

Chapter 15

PRINCIPLES AND APPLICATIONS OF EDUCATIONAL AND PSYCHOLOGICAL TESTING

Gerard J. Fogarty, PhD
Associate Professor of Psychology
Department of Psychology
University of Southern Queensland
Toowoomba, QLD, 4350

The main aim of this chapter is to introduce the reader to the concept of testing and to the principles that underpin the practice of educational and psychological testing. The first part of this chapter is devoted to that aim. A second aim of the chapter is to examine applications of testing under the headings of educational, industrial, and clinical. Examples will be introduced from the author's own work to illustrate various applications. A final aim is to discuss some of the issues and controversies that surround the theory and practice of testing. By the end of the chapter, the reader should have an understanding of the techniques used in test validation and should be familiar with some of the main applications of educational and psychological tests.

PART A: TECHNICAL REQUIREMENTS OF TESTS

Definition of tests

The word "test" refers to any systematic and standardised method of obtaining information about some aspect of human behaviour. The definition covers both educational and psychological tests and, to save space, the term "psychological test" will be used in most places in this chapter. A typical text book definition of a psychological test will usually cover the three defining characteristics (Murphy & Davidshofer, 1988):

- a psychological test is a sample of behaviour;
- the sample is obtained under standardised conditions;
- there are established rules for scoring, or for obtaining quantitative (numeric) information from the behaviour sample.

Thus, a psychological test is a sample of behaviour taken under standardised and highly regimented situations. For example, in the case of a paper and pencil intelligence or personality test, there is usually a set of instructions that have to be followed for every administration of the test, a set of instructions to be followed by the examiner regarding scoring and interpretation, and a set of guidelines covering the use of test results. It is important to follow the instructions exactly, whether in the administration, scoring, or interpretation phase. Any deviation from the standard pattern may cause a change in the test-taker's behaviour that might not otherwise have occurred.

Apart from the aspect of standardisation, which is intended to ensure that everyone is treated in the same way, the definition given above also emphasises the fact that any test is just a *sample of behaviour*. If you grasp that simple fact, you will understand much about testing. The whole point of sampling in any field is to select manageable subsets of elements and draw conclusions about the whole set from the sample. Thus, a quality control inspector may take a fistful of components from a bin at the end of a production line and make inferences about the quality of the whole production process by looking at the components in his or her hand. The assumption underlying the sampling technique is that the characteristics of the whole set will be reflected in the sample. In a highly automated production line situation, where components are manufactured by machines, the sampling process is relatively straightforward. For many aspects of human behaviour, however, obtaining a representative sample of behaviour is usually very difficult. Constructs such as personality, intelligence, motivation, interests, values, and knowledge cannot be observed directly and are extremely complex in their own right. Yet it is mostly constructs such as these that form the subject of educational and psychological testing.

Given the difficulty of the subject matter, it should come as no surprise to learn that an elaborate technology has been built around the practice of testing to ensure that the behaviour sampled reflects the constructs in which we are interested and that the testing instruments themselves are up to the task of accurately sampling this behaviour. The branch of science concerned with the development of educational and psychological tests is known as *psychometrics*. As the term implies, the main task of psychometrics is to measure psychological entities, such as what we know, how we feel, and what we think. The full range of techniques used by psychometricians to measure these complex processes requires a sound grasp of some maths processes and of statistics. We do not need to worry about the more difficult techniques here. However, some of the basic concepts of psychometrics are straightforward and absolutely essential for an understanding of the principles of testing.

Key technical concepts in educational and psychological testing.

As stated above, the area of educational and psychological testing has become highly technical and specialised with its own terminology and techniques. The two most important terms are *reliability* and *validity*. There is usually no point in administering a test that has low reliability and validity. In order to explain the concepts of reliability and validity, however, we need to backtrack a little and introduce a statistic that is used to assess both reliability and validity and is also used in many other applications of testing. That statistic is the correlation coefficient.

Correlation Coefficients

The psychometric properties of tests are often evaluated in terms of correlation coefficients. A correlation coefficient can take values from +1.00 to -1.00. A correlation of 1.00 between any two tests means that they are perfectly related. If you knew how well a person performed on one test relative to the rest of the group taking the test, you would know exactly how well they performed on the other test. For example, if a person topped the group on the first test, a correlation of +1.00 necessarily implies that the person tops the group on the other test. Conversely, if the person was at the bottom of the first test, he or she would be at the bottom of the other test as well. You would not know the person's score, a correlation does not tell you information about actual scores, but you would know the ranking of the person on the second test. Conversely, a correlation of -1.00 also indicates a perfect relationship but this time in an inverse manner. Thus, if a person came top of the group on one test that same person would necessarily be at the bottom of the group on the other test. The actual index of correlation is usually somewhere between these perfect extremes. The closer the index is to +1.00 or -1.00, the stronger the relationship between the tests. The closer to zero, the weaker the relationship until, at 0.0, there is no relationship at all between the test scores, or between that test and some criterion measure.

Apart from its role in assessing reliability and validity, which we will get to shortly, the correlation coefficient is extremely important in virtually all areas of psychological testing. Its popularity stems from the fact that the sample of behaviour obtained by administering a psychological test is often not the behaviour that we want to measure but is strongly related to it. Thus, the selection tests that job applicants are required to undertake usually contain tasks and questions that may not be encountered anywhere in the job itself. What is known about the selection tests is that performance on the tests is *correlated* with actual job performance. Someone who obtains a high score on the test is likely to do well on the job. Conversely, someone who does poorly on the test is likely to be a poor performer in the workplace. In order to be able to make these decisions, there must have been

a time when test scores and performance measures were available for a group of employees, so that the correlation between the two could be calculated. From that time onwards, the behaviour sampled by the test was used as a predictor of actual job performance. This situation has many parallels in educational, organizational, and clinical settings. Tests are used so widely because they sample behaviour that is related to behaviour in other settings. It is the correlation coefficient that is used to indicate the strength of this relationship.

Reliability

The reliability of a psychological test is often defined as the extent to which the scores on the test are free from error. That is, test reliability indicates the extent to which individual differences in test scores are attributable to "true" differences in the characteristics under consideration and the extent to which they are attributable to chance errors. For example, if the petrol gauge in your car gave wildly different readings each time you looked at it within a short space of time, you would begin to suspect that it was somewhat unreliable. The differences you are observing are not "true" differences. That is, the tank is not full one minute and half-full the next; these are "error" readings from the petrol gauge and we would say that it is unreliable. Reliability is usually, but not always, synonymous with consistency: the consistency of scores obtained by the same persons when reexamined with the same test on different occasions, or with different sets of equivalent items (Anastasi & Urbina, 1997). The following treatment of reliability theory aims, within the space of a page or so, to give you a basic understanding of its importance in test theory. For a more detailed treatment of this and other technical terms refer to a standard text such as Anastasi and Urbina (1997) or Murphy and Davidshofer (1988).

Reliability is usually assessed by examining aspects of the consistency of scores yielded by a test. Whilst this approach sometimes gives misleading results it has been adopted by most test constructors. Consistency measures of reliability fall into four kinds: parallel forms, test-retest, internal consistency, and inter-rater reliability.

1. **Parallel Form Reliability.** If two equivalent forms of a test exist, then it is possible to administer both forms to the same group of people and look at the correlation between them. If the correlation is very high, then both forms may be regarded as reliable. One has to exercise great care to ensure that the two forms are truly parallel with questions expressed in the same form, covering the same content, and containing items that cover the same range and level of difficulty. If one of the versions is less reliable than the other, the correlation between the two of them will be depressed. If the two forms are administered close

together, there may also be learning effects that transfer from one test to the next. Another major problem for this kind of reliability is that the effort involved in constructing a single version of a test is often very large indeed and few test producers have the resources to develop parallel forms. Examples of tests with parallel forms include the AL/AQ and ML/MQ tests of intelligence developed by the Australian Council for Educational Research (ACER). AL and ML both measure linguistic reasoning whereas AQ and MQ both measure quantitative reasoning. The AL and ML forms are parallel, as are the AQ and MQ.

2. Test-retest reliability does not require the existence of two versions of a test. Instead, a single test is re-administered to the same group of people after a short interval. Often this type of reliability is called a *stability index* because it reflects the extent to which individuals held their positions in the group over the two testing periods. High reliability does not mean that people obtain the same score but that they tend to maintain their position within the group. Once again, reliability is assessed by looking at the correlation between test and retest scores. Six weeks is often regarded as a satisfactory interval for establishing test-retest reliability. Much shorter than six weeks and there is the risk that respondents will remember answers given on the first occasion and that practice effects will occur. Much longer than six weeks and there is the risk that life experiences may have changed the scores on underlying traits in the intervening period. Six months is usually regarded as about the outer limit for test-retest correlations. This form of reliability is suitable for sensory discrimination and motor tests but a large number of educational and psychological tests cannot be administered twice over the short periods of time demanded by test-retest reliability.
3. Internal consistency reliability is a type of reliability that has some similarities to parallel form reliability. The simplest form of internal consistency reliability is called split-half reliability. Split-half reliability is obtained by dividing the items of a test into two equivalent halves. For example, the first half may consist of the odd numbered items and the second half of the even numbered items. The correlation between the scores from each half is taken as an index of reliability. An alternative method of estimating reliability from internal consistency uses the mean of all the possible split-half reliability coefficients that could result from different divisions of the test. Coefficient alpha is an example of such a coefficient and is probably the most widely used index of reliability.
4. Inter-rater reliability is relevant where interviews, observations, or open-ended questions are used. For example, the author has been involved in research with people with an intellectual disability (PWID). We have been assessing their stress, anxiety, and depression levels. Many PWID

cannot complete normal paper and pencil tests so we have used psychologists to conduct interviews with these people and make ratings of stress, anxiety, and depression. It was important to ensure that the psychologists were being consistent in their ratings. We were able to establish this by having three psychologists each interview and rate a small group of PWID. The inter-rater reliability was .87, which is close to the recommended minimum .90. Had the index of agreement been much lower, we would have had to abandon this method of assessment. Similarly, where people are being observed or where open-ended questions are used, an inter-rater reliability check should be conducted. This begins with the construction of an unambiguous coding system for every type of response. At least two raters then independently use the coding system to score a sample of responses and check that there is a high level of agreement.

If correlation indices are used to measure reliability, the index should be above .90 for inter-rater reliability. Apart from inter-rater reliability, there is no set figure for an acceptable level of reliability, although indices below .70 are generally regarded as unacceptable. Many sources indicate a lower bound of .60 for non-professionally developed tests and .80 for professionally developed tests.

Speeded tests - that is, tests which are designed so that most people cannot complete them within the specified time frame - present special problems for estimating reliability. Essentially, measures of internal consistency cannot be used unless one divides the test into sections and administers each section separately. Test-retest or alternate form estimates of reliability are used with speeded tests.

Standard error of measurement

The reliability coefficient is a helpful statistic when a judgement has to be made about the usefulness of a particular test. If the reliability is too low, say below .60, the test is probably not suitable for general use. The value of the reliability coefficient, however, is not confined to the situation where a choice has to be made between tests. Having made the choice, the reliability coefficient of the test chosen can also be used to give some indication of the confidence one can have in the score obtained for any individual. Because the reliability coefficient tells us the extent to which the scores on a test are free from error that is due to imperfect reliability, if you know the reliability of a test it is possible to set a band of tolerance around a given score and make estimates as to the likely error component. The reliability coefficient plays a part in this through the following formula:

$$SEM = SD_t \times \sqrt{1 - r_{tt}}$$

where SEM = standard error of measurement

SD_t = standard deviation of test scores*

r_{tt} = reliability coefficient for the test

* The standard deviation (SD) is a measure of how the scores are distributed around the mean or average score on the test. A small SD is an indication that most people have scored close to the mean, a large SD indicates that scores are spread more or less evenly across the whole range. The actual formula for calculating the SD is given below:

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{N - 1}} \quad \text{where } \bar{x} \text{ is the mean of the scores.}$$

The SEM can be treated in the same way as other standard error estimates: if you take a band plus or minus one SEM on either side of the obtained score, you can be about 68% sure that the true score lies somewhere within this band. Extend that band to include plus or minus 1.96 SEM's, and you can be 95% sure that the true score lies within this band. Thus,

upper boundary = score + 1.96 x SEM

lower boundary = score - 1.96 x SEM

A test with low reliability will have a large SEM, a reliable test, on the other hand, will have a narrow band. An illustration of a test report used by the author that includes SEM's is shown in Figure 15.1.

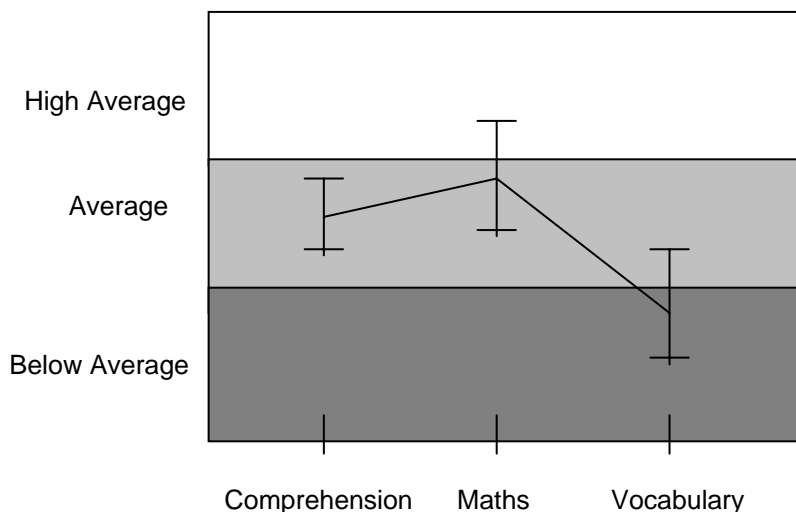


Figure 15.1. Illustration of a test report that includes standard errors of measurement

Test such as the ones shown in Figure 15.1 are very often administered as part of a job selection process. An educator or a supervisor looking at this chart can see what the person actually scored on the Comprehension, Maths, and Vocabulary tests and is also reminded that there is some error in the estimates. The size of the SEM bands gives a direct indication of how much faith we can have in the accuracy of the score. In this figure, the Comprehension test is quite reliable and has a rather narrow SEM. The SEM bands for the Maths and Vocabulary tests are somewhat broader, reflecting a lower reliability estimate for these tests.

The SEM statistic is much used in psychological and educational measurement for the interpretation of individual scores. Most published tests have SEM's listed in the test manuals. Unfortunately, many people still do not know how to interpret them.

Validity

Validity is another important psychometric characteristic of a test and it is often defined by the question: "Does the test actually measure what it is intended to measure?". There are four broad means of establishing a test's validity: content-related validation, face validity, construct-related, and criterion-related. Let's look at each of these in turn.

1. **Content validity.** When constructing tests, the test items should be representative of the behaviour domain to be measured. An arithmetic test, for example, should contain items that are representative of the content area, factorially pure (e.g., not contaminated by other factors such as ability to read instructions), and arithmetic in nature. Test-specifications help to achieve this. These define the content areas, the weightings, and the objectives. Thus, a test at the end of a training course should be related to the material learned in the course and should not overemphasise areas that did not receive much attention in the course.
2. **Face validity.** Face validity refers to the extent to which a test *looks as though* it measures what it was designed to measure. Whilst face validity has no scientific or technical basis, it must not be overlooked if a test is to be accepted by applicants. For example, if a particular position in a manufacturing firm required numerical skills and a selection test was required to test these skills, it would be wise to use questions that are based on the materials of the workplace. If the firm makes spare parts and employees are required to calculate the prices of different combinations of these parts, it is worthwhile constructing questions that use these examples. There may be existing tests with excellent reliability that ask the same types of questions using different objects (e.g., apples

and oranges) but job applicants are more likely to see the relevance of the context-sensitive questions and therefore show greater motivation. Face validity is a weak kind of validity in the sense that it is to do with appearances rather than more fundamental considerations such as whether the test measures what it purports to measure. However, appearances can determine reactions to a test and although it is the weakest form of validity, it is unwise to dismiss face validity as completely trivial.

3. **Criterion validity.** Procedures used to establish criterion validity indicate the effectiveness of a test in predicting an individual's performance in specified activities. Test scores are correlated with actual performance on some independent criterion. The most obvious example would be the relationship between performance on job selection tests and actual performance on the job itself or on a job-related training course. The validation could be *concurrent* or *predictive*. In the first situation, validity is established by administering the test to those for whom some criterion measure is already available. In the second situation, validity is established by first testing then matching against the criterion some time later. Thus, if we were to test a group of computer programmers and correlate the results with supervisor's ratings of work performance, this would constitute a measure of concurrent validity. If, on the other hand, we were to test a group of newly-hired computer programmers and later obtain measures of on-the-job performance for these same people, the correlation between test scores and job performance would constitute a measure of predictive validity. In both cases, a high correlation coefficient would indicate that the test has good criterion validity.
4. **Construct validity.** Construct validity is more abstract than the other forms of validity. It reflects the degree to which a test measures some theoretical construct or trait. To some extent, content validity and face validity also deal with this same aspect of a test and the reader may find the overlapping terms confusing. The terms do overlap but their meanings can be separated. Content validity refers to the material that is included in the tests. A test of mathematical ability should include mathematical questions. A subject matter expert should be able to judge whether or not a test has content validity. Face validity refers to the extent to which the test looks as though it is measuring what it should be measuring. Anyone can make such judgements but they may be wrong and psychometricians do not place much value on face validity, sometimes referring to it in a derogatory fashion as "faith validity". For example, people may expect a test of mathematical ability to contain mostly symbols and equations and to rate it low on face validity if it does not contain a high proportion of such items. However, many

mathematical problems can be couched in verbal terms (e.g., word algebra problems) and a test that contains such questions may still have excellent content validity. Unlike content validity and face validity, construct validity cannot be judged by simply looking at a test. Rather it requires the application of statistical techniques to determine what is being measured by the test. Some of the techniques contributing to construct-related validation are as follows:

- a) **Developmental changes.** Some tests, for example the *Stanford-Binet* (*SB*) test of intelligence, assume changes in performance will occur with age. The basis of test construction is that the *SB* consists of a number of sub-scales, which are like mini-tests assessing different areas of performance, e.g., vocabulary, spatial ability, comprehension, and so on. Within each sub-scale, items are grouped according to age bands: a group of items that the typical two-year old can solve, followed by a group that the typical three-year old can solve, and so on, right through to adult level. In this way, the items become increasingly more difficult. The *SB* is administered on an individual basis with different starting and finishing points for each person. A person will complete an initial test that determines on which level he or she will start. If the person gets these items right and the items right on the next level as well, it is assumed that all of the earlier items would also have been correct and credit is awarded for them. The person then moves through the higher levels of the sub-scale until failing the majority of items at two consecutive levels. It is then assumed that all subsequent items would also be failed and the test is discontinued. Clearly, it is most important that the test constructors have not mixed up the ordering of items. If they have, then the test is invalid. In the case of the *Stanford-Binet*, the first edition of the test appeared in 1904 and there is an enormous bank of data that can be used to justify the ordering of items.
- b) **Factor analysis.** In general, this highly mathematical approach involves calculating the correlations among a set of tests and looking for patterns that suggest some tests "go together", so to speak. If such patterns do exist, the tests that "go together" will define a factor and the statistical packages used to conduct factor analyses will show the correlations of each test with its factor. The correlation of each test with its relevant factor is referred to as the factorial validity of a test. Factor analysis can also be applied at the item level where observations of high intercorrelations among sets of items indicate that they are measuring something in common, hopefully what they were intended to measure. Factor analysis is also discussed in the chapter on intelligence, including an example of its application in that

field. A brief example follows here.

The example comes from research the author conducted with a colleague on the structure of learning styles among adult learners (Fogarty & Taylor, 1997). We used a 30-item test called the *Approaches to Studying Inventory* (ASI; Entwistle, 1983). The ASI contains seven subscales which, as mentioned above, are like tests within a test. Each of the sub-scales is designed to measure one of seven different learning orientations: Achieving, Meaning, Comprehension, Operation Learning, Reproduction, Improvidence, and Globetrotting. The first four of these measure aspects of what might be called a "deep" approach to learning, the last three measure aspects of a "shallow" approach. In terms of what was discussed above, one might expect that these two underlying constructs would be identified in a factor analysis. The table of correlations among the sub-scales of the test is shown below.

Table 15.1
Correlations among *Approaches to Studying Inventory* Subscales

| Sub-scale | 1 | 2 | 3 | 4 | 5 | 6 |
|-----------------------|------|------|------|------|------|------|
| 1. Achieving | 1.00 | | | | | |
| 2. Meaning | .47 | 1.00 | | | | |
| 3. Comprehension | .42 | .52 | 1.00 | | | |
| 4. Operation Learning | .42 | .43 | .41 | 1.00 | | |
| 5. Reproducing | -.06 | -.11 | -.12 | .02 | 1.00 | |
| 6. Improvidence | .09 | .06 | -.08 | .03 | .38 | 1.00 |
| 7. Globe Trotting | -.08 | -.08 | -.17 | -.08 | .37 | .35 |

It is quite apparent from Table 15.1 that the four sub-scales measuring a deep approach to learning tend to correlate among themselves and not to correlate with the remaining three sub-scales, which measure a shallow approach to learning. The three sub-scales measuring a shallow approach show the same tendency to correlate among themselves. Patterns of correlations like this suggest that these seven sub-scales are measuring two unrelated constructs, which we can call factors. A factor analysis will tell us whether or not this is the case and just how well each test measures the underlying construct it was intended to measure. Table 15.2 shows part of the output from a factor analysis of the correlation matrix in Table 15.1.

Table 15.2
Factor Analysis of *Approaches to Studying Inventory*

| Sub-scales | Factor 1 | Factor 2 |
|-----------------------|----------|----------|
| 1. Achieving | .65 | |
| 2. Meaning | .74 | |
| 3. Comprehension | .67 | |
| 4. Operation Learning | .61 | |
| 5. Reproducing | | .61 |
| 6. Improvidence | | .64 |
| 7. Globe Trotting | | .57 |

The mathematics used to generate the figures shown in Table 15.2 need not concern us. What is apparent is that the first four sub-scales of the ASI are related to Factor 1 (Deep) and the last three sub-scales relate to Factor 2 (Shallow). This is what Entwistle intended. On the basis of this study, we can say that there is support for the factorial validity of each of the sub-scales of the ASI.

- c) Convergent and discriminant validity. Support for the validity of a test can also be obtained by showing that it correlates highly with variables with which it should be correlated and also by showing that it has no relationship with variables with which it is not expected to have a relationship. The former is called convergent validity, the latter, discriminant validity. In the example reported immediately above, if there was another well-validated measure of deep and shallow processing, it would be possible to test whether the ASI measure of deep processing has convergent validity by examining its correlation with its counterpart. The correlation should be high. Its discriminant validity could be ascertained by checking to see that the correlation between the ASI measure of deep processing is uncorrelated with the measure of shallow processing from the second test (Entwistle argues that deep and shallow constructs are uncorrelated). The same checks could be applied to the ASI measure of shallow processing.
- d) Known groups validity. This aspect of validity is demonstrated by showing that test scores differentiate, in a predictable manner, between groups of test-takers known to differ on the characteristic being measured. Known groups validity is similar in some respects to concurrent validity. Thus, a test of honesty might be expected to discriminate among criminals (low scores), politicians (moderate scores), and clergy (high scores). Similarly, a test of mental toughness might be expected to differentiate among athletes of

various levels. A test developed to measure sex-role perceptions should demonstrate different average scores for groups of males and females. Failure to find expected differences would raise doubts about either the construct of sex-role perception, the construct validity of the test, or both.

Which type of validity is most important?

Apart from face validity, it is a mistake to think that one type of validity is more important than another. It depends on the purpose of the test. If the purpose is to select good employees, predictive validity is all-important. If the purpose is to assess performance in a course or training programme, content validity is very important. If the purpose is to develop models of how different constructs relate to each other, construct validity is paramount. In some situations, it will be necessary to demonstrate that all forms of validity are satisfied.

By way of illustration, the author was once involved in the development of a selection test for sales personnel. The early stages of the project required a thorough search of the literature to determine just what aspects of ability, personality, and interests were related to success in sales. When these were determined, the test construction process began. One of the early decisions made by management regarding the test was that it had to "look the part". That is, it had to have face validity, not just in relation to the seeming appropriateness of the questions but also in relation to things like the physical appearance of the test. A second consideration for the test developers was that the questions had to sample various content domains. There was a need for questions on numeracy and literacy skills, some questions on various aspects of personality, such as extraversion and persistence, and quite a range of questions tapping demographic characteristics such as age, previous employment, and so forth. These were issues of content validity. When a draft set of questions had been developed, they were trialled on a group of salespersons that included both high performers and low performers. The aim was to establish the concurrent validity and also the known-groups validity of the test by showing that it discriminated between these two groups. The trial also resulted in valuable feedback about the acceptability of the test (face validity). Finally, the test was included in the selection process, allowing the accumulation of a large dataset that included scores on both the test and later sales performance. The predictive validity could then be established.

In the process of validating this selection test, some of these stages were revisited a number of times during the early years of its operation. The validation process should not stop once the test has been implemented. Circumstances do change and it will be necessary to keep checking all

aspects of reliability and validity throughout the lifespan of the test. Test validation is most intensive in the construction and implementation phases, but it is an ongoing activity.

Relationship between reliability and validity

Reliability and validity should not be thought of as just desirable features of a test, they are both essential. The relationship between the two is best depicted in diagrammatic form. Figure 15.2 shows the relationship using the analogy of a target board.

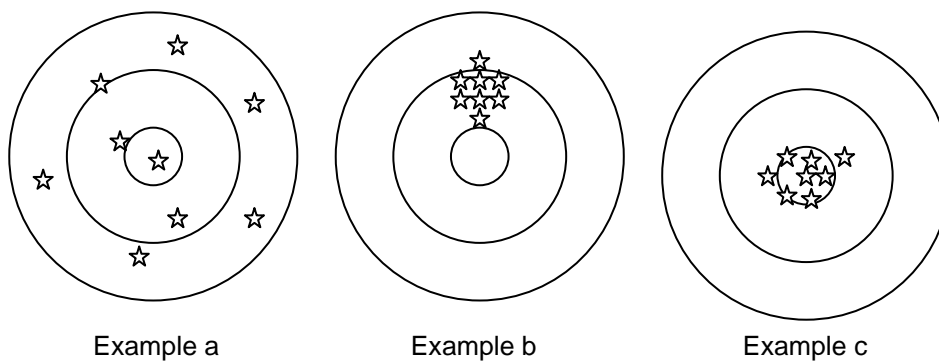


Figure 15.2. The relationship between reliability and validity

Example (a) shows what an unreliable test would look like. Example (b) shows what a reliable but invalid test would look like. It is similar to a rifle that has its sights mis-aligned. The high degree of reliability is shown by the consistency of the strikes. The lack of validity is shown by the fact that the missiles are missing their target, the bullseye. For example, a job satisfaction test given to unskilled workers may measure literacy skills rather than job satisfaction if the test is written in complex language. In psychometric terms, the test is not measuring what it was intended to measure. Example (c) is what a valid and reliable test would look like: the missiles hit the mark and they hit it consistently.

Different Types of Tests and Their Applications

Most psychological tests can be sorted into three general categories (Murphy & Davidshofer, 1988):

- a) Performance tests in which the test-taker performs some specific task, such as writing an essay, answering multiple choice items, or mentally rotating images presented on a computer screen;
- b) Behaviour observation tests that involve observations of a person's behaviour within a particular context;

- c) Self-report measures, in which the test-taker describes his or her feelings, attitudes, beliefs, interests, and the like.
The following paragraphs will expand upon each of these categories.

Tests of performance

One of the most familiar testing situations is one in which participants are given some well-defined task and asked to do their best within a given time frame. The score is usually the number of items that the participant has scored correctly in that time. Cronbach (1970) referred to this as a "test of maximal performance". To illustrate the nature of a test of maximal performance, Ackerman and Heggstad (1997) gave some rather delightful examples of instructions to early examinees around the turn of the century. For example, they cited the description of test procedures given by Binet & Simon (1911-1915), the developers of the *Stanford Binet* intelligence test:

What should be done [if a child does not respond]? The help of the teacher is often useful. If she is intelligent, she knows what to say to her children to reassure them and arouse their courage. A caress to one, a reprimand to another and all goes well. [Cited in Ackerman & Heggstad, 1997, p. 220].

Contrast those instructions with the set given for testing US army recruits during the First World War:

When everything is ready E. (sic) proceeds as follows:
"Attention! The purpose of this examination is to see how well you can remember, think, and carry out what you are told to do... You are not expected to make a perfect grade, but do the very best you can.

Now, in the army a man often has to listen to commands and then carry them out exactly. I am going to give you some commands to see how well you can carry them out. *Listen closely. Ask no questions.*

[Cited in Ackerman & Heggstad, 1997, p. 221]

Both sets of instructions, contrasting though they may be, make it very clear that in these intelligence-testing situations, the individual is expected to demonstrate maximal performance.

Behaviour observations

Some psychological tests involve observing the examinee's behaviour and responses in a particular context (Murphy & Davidshofer, 1988). Tests of this type usually involve an examiner noting the typical behaviour of the individual in a particular situation. Teachers may make assessment of the social skills of children in classrooms by observing how they behave to other children. Employers in job situations may set up typical job simulations and observe how job applicants handle the work in those situations. Many training situations involve highly structured observations by trainers of trainees undergoing particular tasks. Trainee teachers are subjected to regular inspections in a classroom. These situations are a lot less standardised than those described earlier but provided that the instructors are using a checklist of behaviours and recording the student's behaviour using some standardised format, then it is a testing situation. Reliability is usually assessed by checking inter-rater agreement.

Self reports

The final class of test includes a variety of measures that require the examinee to report or describe his/her feelings, attitudes, beliefs, values, opinions, or physical or mental state (Murphy & Davidshofer, 1988). Many personality inventories take this form. It is a very efficient form of data collection and large numbers of respondents can be tested at the same time. Self-report techniques do have a number of drawbacks, which will be dealt with later in this chapter.

Interpretation of test scores

Whatever the form of test, the result will be a score or rating of some kind. The score or rating can provide two types of information. One type is the relative standing of the individual in relation to his or her peers. Such measures are called *norm-referenced*. A second type of measure that can be obtained from tests of maximal performance reflects the degree to which the individual has mastered the skills that characterise the domain being tested. Such scores are called *criterion-referenced*. Other names used for this type of reference framework include content-, domain-, and objective-referenced.

Criterion-referenced testing

A very simple example will suffice to illustrate the difference between the two treatments of a test score. A driver's licence test is a criterion-referenced test. The score obtained does not indicate how well you went compared with everyone else who sat the test but how well you can drive. Thus, the emphasis is on what you can do. The driver's test is a very simple format, you either pass or fail. Some criterion-referenced tests have

many levels. Criterion-referenced testing became popular in the 1970's, especially in the field of education. The advent of computerised instruction systems made it relatively easy to test what an individual knows and to adapt instruction accordingly. Repeated testings indicate the level of mastery attained and the learning modules that should be presented next. At no stage is there any comparison with other individuals, the focus is entirely upon what the learner knows and what the learner can do. A score in this context reflects the level of attainment. Anastasi and Urbina (1997) commented that this form of test score interpretation is best suited for basic skills where instructional objectives can be arranged in an ordinal hierarchy, the acquisition of more elementary skills being a prerequisite for the acquisition of higher level skills (p. 77). Beyond the basic skill level it is extremely difficult to arrange skills in such an ordered sequence and norm-referenced testing, or a combination of norm-referenced and criterion-referenced testing, is more appropriate.

Norm-referenced testing

A much more common form of norms employs a quantitative approach to show the position of an examinee in relation to his or her peer group. The crudest of these simply indicate whether someone is above or below average. For example, an individual may be described as above average on a test of intelligence, or neuroticism, or self-confidence, or whatever it is that is being measured. A more elaborate and quite common scheme uses five intervals corresponding to the top 10%, the next 20%, the middle 40%, the next 20%, and the bottom 10% of the population. University grading systems often use schemes like this.

Percentile scores represent an improvement on the five-fold classification scheme. Percentile scores use a scale numbering from 1 to 100 to tell us where an individual is located on a test. Thus, a percentile score of 45 indicates that 45% of the population obtained this same score or a lower score. A percentile score of 99 means that 99% of the population were equal to or below this score. Percentiles are easy to understand and are found in nearly all test manuals. Unfortunately, they are somewhat distorted near the ends of the distribution. That is, a small difference in raw scores in the middle of the distribution can lead to a big difference in percentile scores whereas the same difference towards the tails of the distribution may result in only one or two percentile points difference.

To overcome this problem, psychologists and educators frequently use norms that are based on the normal frequency distribution. There are four main types of norms based on this distribution: z scores, quotients, stens, and stanines (see Anastasi & Urbina, 1997, pp. 61-66). Z scores report an individual's scores in terms of how many standard deviations the score is

from the mean. The formula is easy enough: simply subtract the mean from the score and divide the result by the standard deviation. In virtually all cases, this will result in a z-score that ranges somewhere between -3.00 and +3.00. Z scores are not often used because they can be negative as well as positive and many people do not like dealing with negative test scores! To overcome this problem, some test manuals use quotients. A quotient is a standard score with the mean set at 100 and the standard deviation set at 15. IQ scores are always reported in this form. A raw score is obtained on the test and then converted to an IQ score on the basis of tables given in the test manual. The term "IQ" comes from Intelligence Quotient and is a leftover from earlier times when intelligence was assessed by forming a ratio between mental age and chronological age. The ratio is no longer used but IQ has come to stand for scores on tests of general intelligence. T scores are similar to IQ scores, but are based on distribution with a mean of 50 and a standard deviation of 10. Sten (standard ten) scores are based on a distribution that has a mean of 5.5 and a standard deviation of 2, with 10 discrete scores being possible. The decimal point has made sten scores somewhat unpopular. Stanines (standard nine) scores are based on a distribution with a mean of 5 and a standard deviation of 2, with nine discrete scores being possible. They can be obtained from z scores by multiplying the z score by 2 and adding 5. Stanine scores are used in many test settings. Both stens and stanines operate by banding together scores in a specific region of the distribution and using the same number to represent all scores falling within that band.

Constructing and validating a test

This final section on technical matters will attempt to bring together much of what has already been discussed by tracing the development of a test in which the author has been involved. The description will cover all aspects of the development process, from the rationale right through to the point where publication of a test manual is possible.

Overview of test development process

Test construction involves first deciding the broad domain to be covered by the test. Once it is clear what has to be covered by the test, the process of item construction begins. The idea is to develop as many items as possible because some of them will not pass the various filters. The first filter is normally a group of experts who will make judgements about the face and content validity of the items. Some will be discarded at this point. Surviving items are generally tested with a representative sample drawn from the population for whom the test is intended. This testing will make it clear whether there are any major problems with intelligibility, clarity, and

appropriateness of items. If the test is one of maximal performance, pre-testing will help to ascertain whether the items are too easy or too difficult for the intended population. Figure 15.3 shows examples of tests that are a) too easy, b) too difficult, and c) of an appropriate difficulty level.

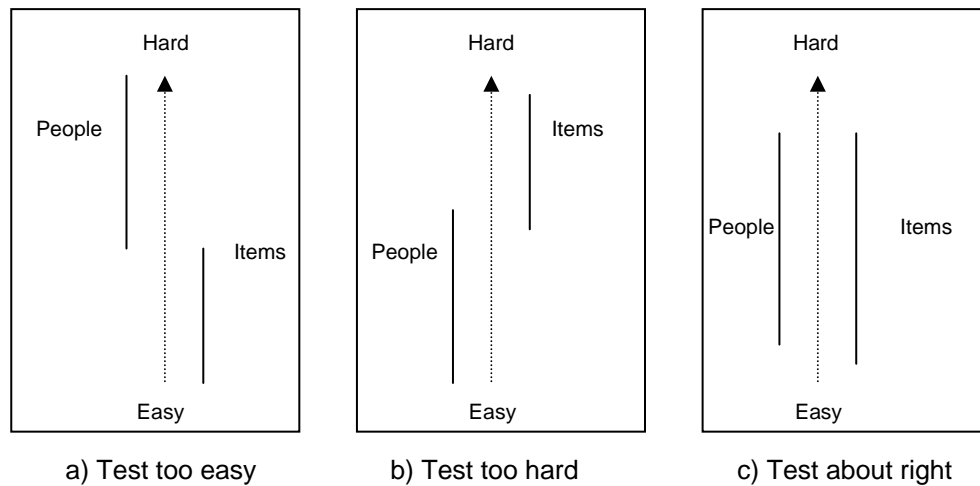


Figure 15.3. Diagrams depicting suitability of test difficulty

Figure 15.3 shows what are often referred to as item-person maps. The dotted line in the middle represents a difficulty continuum with low difficulty and low ability at the bottom of the line. The "People" line represents the ability range of the people taking the test whilst the "Items" line represents the range of difficulty of the items. Item difficulty can be something as simple as the proportion of people who made an error on the item (typically this is not how it is measured but it will do for our purposes). The item-person map shown in example (a) indicates that the items are too easy for this population. In example (b) they are too hard. There would not be any point in giving either of these tests to the populations sampled here. In example (a) everyone would get all the items right and in example (b) everyone would get every item wrong. Example (c) shows the type of item-person correspondence test constructors are trying to achieve.

Item-person maps can also be constructed for self-report tests where there are no right or wrong answers but where people have to judge the extent to which they agree or disagree with statements posed in the items (e.g., Tenenbaum & Fogarty, 1998). It is just as important in these situations to ensure that some of the respondents agree with the statements and some disagree. If all respondents find it too easy to agree with the statements (or to disagree), everyone will end up with almost the same score, and that is a

highly undesirable situation in testing. After all, one of the aims of any test is to discriminate among individuals. This is not possible if everyone has close to the same score. A technique known as Rasch analysis is particularly useful for making these kinds of analyses (Tenenbaum & Fogarty, 1998).

If the test for students passes these preliminary filters, it is time to use it with a larger sample. The reliability of the test can then be assessed. Some further item changes may be required after these analyses. The final stages of test validation will involve checking relationships between scores on the test and external measures (concurrent and possibly predictive validity) and also checking whether the test appears to be measuring the constructs it was intended to measure (construct validity).

An example of test development and validation

To clarify the process of test development further, consider the steps involved in the validation of a test intended to measure stress in people with an intellectual handicap (Bramston & Bostock, 1994; Bramston & Fogarty, 1995; Fogarty & Bramston, 1997).

1. Initial item construction. Original items for this test were derived from 89 adults with mild intellectual disabilities who took part in one of 22 brainstorming groups set up across two Australian states. Participants discussed what aspects of their day-to-day lives upset, bothered, worried, and stressed them. Twenty-four people who worked closely with people with intellectual disabilities completed similar brainstorming exercises and their responses were added to the pool of ideas. After duplications were removed, 60 stressors remained that were then worded into interview questions. The test was then tested widely among adults with mild intellectual disabilities and reviewed for content and clarity of item wording by a panel of 6 people consisting of academics (2), teachers in the disabilities field (2), and parents of adults with intellectual disabilities (2). Based on their responses and the trials, the items were revised and the test reduced to 31 items.
2. Test administration. The resulting *Lifestress Inventory* is a self-report test administered by interview. Respondents are asked to acknowledge if any of the stressful events listed in the test had occurred in the last fortnight (e.g., "Have you argued with anyone recently?"). To counter any tendency towards acquiescence, the stressful option was "yes" for half the items and "no" for the remainder. If the stressful option was indicated by the response, the respondent was asked to rate how stressful the event currently was by pointing to a spot on a 4-point Likert test. The rating test clearly set out the numbers "1-4", written descriptors from "not stressful" to "a great deal of stress", and graphic

representations of each point using buckets with varying amounts of water in them. Normally the graphic representation would not be necessary, but with this population, extra care was taken to ensure that they understood what was required of them. The interviews were conducted by one of two interviewers, both aged in their late 30's, one male and one female. Both were experienced clinicians and familiar with the test. It had already been established that the inter-rater reliability for trained interviewers was .87. Respondents were helped to feel at ease and then asked if they would simply say whether certain events or issues had occurred in their life within the last two weeks. The interviewers were required to verify subject responses with the prompt "Tell me more about that", so that unreliable responses could be detected and scored accordingly. The interviews generally took about 15 minutes each.

3. Validation of test structure and cross-validation. The next stage (Bramston & Fogarty, 1995; Fogarty & Bramston, 1997) involved the analysis of the data, including a factor analysis that identified three subtests of the Lifestress Inventory: general worry, interpersonal concerns, and concerns with coping. Another study was then conducted to refine weak items and to confirm the construct validity.
4. Further validation and refinement of test. The next part of the validation work consisted of yet another administration of the test to a different sample, this time a clinical psychologist also interviewed respondents and ratings of stress were collected from their work supervisors for the purposes of concurrent validation.
5. Production of test manual. The last stage involves the production of a test manual that will contain a complete description of the test, psychometric data relating to reliability and validity, and tables showing typical scores for males and females and different age groups. In other words, norms will be constructed so that it is possible to tell whether a particular score is high or low in relation to the general population of people with an intellectual disability.

If the above process seems rather exhaustive, it is a reminder that it is an extremely painstaking process to put together a test that is reliable and valid. Such projects should not be undertaken lightly.

PART B: APPLICATIONS OF TESTING

It is not possible to describe the full range of testing applications in one book, let alone in a chapter, but it is relatively easy to identify the main types of test in current use and to describe some of the purposes for which they are used. The three main types of test to be considered here are tests of intelligence, tests of achievement, and tests of personality.

Tests of intelligence

Modern intelligence tests can be classified into two categories. The first category contains what are called individual tests of intelligence. The best known of these are the *Stanford Binet*, now up to its fourth edition, and the *Wechsler Adult Intelligence Scale (WAIS-III)*, the fourth version of which has just been released in Australia. Both of these take anything up to 2 hours to administer, although the time taken is typically less. Test items cover a range of abilities employing both verbal and non-verbal item types. Test takers may be asked to give the meanings of words, to complete a series of numbers, to recall a list of numbers, and so on. Both tests yield overall estimates of intelligence (IQ) and estimates of a range of specific abilities. Because they are administered in an interview situation, both tests are capable of yielding a lot of clinical information as well. The Fourth Edition of the *Stanford Binet* can be used with both adults and children. The *WAIS-III* is similar in format and style to the *Stanford Binet* but comes in different versions: a) an adult version (*WAIS-III*); b) a children's version (*WISC-III*); and c) a preschool version (*WPPSI*). Both the *Stanford Binet* and the Wechsler tests are updated from time to time, with new revisions subjected to thorough validation procedures. There are other individual tests of intelligence, some that are arguably more reliable and more valid for certain purposes, but none that would match the *Stanford Binet* or the Wechsler tests in terms of popularity.

The second category consists of the group tests of intelligence. These tests were first developed to handle the very large number of soldiers recruited into the US army during the First World War. Results helped with placement and classification decisions. Most group tests of intelligence now yield scores on a broad range of abilities, such as vocabulary, numerical ability, spatial ability, memory, reasoning, and many others. Testing has become a very commercialised enterprise and major test publishers offer a variety of group tests of intelligence. Examples of some ones include the *Differential Aptitude Test (DAT)*: Bennett, Seashore, & Wesman, 1989) and the Australian Council of Educational Research (ACER) *Advanced Tests AL-AQ* and *BL-BQ* (ACER, 1982). Such tests can be administered to many people at the same time, scoring is generally easy, and norm tables can be

compiled without any great difficulty. For this reason, there tends to be many more group tests than individual tests. Some group tests of intelligence have also been developed for special populations, such as those with language difficulties or hearing or sight impairment.

Applications of tests of intelligence

Intelligence tests, whether group or individual, have been in widespread use since the start of this century. They were developed initially for use in educational settings but quickly found their way into occupational and clinical settings.

Tests of intelligence in clinical settings

Tests of intelligence are routinely used in clinical settings. Level of intellectual functioning provides insights into general level of health. The fact is that a lot of the problems that are referred to a psychologist or psychiatrist either have their origins in intellectual weaknesses or can be better understood following a diagnosis of the individual's intellectual strengths and weaknesses. Invariably, the tests used are individual tests, such as the *Stanford Binet* or one of the Wechsler tests. A major use of intelligence tests in clinical settings is in neuropsychological assessment, where the aim is to assess possible brain damage as a consequence of trauma, usually caused by a car accident.

Tests of intelligence in educational settings

Intelligence tests were originally developed to measure learning potential, something that they still do very well. The author has for many years used measures of numerical ability, verbal comprehension, vocabulary, and abstract reasoning to predict educational achievement. There is no doubt that tests such as these give a fairly accurate indication of success in various subject areas. That is, they have good predictive validity. Used in conjunction with measures of actual academic performance, intelligence tests can help guide people towards appropriate career choices. Although not as popular as they were in the 1950's and 1960's when most school children underwent IQ testing, intelligence tests are still very much part of the educational environment. They are used as the basis for awarding scholarships, for gaining entrance to some prestigious courses (especially in the United States), and they are widely used for diagnostic assessment where learning difficulties are suspected.

Tests of intelligence in occupational settings

The role of intelligence testing in occupational psychology was summarised in a review by Hunter (1986) who pointed out that although

intelligence testing has not been as successful in the occupational field as the educational field, it nevertheless predicts a reliable proportion of job performance, and it does so better than alternative measures, such as interviews or personality assessment. The relationship between intelligence and job performance, however, depends very much on the individual's familiarity with the job. In the early stages when there is a lot of learning occurring, tests of intelligence predict performance quite well, probably because performance is closely linked with the ability to learn rapidly. Once the individual has settled into the job, however, the strength of the relationship between intelligence and job performance starts to decrease. In jobs which impose variable demands and where learning is constantly occurring, intelligence tests will prove more useful for predicting performance.

Furthermore, there are many occupational settings where the tasks are quite complex and in these situations intelligence tests can be useful. Indeed, with the increasing complexity of modern day work situations, it is possible that the predictive validity of intelligence tests will increase in occupational settings. The introduction of automation is a familiar scene everywhere in the workplace. Tasks that were once performed by manual labour are now being completed by a machine. As this happens, the job requirements are shifting from physical strength and motor coordination to cognitive dexterity. In Reich's (1991) terms, we are moving from a world of doers to a world of symbol analysts. The new technologies devalue experience and increase the value of the ability to learn (Hunt, 1995), precisely the sort of thing that is predicted by tests of intelligence. However, the trend is not completely in the direction of greater complexity. Some jobs that formerly required cognitive skills no longer do so because a machine (e.g., a calculator) now takes care of the cognitive work. Time will tell whether intelligence becomes more or less important in the workforce of the future. For a thorough analysis of this issue, see Hunt (1995).

Tests of achievement

For most people, the most commonly experienced tests are the ones that we sit as students in educational institutions or as adults seeking professional or trade qualifications. These so-called achievement tests are designed to measure the effects of a specific programme of instruction or training (Anastasi & Urbina, 1997). They usually take the form of either free-response questions such as essays, or objective questions such as the popular multiple choice format. A problem with many achievement tests is that they are never standardised or validated in the manner suggested in this chapter. It is not hard to see why: most people who construct achievement tests have neither the time nor the expertise to undertake the necessary

analyses. Free-response format tests, for example, should really be checked for inter-rater reliability to make sure that different subject matter experts rate the answers in the same way. For reasons mentioned above, this rarely happens. Multiple choice tests are a different story: software is readily available to score these tests and at the same time give valuable feedback about questions that are unreliable and therefore decreasing the reliability of the whole test. If a test is unreliable, it cannot be valid. Unfortunately, it is probable that many constructors of achievement tests do not even use test specifications when selecting questions for inclusion. At the very least, the specifications should take account of the objectives, the content areas covered, and topic weightings.

Having said this, there are excellent examples of achievement tests that are properly standardised and validated. The *Progressive Achievement Tests (PAT)* published by ACER are widely used in Australia to measure a student's level of attainment in key academic areas such as vocabulary, comprehension, and mathematics. Test norms are available, so it is possible to see how a student compares with other students throughout Australia. Tests like this can be extremely helpful for designing curricula and making decisions about which students can be directed to extension classes and which ones might benefit from supplementary work.

Personality Tests

In terms of widespread usage, the assessment of personality ranks second only to intelligence and achievement testing. There are two basic forms of personality testing: self-report measures and projective techniques such as the *Rorschach* and the *Thematic Apperception Test*. The *Rorschach*, better known to most people as the inkblot test, is one of the earliest forms of personality assessment, having first made its appearance in 1921. The test presents a series of 10 stimulus cards to the test taker, who is required to state what he or she can see in the card. The theory upon which the test is based claims that the way a person perceives and interprets the test material reflects fundamental aspects of his or her psychological functioning, including personality. The *Thematic Apperception Test (TAT)* also makes use of pictures, but employs them in a different way. A series of 19 pictures and one blank card is shown to the test taker who is asked to make up a story about each picture. In the case of the blank card, the task is to imagine a picture on the card and then tell a story about it. The rationale underlying the use of the *TAT* is much the same as that for the *Rorschach*; there is an expectation that people will project much of themselves into the stories they tell. However, although projective techniques are powerful tools for personality assessment they are not used by many practitioners. They take a lot of time to administer and a lot of training before reaching a reasonable

degree of proficiency. Self-report methods of personality assessment have proven to be much more popular.

As the term implies, self-report tests rely upon the test-taker responding to a set of standard statements by indicating whether they agree or disagree with the statements (if the answer is a simple yes-no) or choosing a number to indicate the extent of their agreement or disagreement with the item. There are so many self-report forms around these days that it is extremely unlikely that the reader has not encountered this form of test before.

In the development of self-report personality inventories, several approaches have been followed in formulating, assembling, selecting, and grouping items. Among the major procedures in current use are those based on a) content validation, b) empirical criterion keying, and c) factor analysis.

- a) Content-related validation. These personality inventories are generally formed from lists of known problems which the individual can then tick as affecting them or not affecting them. This is the technique used in the development of the *Lifestress Inventory* described earlier in this chapter.
- b) Empirical criterion keying. This method builds upon the previous method but takes a more statistical approach, looking for items that separate "normal" from "abnormal" response patterns. In a purely hypothetical example, if it became known that schizophrenics showed a fear of clocks, an item assessing attitude to clocks could be included in a test designed to detect schizophrenia. It is not important that we have no idea why clocks might inspire fear in this group. The important thing is that people with the disorder have the fear whilst others don't, so empirical criterion keying would suggest that such an item could be included. The best known example of a personality test developed through the use of empirical criterion keying is the *Minnesota Multiphasic Personality Inventory (MMPI)*. The *MMPI* is a very large self-administered test, comprising numerous sub-scales. The sub-scales were developed empirically by criterion keying of items, the criterion being traditional psychiatric diagnosis. The latest revision of the *MMPI* has resulted in it being separated into two forms, the *MMPI-2* and the *MMPI-A* (for use with adolescents). The *California Psychological Inventory (CPI)* is another very well-known instrument that was based on the *MMPI*. It consists of 434 items to be answered true or false. Half of these items came from the *MMPI*. The *CPI* has been widely used in industry as well as in clinical practice.
- c) Factor analysis. As mentioned earlier, factor analysis is a technique for detecting patterns of correlations among test scores that indicate underlying dimensions that are responsible for scores on the test. Factor analysis can be used to help select items for inclusion in a personality

test or to identify how many dimensions underlie tests developed by either of the first two methods. Cattell's *Sixteen Personality Factor Questionnaire (16 PF)* was developed using this method. The so-called Big Five Factor Model (Costa & McCrae, 1991), perhaps the dominant model of personality in occupational testing settings, was also based on factor analysis. The big five personality factors are:

1. Neuroticism (N): indicates an individual's level of emotional stability, ranging from calm and even-tempered up to maladjustment and emotional distress.
2. Extraversion (E): indicates a person's degree of sociability and preference for interacting with people.
3. Openness (O): measures openness to experience, and is related to divergent thinking and creativity. Low scorers tend to be conventional and conservative.
4. Agreeableness (A): measures how a person views others. Low scorers tend to be competitive while high scorers favour cooperative interactions with others.
5. Conscientiousness (C): indicates a person's ability to control impulses and desires. High C is associated with strong will and high need for achievement, while low C is associated with a more lackadaisical approach to life.

Applications of personality testing

The two traditional areas for the application of personality tests have been clinical settings and occupational settings. Recently, personality tests have become popular in the new field of sport psychology, where they are used to gain insights into factors that affect performance.

Personality testing in clinical settings

Personality testing has a long history in clinical settings, where it has obvious relevance to the analysis of personality disorders. Perhaps the most common use of personality tests stems from the profile that can be obtained following their administration. A profile is a line linking an individual's scores on various parts of a test. Figure 15.1 shows a profile on an ability test. Similar profiles can be constructed for personality tests. The resulting pattern can be inspected for signs of abnormality. A single high or low score on its own may not indicate any problems but a combination of test scores may well be indicative of particular syndromes, such as schizophrenia.

These forms of profile analysis have not lived up to expectations for two main reasons. Firstly, variations among subtest scores could arise from a variety of circumstances, only some of them pathological. Secondly, the diagnostic categories that provided the criteria for profile analysis are

themselves subject to debate. For example, what does it mean to say that a particular pattern indicates schizophrenia when schizophrenia itself is not a clearly defined condition? Profile analysis may tell us something about group characteristics but it is prone to error when applied to individual cases. A more fruitful approach is to treat the pattern information as a source of hypotheses that can be tested against the wealth of other data collected in individualized testing.

Personality testing in occupational settings

Personality testing also has a very strong tradition in job selection testing where the 16 PF and more recently the Five Factor Inventory, a measure of the big five personality factors (NEO-FFI: Costa & McCrae, 1991), have proved very popular. This popularity continues despite evidence that personality tests do not predict job performance very well, even for sales positions (Hunter, 1986). Robertson and Smith (1989) report a validity coefficient as low as .15 in personnel selection testing. In contrast, the coefficient for ability tests ranges between .25 and .45.

The use of personality testing in occupational settings is not confined to selection testing, it has also proved very popular as an aid in training courses. The Myers Briggs Type Indicator (MBTI) is one of the best known personality tests because it is used so often in workshops on career decision making, team building, conflict resolution, time management, relationship counselling, and a number of other applications. The MBTI is based on the theory of psychological types proposed by Carl Jung. Psychological types represent combinations of two or more traits or attributes that are stable and shape the way individuals think and behave. The MBTI classifies people into 16 types and one of the reasons for its success lies in the fact that all types are seen as being valuable with each having particular strengths and weaknesses. There are many clones of the MBTI that also seek to describe people in terms of types. They are frequently used in management courses. Such type indicators can be extremely valuable in workshop settings where they serve as a basis for discussion of the different perspectives individuals may have on work situations, home life, and so on. For career selection purposes, there are as yet no data to support claims that knowledge of type (or personality) is a useful in predicting job performance.

Personality testing in sport settings

The latest field of psychology to embrace personality testing in a big way is sport psychology. Much of the testing centres around what is now known as "sport personology" - the study of personality characteristics as determinants of sporting success. As in occupational testing, the findings so

far have not been very promising. When the personality profiles of elite athletes are compared with those of novice athletes, there are differences. Elite athletes are more aggressive, more focussed, less anxious, and so on, but individual differences on these traits do not predict who is going to be an elite athlete. That is to say, the tests of personality have concurrent validity but poor predictive validity. Talent identification programmes have grappled unsuccessfully with this problem for years. Perhaps more situation-specific personality tests will help to improve the predictive validity of personality tests in both occupational and sports settings. There is no doubt that serious attempts are now being made to develop personality tests that are suited to sports situations. Ostrow (1990, p. 8) has included a graph which shows a quite steady increase in the number of sports-specific tests since 1975. The proportion is now close to 45%. As one might expect, given the nature of sport, these new tests are primarily in the areas of anxiety, motivation, mental skills, and specific sporting factors such as team cohesion (Fogarty, 1995).

Tests of vocational and career interests

Another category of test that has proved to be very popular in educational and occupational settings is the career interest inventory. The best known of these are the tests based on Holland's model of career decision making. The Self-Directed Search (SDS) is the most popular of these tests and has an Australian version which is in widespread use in this country. Holland (1985) believed that the most productive approach to career decision making involved an investigation of the individual's personality type. He proposed a six-category typology: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C). He believed that this six-category system could be used to not only describe the major types of people but also to describe the work environments they are likely to encounter in Western society. Holland's assumption was that people seek environments that allow them to express their interests, and by knowing something about their general orientation we are in a better position to judge where they will be happiest working. Holland's tests, or derivatives of them, are widely used in educational and occupational settings to assist with career decision making.

Miscellaneous tests

There are many more test types than can be described here, some of them adapted to particular situations. Areas not covered in this chapter include stress (e.g., Osipow and Spokane, 1987), values (e.g., Schwartz & Bilsky, 1987), decision styles (e.g., Driver, Brousseau, & Hunsaker, 1990), learning styles (Entwistle, 1983), and perhaps it is better to stop here

because the list could go on and on. The reader is referred to a text devoted exclusively to psychological testing, such as Anastasi and Urbina (1997), for an overview of virtually the whole testing domain. For information on tests available here in Australia, the best place to approach is the Australian Council for Educational Research (ACER), which has a number of test catalogues containing descriptions of individual tests, including details of what qualifications you need to administer the tests and suitable areas of application.

Computerised testing

It would be a mistake to conclude this chapter leaving the reader with the impression that tests are available only in paper-and-pencil format. Testing was one of the earliest areas within psychology to benefit from computer applications with standardised, objective-type personality tests being particularly well-suited to automation (Bartram & Bayliss, 1984). Initially, interest focussed on automated scoring but later expanded to include the computerised administration of existing pencil-and-paper tests. Currently, almost every facet of personality testing has been computerised, from test design and development, through item generation and analysis, to test interpretation and report generation. In a typical computerised test presentation, individual questions or stimuli are presented on a video display unit (VDU) attached to the computer, a set of limited responses is offered, and test-takers record their selected response via a keyboard or some other interface. The advantages of this form of administration over conventional administration are well-documented in several reviews (e.g., Bartram & Bayliss, 1984).

The move from paper-and-pencil tests to computer-based formats, however, represents a major shift in the way tests are administered and it is important that research is conducted to check the equivalence of the two methods. Work has already started in this area, especially on the equivalence of paper-and-pencil versus computerised presentation. Reviews of these studies report conflicting findings, with many uncontrolled variables influencing the outcomes (e.g., Burke & Normand, 1987; Webster & Compeau, 1996). The author's own experience with this form of testing is that it does not appear to make a noticeable difference and that test manuals developed on the basis of paper-and-pencil tests are still applicable to computerised versions of tests (Fogarty, 1998).

One of the major benefits of computerised assessment is undoubtedly the increased efficiency of administration made possible by software that adapts the presentation of items for each user. Thus, when assessing abilities there is no need to present a whole lot of easy items to a very capable person. It is a waste of time. Similarly, there is no need to

present a lot of difficult items to a person who has no chance of solving them. In a traditional paper-and-pencil test, everyone is given the same instructions, the same items, and the same time in which to complete the test. In an adaptive, computerised test situation, the algorithms built into the software can quickly estimate a person's *level* of cognitive functioning on the ability being measured, rather than the person's *total score*. Such an estimation is possible because the difficulty level of each item is known beforehand and the test can draw upon a large bank of items covering all possible ability levels. A typical test scenario is presented in Figure 15.4.

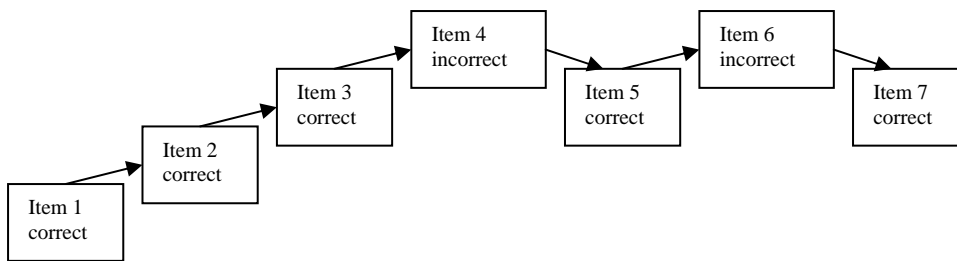


Figure 15.4. Illustration of possible sequence of items in an adaptive computerised test

If you can imagine that the higher items are more difficult you can see that each time the person is correct, a more difficult item is selected. When the item is incorrect, the computer selects an item from an easier level. In the oversimplified representation shown in Figure 15.4, the person's ability level is somewhere near to the levels assessed by items 3, 5, and 7. Above that level, the items are incorrect. Below that level and the items are correct. In actual practice, it is not quite this easy, but the principle is clear. At some point, performance will alternate between success and failure. The test is then ended and the level at which this occurred is reported as the ability estimate for that individual. Adaptive tests are not new in psychology and education, many individually administered tests are adaptive (e.g., the *Stanford Binet*), but computers allow whole groups of people to be tested simultaneously. In some cases, testing time is halved because fewer items have to be presented. Adaptive computer testing can be used with other types of test (e.g., personality), but so far their application has mostly been with intelligence and achievement tests.

Apart from test administration, computers are increasingly being used to write test reports on the basis of scores collected during a computerised administration of the test or entered by the test administrator.

There is a real question mark hanging over the issue of the lack of flexibility of reports written by a computer. Such reports are based on test scores only and omit the large amount of data that can be collected in a face-to-face testing situation. When using tests of maximal performance, for example, a clinician can report the amount of effort put into the test, a computer cannot. In personality testing, many signs of abnormality may be evident in the person's bearing and manner, things that are not available to the computer that generates the report. It may be some time before there is good research data on the validity of computer generated reports. Certainly they should be supplemented by other reports from the training officer, psychologist, or whoever arranged the test.

PART C: LIMITATIONS OF TESTING AND CONTROVERSIES

Up to this point, the chapter has emphasised the positive aspects of testing but there are negative aspects that should also be mentioned before closing. The chief limitation of testing is implied in the definition mentioned at the outset of the chapter: tests provide samples of behaviour. As such, they should never be interpreted as yielding completely accurate descriptions of people. They are accurate up to a point. The degree of accuracy is reflected in the psychometric properties discussed in the first section of this chapter. Even a test with excellent psychometric properties, however, will yield trustworthy data only if the individuals undertaking the test understand what is required of them and are motivated to answer the questions properly.

The fallibility of test results has resulted in some strong criticisms of the practice of psychological testing. There are two main areas of controversy, the first has to do with test users, the second with the tests themselves.

Problems relating to test users

Most examples of test misuse relate to test users. Problems relating to test users can be summed up under the following headings (Anastasi & Urbina, 1997):

- User qualifications and professional competence. The introduction to the technical aspects of testing at the start of this chapter has probably left some readers wondering how much training is required before one can be considered competent to administer tests. The answer is that the amount of training depends on the type of test with some tests such as the MMPI requiring a high level of training and others, such as tests of decision styles, requiring less training. Virtually all forms of testing, however, require a basic knowledge of psychometrics. Following reports by a special panel of the U.S. National Academy of Sciences set up to investigate testing practice (Wigdor & Garner, 1982a, 1982b), much

more attention is now paid to the qualifications of test users. Most test publishers now have categories of usage, with sensitive tests restricted to professions such as psychologists who complete the appropriate training, usually at postgraduate level. Less sensitive tests can be used by teachers, personnel managers, training officers, and other professionals. These people may have some university training but the majority will undertake private training courses run by their companies or by the test publishers. There are two main reasons for restricting test usage: a) to ensure that the test is administered and interpreted by a qualified user and that the test is properly used, and b) to prevent general familiarity with the test content, which would invalidate the test (Anastasi & Urbina, 1997, p. 10).

- Responsibilities of test publishers. The publishers have to make sure that they do not sell restricted tests to people who lack the training or qualifications to use them. In Australia, a lot of responsibility is placed on the publishers and distributors of tests to make sure that they do not fall into the wrong hands. The main distributors in this country are the Australian Council for Educational Research, The Psychological Corporation, Science Research Associates/London House, and private companies such as Saville Holdsworth. The first three of these publish catalogues that show quite clearly the level of training required for each test listed. The Psychological Corporation, for example, uses a three category system. Level A tests require basic professional qualifications, such as those gained through a Bachelor of Education. Level B tests require a specialist professional qualification, such as physiotherapy. Level C tests require advanced training in psychometrics, such as that provided through a Masters degree in psychology. Private companies, such as Saville Holdsworth, run their own training courses which are open to everyone, for a fee.
- Protection of privacy and confidentiality. Given the nature of constructs being assessed by educational and psychological tests (e.g., personality, ability, achievement), there has always been a concern about leakage of test results to people who may use the results inappropriately. For example, an employer who refused a promotion to an employee because he or she had learned from an outside source that the employee had a high score on neuroticism. Such incidents were not uncommon in the early days of testing. The issue of privacy and confidentiality is now covered by statements in the code of ethics for psychologists and statements about *Testing Standards*. For non-psychologists, the threat of legal action helps to keep matters in check.
- Communicating test results. As you have seen, testing can envelop itself in a shroud of technical terms that imposes a barrier between the general

public and test administrators. Some of the controversy surrounding test usage has stemmed from basic misunderstanding of what testing is all about. More attention is now being given to communicating test results in an intelligible form for parents, teachers, and others who may need to view reports.

Problems relating to tests

Problems relating to the tests themselves include the following:

- Self-report inventories, of which there are a great number, are too open to faking. People can often see what are socially desirable responses and may choose to respond in a "socially desirable" way, especially if a job is at stake. Lie scales - questions designed to trap people who are trying to project a favourable image - are often used to overcome this problem but the evidence indicates that lie scales are not all that effective.
- It is unfortunate that so many tests are constructed that do not meet the rigorous guidelines discussed above. The problem is worse in some areas than in others. New fields of education and psychology tend to suffer from a rash of poorly constructed tests. The field of sport psychology is an excellent example of over-zealous test development. Fogarty (1995) reviewed the situation in this field and reported a large number of tests developed for the purpose of a single study but then used in applied settings without any evidence of reliability or validity. Many of these tests are not worth administering.
- Fortunately, most fields of education and psychology are well developed and offer a variety of tests for which there is abundant psychometric information. Sources such as the *Buros Mental Measurements Yearbook* (MMY) offer critical reviews of nearly all commercially available psychological, educational, and vocational tests published in English. Test manuals that accompany commercial tests also contain a lot of valuable psychometric information that can be used to help evaluate a test. In fact, one could almost say that if a test does not have a manual, it may not be a good idea to use that test. Look for one that has the important psychometric information described above and one that, hopefully, is reviewed positively in publications such as the MMY.
- Various forms of bias can occur in tests, especially tests of ability and achievement, such that the tests tend to favour one group over another. The debate on bias has been most bitter in the U.S. where it has been known for many years that significant black-white differences exist on tests of intelligence. If such tests are used to select employees, obviously white applicants are going to have a better chance of success. However, the debates on bias in the 1970's made it clear that bias does not exist simply because one group scores better than another on the selection

test. Job performance differences must also be taken into account. Bias exists if it can be shown that the equation used to predict job performance on the basis of test results is different in certain important ways for the two groups. The issues surrounding the debate are too technical to introduce here, suffice it to note that a great deal of research has failed to show evidence of bias. It seems that where tests predict such outcomes as job performance or academic achievement, they do so equally well for most cultural minorities. The emphasis has now switched from looking at the relationship between test scores and job performance, which seems to be the same for all groups, to looking at other criteria that might lead to more members of the lower-scoring group being selected.

Conclusion

There is no doubt that despite the criticism testing may have attracted from some quarters, it is here to stay. If anything, its popularity appears to be growing as the search for information continues to drive our society. It is to be expected that people will seek to know more about themselves and about each other. If testing is to become as safe and dependable as we would like it to be in the fields of clinical psychology, organizational psychology, and education then the tests themselves must be valid and reliable and test users must be well-trained in the principles and ethics of testing. The first of these considerations can be handled by ensuring that tests are evaluated and that psychometric data are published in independent sources such as the *Buros Mental Measurements Yearbooks*, as well as the test manuals. The World Wide Web will play an important role in disseminating this information. Much information on tests can already be obtained from this source. Professional bodies, test publishers, and legislators need to ensure that people are competent to use whatever tests they employ in their work. Potential users will require sources of objective guidance on what to buy, training will be available from a number of sources (not just the test distributors), and the interests of test-takers will be protected.

Tests have been criticised from various quarters over the past few decades, and such criticism has led to improved testing practices. The criticism has not, however, uncovered any basic weaknesses in theory or methodology. Indeed, one of the main reasons tests are criticised is simply because they can be criticised - they are open to review. Psychological tests do not provide a basis for making completely accurate decisions about individuals. In reality, there is no method which guarantees complete accuracy. However, a special panel of the National Academy of Sciences in the United States concluded that psychological tests generally represent the

best, fairest, and most economical method of obtaining information which is necessary to make sensible decisions about individuals (Wigdor & Garner, 1982a, 1982b).

Review questions and activities

1. Do you think measures of intelligence taken in early childhood would predict academic performance in adult years? Explain.
2. What problems do you foresee with computers not only administering tests but also scoring and interpreting them?
3. On the basis of your own experience, draw up a list of problems related to testing and make suggestions as to how these problems may be rectified. Base your answer on your own personal experiences or, if necessary, your imagination.
4. If you were about to apply for a job, would you feel comfortable if you learned that the selection process included tests of personality and intelligence?
5. Assume that you are working in the personnel section of a small firm and have been asked to prepare selection criteria for a computer programmer's position, what tests would you consider using?

There is a site on the Web that allows you to complete a personality test and then obtain feedback. Another site allows you to test your IQ. Whether they will still be there when you read this, I do not know. Perhaps there will be other sites. Here are the addresses:

Personality test: <http://www.onlinepsych.com/home.html/>

Five-minute IQ test: <http://www.brain.com/>

True-false questions for review. Read each of the statements below and decide if it is true or false.

1. Reliability coefficients reflect the extent to which a test measures what it purports to measure.
2. Face validity is the most important aspect of validity.
3. A test is acceptable for use if it is either reliable or valid. It does not need to be both.
4. A correlation coefficient of -1.00 between two tests means that the tests are unrelated.
5. The standard error of measurement is useful for interpreting individual test scores.

6. If I said that someone had a stanine score of 3 on a test, I would be using relative norms to assess performance rather than some external criterion.
7. Convergent discrimination refers to the extent to which a test correlates with other variables with which it could be expected to correlate.
8. A percentile score of 23 means that a person has scored exactly 23 on a test that is marked out of 100.
9. IQ scores are based on a distribution that has a mean of 100 and a standard deviation of 15.
10. Item-person maps are useful for judging whether the items in a test are suitable in difficulty level for a particular test population.
11. The *Stanford Binet* is a group test of intelligence.
12. Individual tests are easier to administer and score.
13. Neuropsychological tests are designed to assess brain damage.
14. Holland's model of career interests is based on four basic personality types: Realistic, Social, Investigative, and Enterprising.
15. Cattell's 16 Personality Factors model of personality is based partly on factor analysis.
16. The so-called big five personality factors are neuroticism, extraversion, openness, agreeableness, and conscientiousness.
17. Profile analysis is a technique for looking at patterns of correlations among test scores.
18. The biggest problem with the use of tests of intelligence in occupational selection is that they are biased against minority groups.
19. Research findings indicate that people respond quite differently to tests that are presented by computer.
20. The tendency to respond in a socially desirable way can be a problem with self-report inventories.

Comments on review questions and activities

1. One would have to question the reliability of measures of intelligence in early childhood. Even if you considered the measure to be both reliable and valid, it is not likely that measures taken so far apart, especially when the first measure was taken during early childhood, will be strongly related. Although many parents take delight in seeing signs of great intelligence in their very young children, the effects of schooling, motivation, and opportunity will combine with innate potential to determine eventual learning achievements. The length of time itself is not the major consideration: measures of intelligence taken in early adulthood will still be strongly related to measures of learning taken in mid-adulthood. Rather, the enormous changes in intellectual

development between childhood and adulthood will weaken the relationship.

2. Points to consider include the following: a) Can any computer programme capture the enormous complexity of individual test results? Will the reports be too stylised? b) Will the highly impersonal nature of human-computer interaction dissuade people from responding accurately? [Actually, the evidence suggests that some people are more inclined to reveal personal details to a computer than to another person]. c) What happens if something goes amiss during the testing process when a computer is running the test? d) Will the people reading the report have any idea of how it was generated? Or how it can be explained? e) Will the test results and reports be confidential if they are stored on computer disk?
3. Possible problems include: a) Were you aware of the purpose of testing? b) Were you asked if you were willing to participate? c) Did you find out your results? d) Could you understand the results? e) Was the testing harmful to you in any way (e.g., prospects)? f) Were the results treated in a confidential manner?
4. You should feel more comfortable if you can be assured that the test is reliable. If it is reliable, you can have some confidence that the score you obtain is about what you would normally obtain on such a test. You should also feel more comfortable if it is valid. Not only should it measure what it was designed to measure but these processes must be related to job performance. There have been some successful lawsuits in Australia involving disgruntled job applicants suing test administrators because the test used in the selection process was not related to job performance.
5. What would you need to measure? Possibly programming skills. Are reliable and valid tests already available? Yes, there are some. How do you find out about them? Do you need to construct your own test? What else could you measure that might be related to performance in a programming position? Possibly interests. Programmers work a lot with machines, not a position well suited to people who want to spend most of their time working with other people. What about personality variables? Could you justify the inclusion of any tests you choose in a court of law if an applicant claimed that they were unrelated to the position advertised? Do you need to use tests at all?

Answers to Review Questions

1. False. This is the definition of validity. Reliability measures the extent to which test scores are free from error.

2. False. Face validity is important so that people will accept the test but it has no scientific status and is actually the least important aspect of validity.
3. False. Tests must have both qualities. Without reliability, there will not be any consistency in results. Without validity, the test is not measuring what you think it is measuring. Also, validity partly depends on reliability: if test scores contain a lot of error, they cannot be measuring what the test intended.
4. False. A correlation of 0.0 means that they are unrelated. A correlation of -1.0 indicates that the two tests are perfectly related but in an inverse manner: a high score on one test indicates a low score on the other.
5. True, the standard error of measurement takes the reliability of the test into consideration and allows test users to place an interval around an individual's obtained score that has a high probability of including the true score.
6. True. Stanine scores are one means of estimating where an individual is located in comparison to his or her peers.
7. True. A test of vocabulary, for example, would be expected to correlate with a test of comprehension because they are both tests of verbal ability. A test of vocabulary, on the other hand, would not have a high correlation with a test of decision making, because they are quite different constructs.
8. False. A percentile score of 23 means that the score is better than or equal to 23% of the scores obtained on the test. You cannot tell from the percentile score what the actual test score was.
9. True. An IQ score of 115 is therefore one standard deviation above the mean.
10. True. These item person maps are very useful for judging the suitability of a test for a particular population. The range of difficulty for the items should match the range of ability for the individuals.
11. False. It is an individual test that takes between 1 to 2 hours to administer.
12. False. They are usually time-consuming (e.g., *Stanford Binet* up to 2 hours) and only one person can be tested at a time.
13. True, they designed purely to detect particular cognitive weaknesses that might indicate damage to the part of the brain responsible for that function.
14. False. It is a hexagonal model that also contains Artistic and Conventional types.
15. True. Factor analysis has often been used to establish the construct validity of tests of intelligence and personality. Cattell was one of the first to use the technique with personality tests.

16. True. The Big Five model is very popular in occupational testing
17. False. Factor analysis is used to check for patterns of correlations among test scores. Profile analysis looks at the pattern of scores obtained on a test with a number of sub-scales to see whether there are any abnormalities.
18. False. Whilst it is true that for a variety of reasons minority groups tend to have lower scores on tests of intelligence this does not by itself constitute evidence of bias.
19. False. The evidence on this question is uncertain at this stage but it is likely that there are no real differences between the two methods of presentation.
20. True. Sophisticated test takers can often tell what are the desirable answers in a test and may respond according to how they want to appear rather than how they actually think or feel.

References

- ACER (1982). ACER Advanced Tests: AL-AQ and BL-BQ. *Catalogue of Tests and materials*. Melbourne: ACER.
- Ackerman, P.L., & Heggestad, D.D. (1997). *Intelligence, Personality, and Interests: Evidence for overlapping traits*. *Psychological Bulletin*, 121 (2), 219-245.
- Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed). Saddle River, NJ: Prentice Hall.
- Bartram, D., & Bayliss, R. (1984). Automated testing: Past, present and future. *Journal of Occupational Psychology*, 57, 221-237.
- Bennett, G.K., Seashore, H.G., & Wesman, A.G. (1989). *Differential Aptitude Tests: Australian Manual*. NSW, Australia: The Psychological Corporation.
- Bramston, P., & Bostock, J. (1994). Measuring perceived stress in people with intellectual disabilities: The development of a new test. *Australia & New Zealand Journal of Developmental Disabilities*, 19, 149-157.
- Bramston, P., & Fogarty, G. (1995). Measuring stress in the mildly intellectually handicapped: The factorial structure of the Subjective Stress Test. *Research in Developmental Disabilities*, 16 (2), 117-131.
- Burke, M.J., & Normand, J. (1987). Computerised psychological testing: Overview and critique. *Professional Psychology: Research & Practice*, 18, 42-51.

- Costa, P. T., & McCrae, R. R. (1991). *Manual of Revised NEO Personality Inventory and NEO Five-Factor Inventory*. Odessa, FL: Psychological Assessment Resources.
- Cronbach, L. J., (1970). Essentials of Psychological Testing (3rd ed). New York. Harper & Row.
- Entwistle, N.J. (1983). *Styles of Learning and Teaching: An integrated outline of educational psychology for students, teachers, and lecturers*. Chichester: John Wiley.
- Driver, M.J., Brousseau, K., & Hunsaker, P. (1990). *The dynamic decisionmaker: Five decision styles for executive and business success*. New York: Harper & Row.
- Entwistle, N.J. (1983). *Styles of Learning and Teaching: An integrated outline of educational psychology for students, teachers, and lecturers*. Chichester: John Wiley.
- Fogarty, G. (1995) Some comments on the use of psychological tests in sport settings. *International Journal of Sport Psychology*, 26 (1), 161-170.
- Fogarty, G. (1998). Response bias in computerised tests. *South Pacific Journal of Psychology* (in press).
- Fogarty, G., & Bramston, P. (1997). Validation of the Lifestress Inventory for people with mild intellectual handicap. *Research in Developmental Disabilities*, 18 (6), 435-456.
- Fogarty, G., & Taylor, J. (1997). Learning styles among mature-age students: Some comments on the Approaches to Studying Inventory (ASI-S). *Higher Education Research and Development*, 16 (3), 321-330.
- Holland, J.L. (1985). *Making vocational choices: A theory of work vocational preferences and work environments* (2nd Ed.). Odessa, FL: Psychological Assessment Resources.
- Hunt, E.B. (1995). *Will we be smart enough: A cognitive analysis of the coming workforce*. N.Y.: Russell Sage Foundation.
- Hunter, J.E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behaviour*, 29, 340-362.
- Murphy, K.R., & Davishofer, C.O. (1988). *Psychological Testing: Principals and Applications*. Englewood Cliffs, NJ: Prentice-Hall.
- Osipow, S.H., & Spokane, A.R. (1987). *Manual for Occupational Stress Inventory: Research version*. Odessa, FL: Psychological Assessment Resources.
- Ostrow, A. C. (1990) Directory of psychological tests in the sport and exercise sciences. Morgantown: Fitness Information Technology, Inc.

- Reich, R. (1991). *The work of nations: Preparing ourselves for 21st century capitalism*. New York: Knopf.
- Robertson, I.T., & Smith, M. (1989). Personnel selection methods. In M. Smith and I. Robertson (Eds.), *Advances in selection and assessment* (pp.89-112). New York: Wiley.
- Schwartz, S.H., & Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*, 53, 550-562.
- Tenenbaum, G., & Fogarty, G. (1998). Application of the Rasch analysis to Sport & Exercise Psychology measurement. In J. Duda (Ed.) *Advancements in Sport & Exercise Psychology Measurements*, pp. 409-421. Morgantown, USA: Fitness Information Technology.
- Webster, J., & Compeau, D. (1996). Computer-assisted versus paper-and-pencil administration of questionnaires. *Behavior Research Methods, Instruments, & Computers*, 28 (4), 567-576.
- Wigdor, A.K., & Garner, W.R. (1982a). *Ability testing: Uses, Consequences and Controversies, Part 1: Report of the Committee*. Washington, DC: National Academy Press.
- Wigdor, A.K., & Garner, W.R. (1982b). *Ability testing: Uses, Consequences and Controversies, Part 2: Report of the Committee*. Washington, DC: National Academy Press.

Profile

Dr Gerard Fogarty completed a BA (Hons, Psychology) degree at the University of New England in 1973. He then completed a Diploma of Education and taught English and History for three years at Cabramatta High School in Sydney's Western Suburbs in preparation for further training as a school counsellor. After completing these teaching years, he enrolled in a PhD at the University of Sydney, working with Dr Lazar Stankov on a thesis that explored aspects of the structure of human intelligence. Dr Fogarty left Sydney University in 1984 to take up a position with the Head Office of the AMP society where he supervised the development and validation of a new computerised selection system for the 5,000 strong fieldforce of the AMP, a system that includes tests of intelligence, personality, and interests. Dr Fogarty joined the University of Southern Queensland in 1988, where he is still working as Head of the Department of Psychology, lecturing on statistics and psychological measurement. He has published many articles in the areas of intelligence and the validation of psychological tests.