligence tests and MMPI require training before one can use them. The test results are treated as confidential information and there are strict regulations governing their disclosure to third parties. Informed consent is necessary before administration and both the British Psychological Society and the American Psychological Association have set professional standards and guidelines for test usage. Psychological test publishing is big business and copyrights are usually reserved.

There are several thousands of tests and measurements across the entire field of psychology. These range from measurement of marital satisfaction to

Box 8.4 Commonly used psychological tests

1A. Measures of general or global mental abilities
- *Wechsler scales (WAIS, WISC, WPPSI)*
- *Stanford–Binet intelligence scale*
- *Raven's progressive matrices*
- *British abilities test (BAS)*

1B. Measures of specific mental abilities
- *Wechsler memory scale (WMS)*

2. Measures of personality
- *Personality questionnaires*
  - *—Sixteen personality factor (16 PF)*
  - *—Eysenck personality questionnaire (EPQ–R)*
  - *—Neuroticism extraversion openness personality inventory (NEO–PI–R)*
  - *—Minnesota multiphasic personality inventory (MMPI–2)*
- *Projective tests*
  - *—Rorschach test*
  - *—Thematic apperception test (TAT)*

3. Neuropsychological tests
- *Wechsler memory scale–revised (WMS–R)*
- *Wisconsin card sorting test (WCST)*

ratings of clinical conditions such as anxiety and depression. Most commonly the psychological measurements that are useful in clinical practice are: (1) tests of cognitive abilities; (2) personality tests; and (3) neuropsychological tests (Box 8.4). Brief outlines of the tests are given below, more details on the tests are found in the appropriate chapters.

1. Tests of mental abilities

Often called intelligence tests, tests of general cognitive abilities are the most commonly used tests in clinical and educational psychology. These are based on the psychometric theory of intelligence and the concept of *g* or general intelligence and the factor structure of human abilities; the subscales may be used to measure specific abilities such as arithmetic or reading. The two most common scales used to measure general intelligence are the Wechsler scales and Stanford–Binet scale (see Chapter 2). Tests for specific mental abilities include tests such as Wechsler Memory Scale (WMS) (see Chapter 7).

2. Tests of personality

Personality tests may be broadly grouped into two: projective tests and personality questionnaires.

Projective tests

These are idiographic tests, which are said to test the unique aspect of one's personality. The tests are generally based on the psychoanalytic concept of projection that when individuals are presented with an ambiguous stimulus, which has no 'real' meaning, they project their own feelings, thoughts and fantasies into it. These tests use stimuli that are ambiguous and the subject is

free to respond within certain parameters. The reliability and validity of projective tests are relatively low. The two best-known projective tests are the Rorschach's inkblot test and the thematic apperception test.

The Rorschach test. Perhaps the best known projective test, the Rorschach test (Rorschach, 1921) consists of ten bilaterally symmetrical inkblots (faithfully reprinted in all textbooks in psychology), which the subject is asked to interpret. A variety of interpretations are possible for each ink blot. The individuals' responses are scored according to three general categories: (1) the *location*, or the area of the inkblot on which they focus; (2) *determinants*, or specific properties of the inkblot that they use in making their responses (e.g. colour, shape); and (3) the *content*, or the general class of objects to which the response belongs (animals, humans, faces). Interpretation of the information obtained in this way is thought to be a way of obtaining access to the persons psychological make up. Although popular with psychoanalytically oriented workers and therapists, the major problem for those studying personality characteristics is its low reliability and validity. Hence the test has fallen out of favour.

However, a sophisticated scoring system developed by Exner (1974) has overcome some of these objections. It has revitalised the test by providing a more reliable scoring system. So much so, in fact, that some have argued that the test is now as good as the MMPI in predicting a variety of clinical phenomena.

Thematic Apperception Test. In this test (TAT; Murray, 1943) the subject is presented with a series of cards (about 10) with pictures on them. For example, Card 1 of TAT depicts a boy staring in contemplation at a violin and Card 2 shows a rural scene with a young girl in the foreground holding a book, a pregnant woman watching and a man labouring in a field in the background. The subject is told that it is a test of imagination (not intelligence) and is asked to make up a dramatic story for each of the cards. More specifically, he or she is asked to: (1) tell what has lead up to the events shown in the picture; (2) describe what is happening at the moment; (3) say what the characters are feeling and thinking at the moment; and (4) describe what will happen next. The stories are recorded verbatim, including pauses, questions and other remarks. The areas of interpretation are broken down into three categories: story content; story structure; and behaviour observation.

The assumption is that when presented with such ambiguous stimuli the subjects' stories reflect their personality and characteristic modes of interacting with the environment. Subjects are thought to create TAT stories based on a combination of at least three elements: the card stimulus, the testing environment and the personality or inner world of the individual. The term *apperception* was coined to reflect the fact that subjects do not just perceive, rather they construct stories about cards in accordance with their personality, characteristics and life experiences. Several scoring systems have been developed in attempts to standardise the interpretation of TAT material. The TAT has been used in research and psychotherapy. McClelland (1985) used TAT extensively in his studies on Need for achievement (see Chapter 5, p. 122). It has been used in prediction of overt aggression, identification of defence mechanisms and for mapping interpersonal relationships.

Personality questionnaires

These are the most popular personality tests. The commonly used questionnaires are based on the trait theory of personality and their contents are derived through factor analysis. They are easy to use, norms are easily available and are easy to interpret. However, validity of such tests is dependent on the theory and model of personality that underpins the test, the assumption being that personality dimensions are stable and that behaviour is consistent over a range of situations. One other problem is that the questions, and hence the answers, are to a degree crude and lack refinement. For instance, the following question is from the Eysenck personality inventory: 'Do you find it very hard to take no for an answer?' This is answered 'Yes/No' and is scored on the neuroticism scale if positive, which then raises a second question: 'Under what situation and with whom?'

Factor analysis. The principal psychometric technique used in most studies of personality traits is *factor analysis* and the key statistic involved is Pearson's correlation coefficient (r). Factor analysis is a method for investigating the dimensions or factors that underlie the correlation between more than two variables. It has been used extensively in the psychometric study of personality and intelligence. The purpose of factor analysis is to explore how the variables are interrelated and thereby examine the meaningful elements that they have in common and are responsible for their correlation.

A hypothetical example may illustrate the principles. Supposing we collected data on five scales using questionnaires that describe a sample of people on the following aspects: friendliness; helpfulness; empathy; dominance; and assertiveness. Next we work out the correlations among the five scales and arrange them in a correlation matrix as shown in Table 8.4.

As can be seen from the correlation matrix, the three measures, friendliness, helpfulness and empathy appear to show positive correlations between .70 and .80, whereas the correlation between the first three scales and the last two, dominance and assertiveness is very low (less than .25). But the last two measures, dominance and assertiveness, show high correlations (more than .65). From these observations we could reasonably conclude that a common construct underlies the first three traits while a different factor accounts for the similarities of the last two traits. All that is left for us to do is to name these two first-order factors.

Table 8.4 Correlation matrix for five scales (see text)

Scale	*Friendliness*	*Helpfulness*	*Empathy*	*Dominance*	*Assertiveness*
Friendliness		.75	.70	.15	.20
Helpfulness			.80	.20	.25
Empathy				.10	.20
Dominance					.75
Assertiveness					

Thus, it could be said that two factors have been 'extracted' from the items and, conversely, the respective items of the scales could be said to have been 'loaded' with the particular factors. They are said to be 'orthogonal factors' and can be represented graphically by two lines at right angles to each other. After the initial analysis, in some instances, the factors that emerge (the first-order factors) may be more or less correlated with each other. These can be subjected to further factor analysis and second-order factors obtained. Usually second-order factors are not correlated to one another and can be considered to be different, independent elements that constitute the universe of personality. Factor analysis is a complex area of statistics. An easy introduction to the method may be found in Kline (1994).

Common measures of personality

Despite the potential objections to the use of personality questionnaires to 'measure' personality traits, they are popular with psychologists, especially researchers in the field. The personality tests most often used by psychologists are mentioned below (see also Chapter 3).

16 Personality factor questionnaire. The 16 PF (Cattell, Eber & Tatsuoka, 1970) is a self-administered questionnaire consisting of 187 trichotomous items. It is thought to be a psychometrically sound instrument based on data from almost 25 000 people and has been used mainly in research. It is thought to provide a good instrument for measuring normal adult personality. The original 16 PF provides a profile of the subject's personality on all 16 factors (see Chapter 3). Cattell considered the 16 factors to be primary, meaning that all 16 were necessary for a comprehensive measure of personality. However, recent workers have extracted five second-order factors. The latest version of 16 PF, the 16 PF-5, is designed to be scored on the five factors: extraversion; neuroticism; tough poise; independence; and control.

Eysenck personality questionnaire–revised. The EPQ-R (Eysenck & Eysenck, 1991) is a well-known measure of the three Eysenck personality factors (see Chapter 3) and it has been used in hundreds of studies. It has 100 dichotomous 'yes/no' items. Along with scales measuring the three personality dimensions of extraversion, neuroticism and psychoticism, it also includes a lie scale. It has been normed on almost 5000 subjects, and has high reliability. Its factor analytic validity for the dimensions, especially extraversion/introversion, is good. However, the EPQ has been not popular with clinicians.

NEO personality inventory. The 'big five' dimensions of personality (see Chapter 3) form the basis of the NEO personality inventory (NEO-PI-R; Costa & McCrae, 1992). The NEO-PI-R is a self-administered questionnaire consisting of 240 items designed for use with adults. Each item is rated on a five-point Likert scale, from 'strongly agree' to 'strongly disagree'. Each facet (lower level trait) is assessed by eight statements, such as: 'I am not a worrier' (neuroticism, N) and 'I like to have a lot of people around me' (extraversion, E). A shorter version of the questionnaire comprised of sixty items is also available (NEO-FFI). The test is considered to be reliable and well validated. The main problem with the instrument is that the data derived from it is too general for clinical decision making. Nevertheless, theoretical research using

the NEO-PI-R holds future promise for building bridges between (normal) personality psychology and psychiatry.

The Minnesota multiphasic personality inventory. This instrument (MMPI; Hathaway & McKiney, 1983) is one of the most extensively used and researched tests of personality assessment, especially in the USA. It is a 567-item pencil and paper test. Test items are of the true/false type grouped into validity, clinical and content scales (Box 8.5). Norms for the various scales have been developed on the basis of responses from the normal and clinical groups. It stands out from other personality tests in that it was developed using the criterion keying method rather than factor analysis. For example, patients who had received a diagnosis of depression (and no other) comprised the criterion group. Having made sure that they had an agreed diagnosis of depression, the responses of these patients to the item pool were compared with a group of normal controls. Items that discriminated significantly between the two groups were selected.

The MMPI was originally developed with the purpose of assisting in diagnosis of psychiatric conditions. Experience in using the MMPI has shown that patients usually show elevation in more than one scale. Thus, the MMPI provides a profile of the subject's scores and, over the years, a massive amount of data has accumulated that enables clinicians to identify patterns of *profiles*

Box 8.5 Scales in the Minnesota multiphasic personality inventory (MMPI–2)

1. Validity scales – seven scales
2. Clinical scales
   - *Hypochondriasis (Hs), Scale 1*
   - *Depression (D) Scale 2*
   - *Hysteria (Hy) Scale 3*
   - *Psychopathic deviance (Pd), Scale 4*
   - *Masculinity–femininity (Mf), Scale 5*
   - *Paranoia (Pa), Scale 6*
   - *Psychasthenia (Pt), Scale 7*
   - *Schizophrenia (Sc), Scale 8*
   - *Hypomania (Ma), Scale 9*
   - *Social introversion (Si), Scale 0*
3. Content scales
   - *Anxiety (ANX)*
   - *Fears (FRS)*
   - *Obsessiveness (OBS)*
   - *Depression (DPS)*
   - *Health concerns (HEA)*
   - *Bizarre mentation (BIZ)*
   - *Anger (ANG)*
   - *Cynicism (CYN)*
   - *Antisocial practices (ASP)*
   - *Type A (TPA)*
   - *Low self-esteem (LSE)*
   - *Social discomfort (SOD)*
   - *Family problems (FAM)*
   - *Work interference (WRK)*
   - *Negative treatment indicators (TRT)*

(profile analysis) that assist the clinician in treatment planning and predicting future functioning. The MMPI has been studied intensively and has more than 10 000 research articles to its credit. Computerised interpretive services are available. The scale was updated and revised to increase the utility of the instrument. The MMPI-2 published by Pearson Assessments in 1989 has surpassed its predecessor in its value to clinicians and is now one of the most popular personality tests used by psychologists. A common criticism of the MMPI is the usage of scale names such as depression, which may suggest that the subject shows features of the clinical syndrome. This is not true. Elevated scores on the depression scale reflect characteristics such as apathy, tendency to worry and self-depreciation. Thus the scales represent measures of personality traits rather than diagnostic categories.

3. Neuropsychological tests and assessments

These are tests designed to assess the psychological functioning of patients with cerebral damage. Often they are used to identify deficits in psychological variables such as memory, executive function or intelligence following brain damage or cerebral impairment. They are also used to localise brain lesions (see Chapter 7).

Clinical judgement

When assessing a patient or client the clinician attempts to gather information from a variety of sources (history, examination, past medical notes, accounts from informants and any psychological tests), synthesise and integrate these data and make an informed judgement on the diagnosis and make a sound aetiological formulation of the case. Each of the processes involved in clinical decision making affect their reliability and validity. Errors in clinical judgement are not uncommon. It is crucial to understand the process of decision making and be aware of relevant literature on clinical judgement if we are to improve the accuracy of it. Research has highlighted the following areas (Garb, 1998):

1. The accuracy of judgement is increased by semistructured approaches to interviewing rather than subjective methods based on theoretical knowledge or experience. This is particularly important where judgements on risk are involved. When diagnoses are made it is important to adhere to the specific criteria set out in ICD-10 or DSM-IV.
2. It has been shown that the more confident clinicians feel about their judgements, the less likely they are to be accurate. Interestingly, clinicians with greater knowledge and experience are less confident regarding their judgements and tend to keep an open mind.
3. Clinicians should be aware of the two possible biases to which they are liable when making clinical decisions. One results from the *primacy effect*, where the initial presentation (or referral information) influences information collected later. The implication is that clinicians need to pay sufficient attention to data collected at every stage of the assessment. Another source

of error arises from *confirmatory bias*. Here the clinicians selectively seek out information that confirms their initial judgements. Doctors are particularly prone to make 'instant' diagnoses and thereafter attempt to derive information to justify the decision. Since a diagnosis is nothing but an initial hypothesis, it is crucially important that clinicians carefully attempt to elicit information that may disconfirm their hypothesis as well as support it. Taking a meta-view of the consultation process, which includes the clinician and clinical situation, facilitates balanced decision making and avoids various forms of bias.

REFERENCES

Anastasi A (1982) Psychological Testing (5th ed.). New York: Macmillan.

Cattell RB, Eber HW, Tatsuoka MM (1970) Handbook for the Sixteen Personality Factor Questionnaire. Champaign, IL: Institute for Personality and Ability Testing.

Costa PT, McCrae R (1992) Revised NEO Personality Inventory (NEO-PI-R) (NEO-FFI) Professional Manual. Odessa, FL: Psychological Assessment Resources.

Exner JE Jr (1974) The Rorschach: a comprehensive system. New York: Wiley.

Eysenck HJ, Eysenck MW (1991) The Eysenck Personality Questionnaire – Revised. Sevenoaks, UK: Hodder & Stoughton.

Garb HN (1998) Studying the Clinician: judgement research and psychological assessment. Washington, DC: American Psychological Association.

Hathaway SR, McKiney JC (1983) The Minnesota Multiphasic Personality Inventory (Manual). New York: Psychological Corporation.

Kline P (1994) An Easy Guide to Factor Analysis. London, UK: Routledge.

Kline P (2000) The Handbook of Psychometric Testing (2nd ed.). London, UK: Routledge.

McClelland D (1985) Human Motivation. Glenview, IL: Scott, Foresman.

Michell J (1997) Quantitative science and the definition of measurement in psychology. British Journal of Psychology, 88, 355–383.

MMY (2003) The Mental Measurements Yearbook (15th ed.). [buros.unl.edu/buros/catalog.html]

Murray HA (1943) Manual to the Thematic Apperception Test. Boston, MA: Harvard University Press.

Rorschach H (1921) Psychodiagnostics. Berne: Hans Huben.

FURTHER READING

Anastasi A (1982) Psychological Testing (5th ed.). New York: Macmillan.
- A definitive text on the principles of psychological measurements.

Harding L, Beech JR (eds) (1996) Assessment in Neuropsychology. London, UK: Routledge.
- A good introduction to neuropsychology.

Kline P (1994) An Easy Guide to Factor Analysis. London, UK: Routledge.
- Factor analysis made simple.

Kline P (2000) The Handbook of Psychometric Testing (2nd ed.). London, UK: Routledge.
- A good introduction to the subject including summaries of common tests.