# Psychological Testing: A User's Guide

## Introduction

This guide is about using psychological tests and the principles of good use. Previous test guides issued by The British Psychological Society (BPS) have tended to focus on technical issues underlying tests themselves. While these are essential to the effective use of psychological tests, they represent only one aspect of good practice in testing. The Steering Committee on Test Standards (SCIS) has therefore taken the opportunity of this revision to broaden the scope of the guidelines and consider other issues. Among these are:

• Competence in test use and the recent introduction of standards for the use of tests in occupational assessment.

• Responsible test use and a commitment to such use introduced later in this guide.

This guide is organised in four sections:

• What to look for in a psychological test.

• What qualifies as competence in the use of psychological tests.

• A commitment to responsible use of psychological tests.

• Further information.

## What to look for in a psychological test

Tests are designed for a purpose and the use of a particular test will vary according to the objectives of assessment. Some broad distinctions between different categories of tests can be made as follows:

**Tests of Attainment** are used to assess knowledge and skills acquired through education and instruction. Examples include tests of mathematics knowledge,

foreign language proficiency or mastery in a craft. Such tests tend to be narrowly defined in content and targeted at the achievement of specific standards. Like ability tests, such tests are generally designed so that there is only one correct answer to each test question. The test score is usually the total number of questions answered correctly.

**Tests of Ability** assess broader areas of what a person can do. While scores on such tests are influenced by education and training, they are not designed to assess specific areas of knowledge or skill. Examples of such tests are measures of verbal reasoning (the ability to comprehend, interpret and draw conclusions from oral or written language), numerical reasoning (the ability to comprehend, interpret and draw conclusions from numerical information), spatial reasoning (the ability to understand and interpret spatial relations between objects) and mechanical reasoning (understanding of everyday physical laws such as force and leverage involved in the use of tools and equipment).

**Aptitude Tests** which are used to assess how well an individual is likely to perform in a training programme or in a job. Attainment tests, ability tests and personality tests are all used to predict future performance, and so the term aptitude has more to do with prediction than with a specific category of test.

**Tests of Disposition** which are used to assess how a person is likely to react emotionally to situations and other people, the types of relationship they prefer to establish with others, and the motivational factors that influence a person's actions. In contrast to attainment and ability tests, tests of disposition do not generally contain questions to which there is only one correct answer. Rather, the answers given to questions reflect a tendency to interpret situations or respond to other people in particular ways. Typical qualities assessed by such tests are anxiety, sociability, perseverance, dominance, fear of failure and resistance to stress. Personality tests are the most widely known form of this type of test.

**Tests of Interest and Preference** are similar in their design to tests of disposition, but focus on the activities, hobbies and types of work that a person might enjoy or

might be best suited for. They are frequently used in careers counselling to gauge priorities in life (career, family, life style) and orientation towards work (service to others, invention, entrepreneurship) in order to help focus a person's thinking and to suggest possibilities the person may not have considered previously.

**Tests of Psychological Dysfunction** are among the most complex form of psychological test in dealing with areas that are both sensitive and difficult to diagnose. They are also among the most diverse group of tests in covering a number of conditions and symptoms, and their use requires both general clinical expertise as well as specific knowledge of a particular test. They include assessments of neuropsychological damage resulting from physical trauma or from pathological conditions.

In addition to these categories of tests, broad distinctions can also be made in terms of the settings in which psychological tests are most frequently used. These are:

**The Occupational Setting** in which tests are used in careers guidance, to help select personnel to assess their training and development needs, and in promotion.

**The Educational Setting** in which tests are used to assess levels of educational attainment, learning and instructional needs, and for entry into secondary and tertiary levels of education.

**The Clinical Setting** in which tests are used to diagnose emotional and behavioural conditions as an aid to determining appropriate treatments.

In all three settings, the common theme is that of tests providing information about individual regarding extent (how much of X?) and appropriate action (job offer, instructional programme, treatment). Tests are used for three principal reasons:

- They provide a standardised method for assessing and diagnosing individuals.
- They provide such information more efficiently than most other methods of assessment (e.g. interviews or observation).
- They provide access to the measurement of qualities that are difficult to assess through other means.

In contrast to physical measurements such as height, length, mass or speed, psychological tests measure qualities that are less tangible. Even when there is observable evidence of a condition such as a reading problem or behavioural disorder, the extent and causes of such problems may not be clear from the physical evidence available. So, in contrast to the manifest, observable features of physical measures (i.e. they can be experienced directly by our senses), psychological tests often measure qualities that are hidden, covert or latent (i.e. they cannot be directly or so easily experienced through our senses). As such, psychological tests may provide the only reliable and efficient means of assessment.

This leads us to the first two questions that should be asked of a psychological test: *What does it measure and how does it measure it?* What these questions demand is an explanation of the rationale behind a test; a description of why the test was developed and why it was constructed in the way it was. A test manual should contain such information in the form of a brief history of the test. This history should include any relevant theory supporting the test, the steps taken to construct the test, details of research and summaries of the results of such research. The manual should also state whether the test was designed for a broad, general range of uses, or whether it was designed for use with specific groups of individuals (e.g. ages, occupations, types of condition, as an aid to specific diagnoses or decisions).

With a statement of what the test is supposed to measure, we can then look for numerical evidence of how successful the test construction process has been. We will now examine some of the general quality checks that should be reported in a test manual.

**Reliability.** When a test is administered, the outcome is an *observed score* on the quality measured by the test. However, the observed score may be misleading without information on how successful the test has been in measuring that quality accurately.

Two types of score need to be considered: the person's *true score* on the quality measured, and how closely the observed score matches the person's true score.

Statistical indices known as *reliability coefficients* provide the evidence required to judge how accurately or precisely a person's true score has been estimated by a test. Reliability coefficients take the form of a proportion, and a value of +0.7 is generally held as a minimum requirement for the use of a test. Such a value states that 70 per cent of the differences between people as measured by a test are due to true measurement, while 30 per cent of such differences are due to measurement error. Of course, designers should seek reliability coefficients greater than +0.7 in constructing a psychological test.

There are several different types of reliability coefficient, which address different questions regarding the accuracy of test scores. The test manual should state which type of reliability coefficient is reported and why. The evidence supporting the use of a test is clearly strengthened when the test manual contains information on more than one type of reliability coefficient. The reliability coefficients most commonly reported are:

- **Internal Consistency** which evaluates the extent to which the questions in a test are consistent in their contribution to the accuracy of the overall test score. Low internal consistency would suggest that the test contains poorly constructed questions or questions that measure qualities other than that intended in the design of the test. This type of reliability coefficient is appropriate when the test is made up of discrete questions which are each answered independently. It is not appropriate when the items in a test are ranked one against the other as in some personality and interest inventories, or when the score is based purely on the speed of responses or continuous performance on a task as in tests of clerical speed or hand-eye co-ordination.

- **Alternate Forms** which evaluates the extent to which the ordering of people from higher to lower test scores is consistent across two or more versions of a test. If a person is in the top 10 per cent of people tested on Version A of a test, then one would reasonably expect that person to score in the top 10 per cent on Versions B or

C of the same test. This coefficient is inappropriate unless there is an explicit statement that different tests are alternate forms or versions of the same test.

• **Test-retest or Stability** which evaluates the extent to which the ordering of people from higher to lower scores is consistent across time. If it is expected that the quality measured by a test is unaffected by factors associated with time, then a person's standing at Time A should correspond to their standing on the same test at Time B. The stability coefficient is not appropriate if change over time is an important feature of the quality measured by a test. An example would be a measure of mood which could vary day-today or hour-to-hour.

**Validity.** While reliability is concerned with how accurate or precise a test score is, validity is concerned with what the test score actually measures. It is insufficient to merely state that a test is a measure of, say, mechanical aptitude, tolerance of stress or proficiency in mathematics. Statements like these must be supported by documentation of research data that demonstrates a test score is a meaningful measure of the quality or qualities the test was designed to assess. Three common approaches to the collection of validity evidence will be briefly described.

• **Content Validation.** Most frequently used in developing measures of educational attainment, this approach to validity involves the evaluation of test questions by subject matter experts. Working from a statement defining the content area (what is to be tested?) and standards of achievement (what level defines attainment?), information is collected on whether the experts consider the questions to be an adequate representation of the area to be assessed, and whether the scores taken from the test are an appropriate and accurate means of distinguishing between different levels of attainment (e.g. mastery versus non-mastery of a subject or skill).

• **Criterion Validity.** This is the most common form of validity evidence provided for tests used in occupational settings. Scores on a test are compared to performance in training or in a job to determine whether higher test scores are related to higher training or job performance. Obviously, the stronger this relationship is then the stronger the evidence that the test score predicts subsequent performance. This

evidence is usually reported using the *correlation coefficient* or *validity coefficient* statistic which, although it also takes the form of a proportion, should not be confused with the reliability coefficient. A correlation or validity coefficient may take either positive or negative values, the relevance of the sign depending on the intended direction of prediction: positive if higher test scores are associated with higher performance (e.g. as in predicting training performance from an ability test); negative if lower scores are associated with higher performance (e.g. as in predicting many types of training or job performance from measures of anxiety). Care must be taken to ensure that the sign of the correlation reflects the intended direction of prediction. The significance of a correlation depends on the size of the sample it is based on, but a value of +1-0.3 calculated on a validation sample of 100 or more subjects would support the predictive value of a test. Clearly, larger values reported using larger sample sizes provide stronger evidence of criterion validity.

• **Construct Validity**. Although the concept of construct validity has become closely associated with certain statistical methods for looking at the relationships between various scores (e.g. factor analysis), it also refers more generally to the breadth of research supporting the validity of a test. Information collected from content and criterion validity studies contribute to evidence of construct validity by providing a better understanding of what a test measures and what its limitations are.

**Interpretation.** The observed score is rarely used to interpret the information from a psychological test. Rather, this score is usually transformed into what are referred to as norm scores (in the case of tests of ability or disposition) or criterion scores (in the case of tests of attainment or dysfunction). To be able to interpret these transformed scores, a test user must understand the process by which these scores are arrived at and what they represent. Many tests of disposition and interest generate several scores rather than one single score. Accurate interpretation of these scores depends on understanding the pattern of relationships between them. The test manual should clearly state the procedures for creating transformed scores, why these procedures were chosen, and how these transformed scores are to be interpreted.

**Bias.** It is possible that factors such as sex, ethnicity or social class may act to obscure, mask or bias a person's true score on a test. If this is the case, the observed test score may not be an accurate or valid reflection of the quality assessed through the test. This has been a concern of test designers for a considerable time, and an entire body of psychometric research has been devoted to developing methods for evaluating whether a test score is biased against different population subgroups. Test manuals should state whether the test has been evaluated for potential bias, what methods have been used to carry out such an evaluation and the results obtained.

To sum up, what should be found in a test manual is clear evidence of the psychometric properties of the test showing how extensive the research supporting the test is (e.g. on how many people and in how many settings the information was collected), and how strong the research evidence is (i.e. the extent to which the test has been shown to be reliable, valid and free from bias). For this purpose, BPS Books publishes independent reviews of tests available in the UK.

Of course, all that has been provided in this section is an overview of the technical issues underlying test development, and a good understanding of test information requires more than a passing acquaintance with psychometric concepts and methods. This leads us to the question of competence in test use and to the next section.

## What qualifies as competence in the use of psychological tests?

Determining competence depends on two things; evidence of someone's performance in carrying out an activity, and standards against which to judge how well someone has performed the activity. Defining such standards for the use of psychological tests has been the main focus of the SCTS since 1987. To date, two sets of standards have been produced for the use of tests in the occupational setting. These are the Level A Standards which cover basic psychometric principles and the

skills required to use attainment and ability tests, and the Level B Standards which cover more advanced psychometric principles and the skills required to use tests of personality and interest. In addition to the standards themselves, an open learning pack has been developed for Level A, as well as guidance on how to assess someone's competence in the use of tests in occupational Settings. Standards are currently being developed for the use of tests in educational and clinical settings.

The standards developed by the SCTS require that, to be declared competent, an individual must be able to demonstrate knowledge and understanding of the psychometric principles underlying test construction, knowledge of the types of tests that are available, when it is appropriate to use them, and to be able to administer, score and interpret tests in order to provide accurate and meaningful feedback to others.

Once people have demonstrated their competence to the satisfaction of a Level A or Level B assessor, then they may apply to the Society for the appropriate Certificate and enter their names on the Register of Competence in Occupational Testing (RCOT). However, in order to obtain a certificate and enter the RCOT, an individual must have had their competence affirmed by a Chartered Psychologist who had their assessment practices verified by a verifier appointed by the Society.

So, what should someone look for as evidence of competence in test use? If services are being offered in occupational testing, then one should look for possession of the Society certificates: Level A for ability tests, Level B for personality tests, and Level A and B if both types of test are being considered for use. If someone is looking for advice in the use of tests, the Society recommends that they should consult a Chartered Psychologist as they are governed by the disciplinary procedures of the Society. The Society publishes *The Directory of Chartered Psychologists* through which the relevant expertise can be found.

Today, test suppliers require those seeking access to test materials to hold the Level A or Level B certificates. Other restrictions may apply to specific diagnostic tests (e.g. for clinical assessment) where more advanced training and experience are

required. But, what if a test has been used inappropriately? What recourse do individuals have? The approach of the SCTS continues to focus on prevention as the best cure, which leads to a commitment to responsible use of tests outlined in the next section.

## A commitment to responsible use of tests

The process of testing generally involves the following parties:

• **The Developer** who designs and develops the test.

• **The Supplier** who publishes and provides access to the test. The supplier might also be the developer, but it is common for suppliers to publish tests developed by psychologists independent of the supplying organisation.

• **The User** who administers, scores and interprets the test.

• **The Candidate** who is the person to whom the test is administered.

• **The Client** who is the person to whom the results from testing are reported. The client might also be the candidate, but in many instances the results for a candidate will be reported to a third party.

Each of these parties shares responsibilities for the process of testing. The developer and supplier share the responsibility for ensuring the quality of the test and for the adequacy of documentation provided for the use of the test. The user has the responsibility of ensuring that he or she understands why a client wants to use psychological tests, that testing is a suitable means of achieving the client's goals, and that the use of the tests and test scores are fair to the candidate.

These responsibilities are essentially why the SCTS was established to develop and monitor standards for test competence, and why the Society has codes of professional conduct for its members,

However, these responsibilities clearly focus on the developer, supplier and user and do not reflect the responsibilities of the client and the candidate. So what are

their responsibilities? The client has the responsibility of ensuring that those offering advice on testing and services in test use are competent to do so.

It is both the user's and the client's responsibility to ensure that the purpose of testing has been clearly communicated to the candidate, that the candidate understands the procedures that will be used for testing, how the test information will be used and to whom it will be communicated. As such, it is also the candidate's responsibility to ensure that he or she understands why the tests are to be used and to raise any concerns that he or she has in advance of testing.

In effect, testing is a social contract in which all parties should seek a common shared understanding of the process. At present, the only recourse that a client or a candidate has if they feel that a test has been used inappropriately is to raise the issue with an appropriate professional body such as the Society or to seek legal advice. As stated earlier, the Society believes that the best cure is prevention. To this end, we propose the following set of simple questions as a means by which each party can contribute to responsible use of psychological tests.

- ***The purpose of testing is clearly stated and communicated to all parties involved in the testing process.***

- What is the purpose of testing? What are the outcomes that will be achieved through testing?

- Why are these specific tests being considered or recommended? What evidence is there that these tests are relevant to the outcomes being sought? What evidence is there that these tests are appropriate for the people who are to be assessed?

- ***The procedures for testing are clearly stated and communicated to all parties involved in the testing process.***

- Who will administer the tests? What evidence is there that they are competent to administer them?

- When and where will the tests be administered? Is this a suitable environment for the administration of the tests?

- ***How the test information will be used is clearly stated and communicated to all parties involved in the testing process.***

- Who will score the tests? Who will interpret the scores? What evidence is there that the scorer / interpreter is competent to score / interpret these tests?

- How will the test score be communicated? What actions will be taken to ensure that the communication of test scores is accurate and meaningful?

- How will the confidentiality of the test scores be protected? Who will have access to the test scores? Why are they being given access to the test scores?

- ***Procedures for dealing with inquiries and complaints about the process of testing are clearly stated and communicated to all parties involved in the testing process.***

- Who will handle inquiries and complaints? Are they competent to handle inquiries or complaints?

- What actions will be taken in response to an inquiry or complaint? Will these actions ensure that the inquirer or complainant is treated fairly and ethically?

## Further information

This guide has provided an overview of the terms and issues involved in psychological testing. An information pack can be obtained from the Society's office on the Level A and Level B Standards, the Level A Open Learning Pack, and the Society's reviews of Level A and Level B tests.

Guidelines on testing are also available from other professional bodies such as the Institute of Personnel Development, IPD House, Camp Road, Wimbledon, London SWl9 4UX, Tel: 020 81971 9000.

The Society does not operate any form of accreditation or approval service with respect to publishers and distributors or to tests themselves. It is therefore unable to offer advice on the choice, use or origin of tests. Individual distributors publish their own catalogues of psychological tests, however, if a test is not listed the supplier can be found in the standard reference work that lists most tests in print, the *Mental Measurement Year Book* originally edited by the late Oscar Buros (available from good university libraries). A non-evaluative list of test publishers and distributors is available from the Society's office.

Information about occupational testing courses (level A and B) can be found in the Society publication *Selection and Development Review* issued six times a year. Details on subscribing to SDR can be obtained from the Society office.

The SCTS would be pleased to receive comments about this guide to help us inform others about the fair and valid use of psychological tests. Any comments and inquiries should be forwarded to this address:

The Chair, Steering Committee on Test Standards, The British Psychological Society, St Andrews House, 48 Princess Road East, Leicester LEI 7DR.

Tel: 0116 254 9568; Fax: 0116 247 0787; web site: www.bps.org.uk; e-mail: mail@bps.org.uk.

This document was pre pared for the Steering Committee on Test Standards by Eugene Burke, November 1995. The SCTS would like to thank all of those who kindly took the time to comment and suggest improvements to this guide.