

Psychometric testing



Introduction

The technology of psychometric tests

Testing intelligence

Measuring personality

Summary

Introduction

Psychometric testing produces some of the greatest differences of opinion and some of the strongest arguments to be found in psychology. The arguments often get confused because of the mixture of scientific and political issues that come into play. There are arguments about:

- practical issues: do the tests give accurate and consistent results? Do the complex statistical procedures illuminate or disguise what is going on?
- theoretical issues: do the tests measure underlying psychological qualities? For example, is there such a quality as 'intelligence' and is it measured by an IQ test?
- political issues: can the tests be used to look at differences between groups of people, for example different social classes, or different ethnic groups, or

the differences between men and women? What are the social and political consequences of using tests to categorise people?

In this chapter we will look at all these main issues, though it must be said that some of the arguments are quite complex and cannot be argued in full in the space that is available. The chapter starts with a brief review of the technology of psychometric testing which has been developing for over 100 years. We will then look at intelligence testing and its rather dubious past, and finally we will take a brief look at some issues in personality testing.

The technology of psychometric tests

The term ‘psychometric’ means ‘measuring the mind’, though many psychometricians would be very uncomfortable with a term such as ‘mind’. A psychological test is a task or set of tasks that can be given in a standard format to an individual, and which produce a score that can be represented as a number (or a category). It can involve almost any activity. Most commonly, for reasons of practicality, it involves filling in a questionnaire. Tests are used to measure cognitive functions (for example, IQ tests), personality (for example, Eysenck’s EPI, see below), mood (for example, the Beck Depression Inventory), attitudes (for example, political opinion polls), aptitude for various jobs (for example, the Comprehensive Ability Battery), illness behaviour (for example, the McGill Pain Inventory), and many other qualities. Psychometric tests are extensively used in everyday life, and you are likely to come into contact with them on a fairly regular basis.

Performance and ability

One of the first issues to consider is the distinction between **performance** and **ability**. Performance is what you actually do, and ability is what you are capable of. It is a common experience of students that their teachers say ‘you have the ability, but you are not doing the work’. They mean that the reason you got a Grade E in your examinations was due to poor performance and not poor ability (and obviously not the teacher’s fault either). Any test we give to someone can only measure their performance on that test, and not their ability. We can only infer their ability from that performance. So when we are measuring

intelligence, we are, in fact, measuring *performance* on the particular test and not the underlying *intellectual ability*.

The problem that arises when we interpret the results of a test is the many factors that affect performance. These factors include:

- language of the test: words mean different things to different people, and performance can be affected by our level of understanding of the questions.
- test situation: some people work well in quiet environments with few distractions, whereas others can only concentrate if the television is on in the background.
- expectations: if you expect that you will not be able to answer the questions you will be more likely to give up easily.
- motivation: if you are competitive and want to win everything then you will try harder to do well (if it is an IQ test, for example)

Measuring people

To measure something you have to compare it against something else. If we are measuring a table, it is easy because we can use a ruler, but if we are measuring people what can we use? There are three ways in which we can use tests to measure people:

- 1 direct measurement: where we use a physical measure such as grip strength or reaction time. Although these measures can be useful, there are only a limited number of direct measures we can make of people.
- 2 criterion-referenced measurement: where we compare the performance of an individual against an ideal performance.
- 3 norm-referenced measurement: where we compare the performance of an individual against performance of other people, most commonly the peer group. This is far and away the most common way of using psychological tests.

An example of norm-referenced measurement can be seen in IQ testing. As children get older, they develop their educational skills and so get more answers correct on an IQ test. This is not taken to mean that their IQ is improving, even though their intellectual skills are improving. The individual is compared

against children of the same age, so their IQ score represents their ranking in their own age group. This measure has a number of benefits in that it tells you how well you are doing compared to people like you. It also, however, gives a lot of emphasis to the differences between individuals, even when those differences might be very small.

Test reliability

If we are using a psychometric test, we need to know whether it will give us a consistent result. So, if we give someone the test on a Wednesday afternoon, we hope to get the same results as if we had given the test on Friday morning. It would be remarkable if we got exactly the same result because all forms of measurement have an element of error in them, but we hope this error is relatively small. Psychologists use a number of techniques for assessing **reliability** including:

- 1 Test-retest: in this case we administer the test on two occasions and compare the scores using a correlation technique. Test-retest reliability is affected by a number of factors including:
 - (i) changes in subjects: for example if we leave a gap of three months between the test and re-test of an IQ test we can reasonably expect children to develop their intellectual skills in that time, and since children develop at different rates, their rank order will also change during that time. This will reduce the test-retest reliability calculation.
 - (ii) measurement error: for example, changes in mood brought about by a hangover for one administration of the test, poor test instructions so the subjects do not grasp what is required, and guessing, which people always do but produces a degree of unreliability in the scores.
 - (iii) other factors: for example, if the time gap between the two test times is very brief then the subjects are likely to remember their answers, or if the questions are very easy, then you will get a high reliability value because the subjects get them nearly all right on both occasions.

- 2 Internal consistency reliability: in this case, we compare two parts of the test to see how similar the scores are.

Before the arrival of desktop computers, the most common way to calculate internal consistency was through split-half reliability. This method correlates the response to half the items (for example, the odd-numbered items) to the other items (the even-numbered items). This gives a rough approximation of the reliability value, although a number of corrections are made to the scores to take account of such things as test length (the longer the test the greater the reliability, so if you split the test into two you inevitably reduce the reliability score).

A more sophisticated version of the split-halves technique looks at all the possible split halves and gives a value of reliability called Coefficient Alpha. This makes an estimate of the amount of error in the test and hence what the 'true' score would be without any error. It then compares your score with this 'true' score to give a measure of reliability.

All of the above reliability techniques have serious errors because they make assumptions about the data that are rarely appropriate (Shevlin, 1995, 1998). Psychometricians, therefore, use more complex measures of reliability that use a similar principle to that used by Coefficient Alpha but do not make so many assumptions of the data.

The above is an incredibly brief summary of the statistical procedures surrounding reliability, but believe me, you do not want to see the full details or the mathematical formulae. I can already sense many reader's eyelids becoming more heavy as they work their way through this section. (Note: if you are interested in the technology of testing, then you should refer to Kline, 1993, or Kaplan and Saccuzzo, 1993, and then seek help from your general practitioner or personal counsellor).

We can take two general points from the above. First, there are a number of statistical controversies in psychometrics about how to evaluate the data. Second, we can see that measurement always contains error, but psychologists have been developing statistical tools for over one hundred years to reduce this error, and the whole testing business is statistically very sophisticated. As a result, there are a number of reliable psychometric tests that are used to measure a wide range of psychological variables.

Test validity

A test is said to be valid if it measures what it claims to measure. So we consider whether a depression scale measures depression, and whether an IQ test measures intelligence. This is not as obvious as it sounds, and it is a complex process to measure this validity.

There are a number of types of validity, but perhaps the most important is construct validity. If a test has construct validity it matches up with a psychological concept or theory. If a test has construct validity then the following should occur:

- the scores of the test should correlate with other tests of the same psychological quality.
- the scores of the test should not correlate with tests of different psychological qualities.
- the scores of the test should predict future performance.
- the test items should have some connection to the appropriate psychological theory.

All this is easier said than done, and psychometricians go to great lengths to establish the validity of their tests. Many of these tests, however, continue to attract controversy about what they actually measure.

Standardisation

The idea of standardisation rests on the principle that abilities, both mental and physical, are distributed throughout a population according to a normal distribution curve. This curve describes a set of scores where a few people obtain extreme high scores, a few people obtain extreme low scores, and most people score around the average. If we assume the scores are normally distributed then we can make judgments about the performance of an individual using the statistics of standard deviation and z-scores.

Standardising a test involves establishing how the scores of this test are distributed among the population, and making sure that the test, if administered

to enough people, will produce a normal distribution. This involves testing large numbers of people and establishing what the normal scores for those types of people might be. From this, it is possible to develop population norms, which identify what would be an average score, what would be above-average, and what would be below-average; so standardisation, at least in theory, allows us to judge how typical, or uncommon, someone's result is.

The major issues here concern first, the selection of samples for the standardisation procedure, and second, the assumption that the scores are normally distributed. If we wanted to know the average IQ score of the people in this country, we could give everyone a test and work out the mean score. This is clearly impracticable, so we select a sample from the whole population and use the results of that sample to estimate the average score of the populations. Clearly there will be some error in this estimate, but the amount of error will depend on how representative our sample is.

Summary

The above brief discussion of the technology of psychological tests gives a flavour of the sophistication of test construction and the practical problems involved in constructing and interpreting tests. The controversies are mainly of a technical nature and concern the potential error in tests and how predictive they can be of future behaviour. The arguments might appear a little remote and technical, but without some understanding of the technicalities it is not possible to argue about the other issues that arise out of psychological testing. For example, it was the failure of psychologists to consider scientifically the data offered by British psychologist Cyril Burt that allowed him to get away with scientific fraud on the topic of intelligence testing (see Hearnshaw, 1979).

Testing intelligence

The testing of intelligence is perhaps the single most controversial issue in psychology. In the space available, I can only touch on a few of the issues and try and give you some idea of the strength of feeling that this topic produces. The following two quotes give some idea of the spread of opinion on this topic.

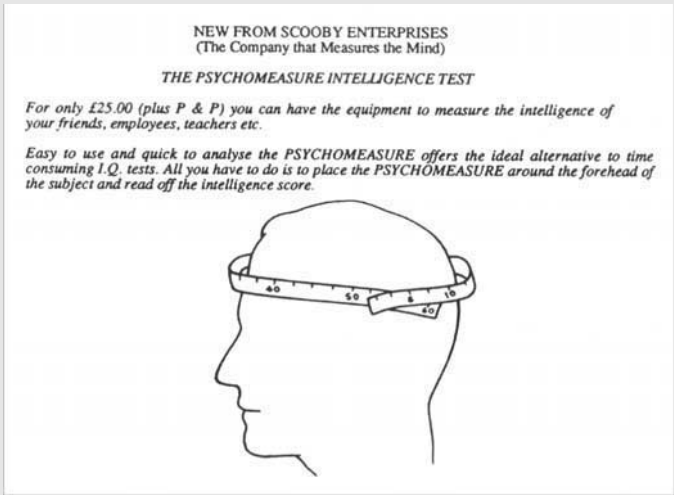


Figure 5.1 The Psycho-Scooby measure of intelligence

- 1 Which of the following statements best describes the PSYCHOMEASURE INTELLIGENCE TEST:
 - a The test is reliable and valid
 - b The test is reliable but not valid
 - c The test is valid but not reliable
 - d The test is neither reliable nor valid
- 2 Describe one way that you could assess the reliability of the PSYCHOMEASURE test.
- 3 Describe one way that you could assess the validity of the PSYCHOMEASURE test.
- 4 Most IQ tests have a bias in them that gives certain people better scores than others. Describe ONE bias in the PSYCHOMEASURE INTELLIGENCE TEST.

The measurement of intelligence is psychology's most telling accomplishment to date.

(Herrnstein, 1973)

The IQ test has served as an instrument of oppression against the poor.

(Kamin, 1977)

The early tests

The earliest psychometric testing is usually attributed to Francis Galton who set up a stall at the International Health Exhibition in London in 1884 and tested visitors mental abilities for the sum of threepence (see Figure 5.2). As many as 9,000 people took the tests, which included measures of reaction times to sounds, lights and touch, and other easily measurable motor activities and sensory judgements. I mention Galton in particular because of his belief that intellectual abilities were inherited and his enthusiastic support of eugenic solutions to the problems of society. Eugenics refers to the attempt to improve the quality of human beings through selective breeding; so for example, if we wanted to improve the general level of intelligence in the country, we would encourage intelligent people to have lots of children and unintelligent people to have none. Ideas like this can have damaging social consequences, so it is important to challenge them.

There are many problems with the eugenics approach to intelligence including the assumptions that:

- there is a human quality that we can call intelligence; this may appear self evident, but it is possible to argue that we have many different types of intelligent behaviour rather than just one.
- intelligence can be reliably and validly measured; though as we have already seen (see above), the issues of reliability and validity remain controversial to this day.
- intelligence is a fixed quantity and cannot be improved; though it might be that intelligence can be improved through coaching.
- the differences in intelligence between people are mainly due to genetic factors; most controversial of all the assumptions, it also has the least scientific evidence to support it.

ANTHROPOMETRIC LABORATORY

**For the measurement in various
ways of Human Form and Faculty.**

Entered from the Science Collection of the S. Kensington Museum.

This laboratory is established by Mr. Francis Galton for the following purposes:—

1. For the use of those who desire to be accurately measured in many ways, either to obtain timely warning of remediable faults in development, or to learn their powers.

2. For keeping a methodical register of the principal measurements of each person, of which he may at any future time obtain a copy under reasonable restrictions. His initials and date of birth will be entered in the register, but not his name. The names are indexed in a separate book.

3. For supplying information on the methods, practice, and uses of human measurement.

4. For anthropometric experiment and research, and for obtaining data for statistical discussion.

Charges for making the principal measurements:
THREEPENCE each to those who are already on the Register.
FOURPENCE each to those who are not:— one page of the Register will thenceforward be assigned to them, and a few extra measurements will be made, chiefly for future identification.

The Superintendent is charged with the control of the laboratory and with determining in each case, which, if any, of the extra measurements may be made, and under what conditions.

H & W. Brown, Printers, 20 Fulham Road, S.W.

Figure 5.2 Advertisement for Galton's early tests

All these assumptions are controversial and we will consider them further below. The important issue to bear in mind at this point is that intelligence testing has been consistently linked with genetic explanations of individual differences, and with scientists who proposed genetic solutions to social problems.

Binet's pioneering tests

The first tests that we can recognise as IQ tests were developed in France by Alfred Binet, who started his scientific studies by examining the relationship between head size and intelligence. He discovered that there was little connection between size of head and intelligence. He was later commissioned by the Minister of Public Education to develop a technique to identify children in need of special education, and from this the intelligence test was born. The test was used to give an estimate of a child's mental age by comparing the child's performance on various tasks with the performance of children of various ages. It was later suggested that the mental age of the child should be divided by the chronological age to give an index of intelligence, and so the notion of IQ was developed (the formula is given below). This is an example of norm referencing (see above).

Intelligence Quotient (IQ) = Mental Age/Chronological Age $\times$ 100

Binet believed that children who were in need of extra help could be identified by these tests, but he vigorously argued against the idea that intelligence is a fixed quantity that cannot be improved by further help. This approach got sadly lost in the translation of tests into the English and in their transportation to America. In contrast to the approach of Binet, the fiercest supporters of intelligence testing in the English-speaking world were scientists who believed that individual differences are mainly due to genetic factors, and who proposed eugenic solutions to the perceived problems of society. For example, Lewis Terman, who introduced the IQ test to America while he was professor of psychology at Stanford University, wrote:

If we would preserve our state for a class of people worthy to possess it, we must prevent, as far as possible, the propagation of mental degenerates.

(Lewis Terman, 1921, cited in Kamin, 1977)

The big words disguise the sentiments of the quote. To paraphrase Terman, he is saying we must stop poor and uneducated people from having children. All this would seem unpleasant but unimportant, were it not for that fact that over half of the states in the USA brought in sterilisation laws for the ‘feeble minded’ and carried out tens of thousands of operations (Kamin, 1977).

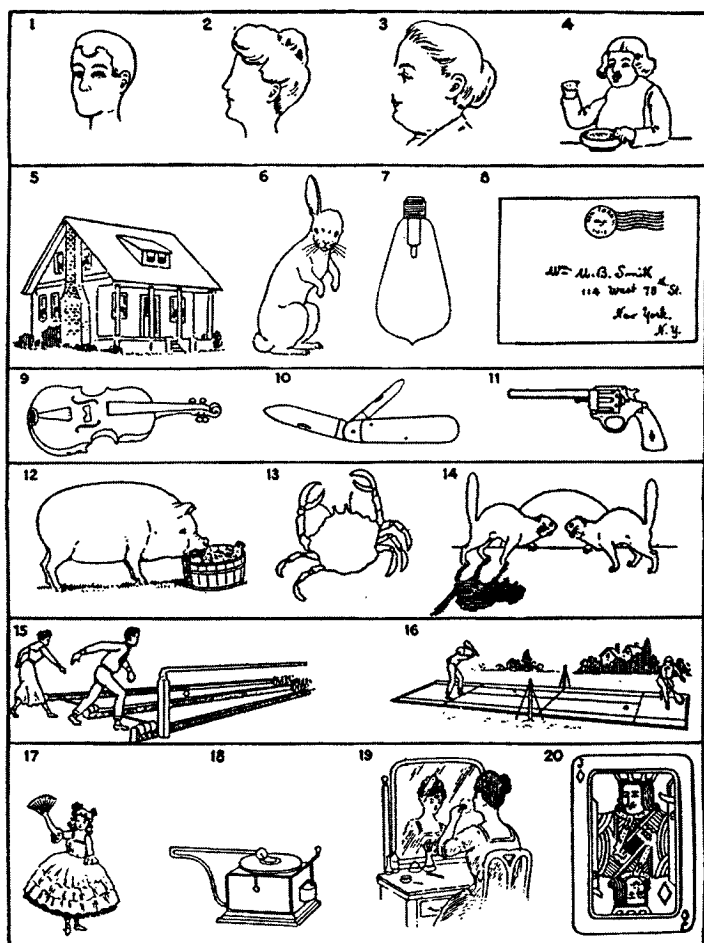
The first mass IQ testing

American psychologist Robert Yerkes was concerned to establish psychology as a ‘hard’ science and thought that mental testing looked a promising route to achieve this. Unfortunately, in 1915 mental testing did not enjoy much credibility, so Yerkes tried to change this. The outbreak of the First World War (1914–18) in Europe and the subsequent involvement of the USA brought about a massive mobilisation of armies. Yerkes managed to persuade the American military to give mental tests to all army recruits, and as a result he was able to preside over the testing of 1.75 million recruits.

There were three types of test: literate recruits were given a written test called the Army Alpha, men who were illiterate or who failed the Alpha were given a pictorial test called the Army Beta, and failures on the Beta were to be recalled for an individual spoken examination. The Alpha had eight parts made up of the items we recognise today as IQ tests, such as analogies, filling in the missing number and unscrambling a sentence. The Beta test had seven parts including number work and the picture completion task shown in Figure 5.3. Each test took less than an hour and could be administered to large groups.

Yerkes asserted that the tests measured ‘native intellectual ability’ (cited in Gould, 1981, p. 349) but the level of cultural and educational knowledge required is illustrated in the examples given below.

- ‘Washington is to Adams as first is to...’
- ‘Crisco is a: patent medicine, disinfectant, tooth paste, food product’



Part six of examination Beta for testing innate intelligence.

Figure 5.3 Some items from the pictorial part of Yerkes' mental tests

Source: Gould (1984)

- ‘Christy Mathewson is famous as a: writer, artist, baseball player, comedian.’

There were a number of problems in the administration of the tests. In particular, many who were illiterate in English were still allocated to the Alpha test and so scored zero or near to zero. This created a systematic bias in the test since recent immigrants who had a poor grasp of English, and Black men who had not been given much, if any, formal education were unable to score on the Alpha test. Another problem was that even the Beta test required the use of a pencil and the writing of numbers, and many men had never held a pencil in their lives.

The tests generated a lot of interest, and by 1921 when Yerkes published his findings he was able to refer to ‘the steady stream of requests from commercial concerns, educational institutions and individuals for the use of army methods of psychological examining or for adaptation of such methods to special needs.’ Mental testing and psychology had achieved the credibility that Yerkes wanted.

Gould (1981) reports that three ‘facts’ were created from the testing data;

- 1 The average mental age of white Americans was about 13. Unfortunately, this had been defined as the intellectual level of a moron, so the tests appeared to indicate that the average American was a moron. (Note: I advise readers to hold their ethnocentrism in check at this point.)
- 2 European immigrants could be graded by their country of origin.
- 3 The average score of Black men was lower than the average score of White men.

These three ‘facts’ can be adequately explained by the administration difficulties of the testing and the level of literacy of the groups of people taking the tests. In fact, a re-analysis of the data showed that performance depended on the length of time a person had lived in America, suggesting that culture and language played a large part in test performance. However, a much more sinister explanation was given for the results. It was argued that White people were superior to Black people, and that Americans were superior to many European peoples.

Once again, political beliefs triumphed over scientific analysis and the eugenic explanation took hold. One of the consequences of this was the passing of the Immigration Restriction Act in 1924 by the American Congress, which selectively stopped certain national groups from emigrating to the USA. The scientists who supported the eugenics argument lobbied the politicians and, according to Gould, ‘won one of the greatest victories of scientific racism in American history.’ (1981, p. 352)

Statistical techniques and intelligence testing

The need to interpret mental tests led to the development of a range of statistical techniques. Karl Pearson developed the correlation, and Charles Spearman developed factor analysis. Factor analysis has a particular importance in intelligence because it is used to suggest that there is such a thing as intelligence. This is not as obvious as it sounds. Just because we have a word for something, it does not mean that it exists. What we can observe is people behaving intelligently, but that does not mean they have a thing called intelligence.

If we give people a range of mental tests they do not do equally well on all of them. For example, people who are good at maths are not necessarily good at English or art. We could explain this by saying that these skills are entirely different, or we could say that there is an underlying skill of intelligence that affects our performance on all mental tests. The idea that intellectual differences are inherited depends on the simple idea that there is a single quality that affects performance. If there is no single quality of intelligence, then it clearly can not be inherited. This is the importance of the statistical analysis of test scores. If we can show there is a single intellectual quality then we can argue that intellectual differences are the result of genetic factors and we can suggest eugenic solutions. If, on the other hand, there is no single quality we can call intelligence, then the whole argument about inherited differences collapses.

It is important to note that Spearman, like many other intelligence testers of his time, favoured eugenic solutions and so was looking to establish that intelligence had a single characteristic. He developed the technique of factor analysis which analyses the relationship between a large number of correlations. He took the data from a range of mental tests and showed that there was one factor that explained the variations in scores on most of the tests. He

called this factor ‘g’ (for general intelligence). This all looks very impressive and the complexity of the mathematics means that many people find it hard to argue against this finding. Gould (1981), however, provided a critique of Spearman’s factor analysis of mental test data that challenges the existence of ‘g’ and therefore the existence of a underlying quality of intelligence. One of the arguments mounted by Gould is that you can use the same data to show the existence of several factors rather than just one.

The statistics, then, are not conclusive and the controversy remains about whether there is an underlying characteristic. Even if there is an underlying factor, it is by no means clear that we can call it intelligence. It could well be some cultural or environmental variable, for example, that affects our performance on all these tests.

Intelligence testing and race

As mentioned above, early IQ tests were used to suggest differences between different racial groups. This is an argument that has rumbled on for over 100 years without ever resolving itself. There are many problems with attempting any exploration of racial differences (see Chapter 4 on bias in psychology), and none more so than the issue of intelligence. This is particularly controversial because differences between one group and another are often presented as differences in ability rather than differences in performance. In truth, the only evidence we have are performance scores, and we have to explain why groups perform differently. This could be due to a range of cultural, educational and motivational factors, and these explanations are much more plausible than any genetic explanations for the reasons described by Jones (see Chapter 4 of this text).

There is also the matter of how the genetic effect is calculated, and the statistic that is commonly cited is **heritability**. This statistic estimates how much the variation within any given population is due to genetic factors. It does not, however, tell us about why two populations will differ and so contributes nothing to our understanding of this issue (see Rose *et al.* 1984). It also fails to tell us anything about how much genetics affects the characteristics of an individual.

Problems like these make assertions about racial differences very difficult and, in fact, one of the leading psychometricians, Kline (1991) suggested that:

The only advantage in setting out the different scores on IQ tests of racial groups is to give ammunition to those who wish to decry them. It adds nothing to theoretical understanding or to the social or educational practice.

(Kline, 1991, p. 96)

IQ testing and expectation effects

One of the concerns about the use of psychometric tests, and IQ tests in particular, is that the results of the tests produce expectation effects. If, for example, an IQ test tells you that someone is a genius you might interpret their subsequent behaviour in terms of the test result and so expect them to behave in an intelligent way. It is fair to suggest that individuals may respond to the expectations of others, so if people expect you to be intelligent you are more likely to act intelligently.

In a famous but controversial study, Rosenthal and Jacobson (1968) set out to discover whether pupils' performance over the course of a school year could be affected simply by the expectations that their class teachers had of them. A proportion of children in an elementary school in the USA were labelled as 'bloomers'; that is they were identified to their class teachers as children who, on the basis of a psychological assessment, were expected to develop particularly rapidly in terms of intellect over the course of the following year. Unknown to the teachers, the bloomers had been chosen at random from each of the classes in the school. That is, there was actually no difference at the outset of the study between them and their classmates. If at the end of the year they had actually developed more rapidly than their classmates then this could be put down to the expectations that the teachers had of them. Presumably if pupils can perform better because of high expectations, the more worrying effect of pupils performing worse because of low expectations is also likely to be at work in everyday life.

The results appeared to support Rosenthal and Jacobson's hypothesis. However, there have been numerous criticisms of the study. For example, the IQ test they used was not standardised for the age-range of children on which it was used. Replication studies have found mixed results, and the effect described by Rosenthal and Jacobson is not as robust as they thought.

Rosenthal and Jacobson thought that one person's expectation of another person can become a more accurate prediction of their behaviour just because the expectation is there. The process behind this idea is more complex than it appears at first thought. The process of expectancy includes the following processes:

- (i) teachers are assumed to have some expectations of their pupils, and that these expectations are based on some sort of evidence or hearsay;
- (ii) these expectations affect the behaviour of the teacher in a visible way;
- (iii) pupils perceive this 'communication' of the teacher's expectation;
- (iv) the pupil's understanding of their teacher's expectations must affect their performance in class.

So the difficulty in replicating the study might be due to failure at any of these four stages. However, there has been some empirical support for the idea of these expectation effects and a good example of this is a quasi-experiment carried out by Seaver (1973). This study looked at teachers' expectations of children that were based on having already taught a sibling of the child. The question was whether teachers would expect the younger sibling to be like the elder child and whether this expectation would affect performance. Many of us have had the experience of being compared at school to a cleverer/better behaved/more responsible elder brother or sister. The teachers seem to have expectations of how we will behave based on their experience of our siblings. The study by Seaver confirmed this, and found that in classes where the teacher had experience of the elder sibling the student's results were more similar to their sibling's than they were in classes where the teacher had no knowledge of the sibling.

Summary

The testing of intelligence continues to generate a number of controversies including:

- whether there is such a thing as intelligence;
- whether we can accurately measure intelligence;

- how much genetic factors account for the differences in individual scores on IQ tests;
- how much genetic factors account for the differences in scores on IQ tests between groups of people;
- the political consequences of IQ data;
- the interpretation of complex statistical procedures;
- the effects of expectation on IQ performance;
- the problems of inferring intellectual ability from performance on IQ tests.

Measuring personality

Cattell and Butcher (1968) suggested there are three ways to approach the study of personality:

- 1 the literary approach in which the playwright or novelist describes people in a way that gives the reader an insight into human behaviour and experience;
- 2 clinical observation in which the clinician attempts to systematically classify normal and abnormal personalities;
- 3 the statistical tradition based on correlational methods and, in particular, factor analysis.

In this section, we will look at this third approach and consider some of the issues around measuring and quantifying personality characteristics. The measurement of personality does not create the same level of controversy as does the measurement of intelligence. The main reason for this is that the political impact of personality scores is not as great as IQ scores. Many of the same controversial issues, however, can be found in the measurement of both qualities, for example, the use of factor analysis, the validity of the tests, the issues surrounding the inheritance of personal qualities, and the historical connection between psychometric testing and eugenics.

Personality questionnaires

The popular way of measuring personality is through a questionnaire. The advantages are obvious in that questionnaires are relatively easy to construct,

they are relatively easy to use with people, and it is relatively easy to establish norms. In short they are psychometrically efficient (Kline, 1993). The big problem to consider, however, is the one of validation. What do these reliable tests measure'?

The problem comes with the use of factor analysis. When factor analysis is used to look at intellectual performance there is a theory that predicts how many factors there will be (in this case, one). In personality theory, however, the expected number of factors is by no means so obvious. Nobody would predict that all the differences between individuals can be explained by just one feature of personality, but how many are there? It is possible to come up with a number of factors just by the statistical method of factor analysis. Say, for example, that we put in the data from all our personality questionnaires and come up with four main factors that can be used to explain the data. The problem for the psychologists is to give these factors a name and show that they really exist. The alternative explanation is that the factors are just a trick of the numbers and have no existence outside the computer of a statistician.

Theories of personality

Factor analysis was used by Cattell to identify sixteen factors of personality, and he claimed that the 'sixteen personality-factor questionnaire' (16PF) measured the major factors of personality and could account for the differences between people. Cattell, therefore, discovered his sixteen factors through statistical means and the debate is about whether these are real factors that exist in people.

Another theory of personality that is also based on psychometric testing is that of H.J. Eysenck, in 1947, who described two major dimensions of personality: (i) introversion/extroversion, and (ii) neuroticism/stability. He argued that these are fundamental traits which account for a large proportion of individual differences. Eysenck's theory has generated a lot of research, possibly because it has two factors rather than sixteen.

Other researchers have identified different combinations of personality traits. McCrae and Costa (for example, 1985) suggested that there are five basic factors of personality, and there is now a weight of evidence that supports this view (Kline, 1993). These factors are derived from factor analysis, but the same factors appear in a range of studies giving support to the argument that the factors

Table 5.1 The five robust factors

1	Extroversion	Sometimes referred to in a slightly wider form as 'surgency'.
2	Emotional stability	Broader in scope than Eysenck's neuroticism, but essentially similar.
3	Agreeableness	Including traits of generosity, friendliness and interpersonal criticism
4	Control	Including traits of disorganisation, meticulous approaches to work, etc.
5	Culture	Including the traits of curiosity, creativity, intelligence and knowledgeability.
Source: McCrae and Costa (1985, p. 716)		

represent real psychological variables rather than just a statistical mirage. The five factors are shown in Table 5.1.

Expectation and the Barnum Effect

A serious problem with the Eysenck Personality Inventory (EPI), and also with similar tests, is to do with the expectation of the people taking the tests. Furnham and Varian (1988) looked at how people predict and accept their own test scores. In their first study, they asked undergraduates to try and predict their own and a well-known other person's personality scores on the EPI. They were fairly good at this. Then, in a second study, they gave some undergraduates false feedback about their scores after they had completed the EPI. The undergraduates were more likely to accept positive feedback (for example, 'you are a warm and considerate person') as accurate than negative feedback (for example, 'you are a miserable, smelly moron'), even though it did not have any connection with their actual scores.

This leads us to an inevitable discussion of the Barnum Effect (so named after the famous American showman, P.T. Barnum). In brief the Barnum Effect refers to a powerful tendency to believe information given to us about our personal qualities. This can be used to good effect by, for example, fortune tellers, astrologers, and handwriting 'experts'. If the 'expert' can say what

people are prepared to accept, and can phrase it in such a way that it implies some intimate insight, then there is a good, if dishonourable, living to be made.

An early demonstration of this was provided by Forer (1949) in a classroom demonstration of gullibility. Forer described a personality test to his students and allowed them to persuade him to let them take the test (the first rule of a successful con is to appear reluctant!). Thirty-nine students completed the test. One week later, each was given a typed personality sketch with their name on it. The researcher encouraged the class to keep the results confidential, and the students were asked to indicate whether they thought the test results were accurate. In fact, they had been given identical personality sketches (see Table 5.2) which had no relationship to their test responses, yet all of the students rated the test as a perfect or near-perfect tool for investigating personality. This is a demonstration of the Barnum Effect. If we return to the study by Furnham and Varian (1988), we could argue that this is a demonstration of this same Barnum Effect, and that the EPI is successful, at least in part, because it provides plausible personality sketches.

Other problems with personality questionnaires

Kline (1993) identifies a number of sources of error in personality questionnaires, including:

- 1 acquiescence: people have a tendency to agree with items regardless of the content.
- 2 social desirability: people prefer to put themselves in a good light and so have a tendency to respond in a way that makes them appear all right. For example, we are unlikely to admit to only rarely washing even if that was true.
- 3 middle categories: many questionnaires ask people to respond on a five-point scale, and there is a tendency for people to use the middle value.

Personality tests are used extensively in occupational and clinical settings. A sophisticated technology has been developed to ensure the reliability of the tests and to produce evidence for their validity. Sometimes, however, the belief in the existence of certain behavioural patterns or personality characteristics can be more

Table 5.2 The statements used by Forer (1949) in his demonstration of the Barnum Effect

1	You have a great need for other people to like and admire you.
2	You have a tendency to be critical of yourself.
3	You have a great deal of unused capacity which you have not turned to your advantage.
4	While you have some personality weaknesses, you are generally able to compensate for them.
5	Your sexual adjustment has presented problems for you.
6	Disciplined and self-controlled outside, you tend to be worrisome and insecure inside.
7	At times you have serious doubts as to whether you have made the right decision or done the right thing.
8	You prefer a certain amount of change and variety and become dissatisfied when hemmed in by restrictions and limitations.
9	You pride yourself as an independent thinker and do not accept others' statements without satisfactory proof.
10	You have found it unwise to be too frank in revealing yourself to others.
11	At times you are extroverted, affable, sociable, while at other times you are introverted, wary and reserved.
12	Some of your aspirations tend to be pretty unrealistic.
13	Security is one of your major goals in life.

powerful than the scientific evidence. For example, the study of coping has been dominated by the Folkman and Lazarus (1990) description of coping styles (see Table 5.3). Their eight distinct styles of coping were derived from factor analysis, and their model is extensively applied and extensively researched on. A close look at the data, however, reveals that the eight coping styles are not as distinct as Folkman and Lazarus suggest, and it would be a more accurate

Table 5.3 The eight coping strategies identified by Folkman and Lazarus through factor analysis but challenged by Ferguson and Cox

<i>Problem-focused strategies</i>	
1	Confrontive coping
2	Planful problem solving
<i>Emotion-focused strategies</i>	
3	Distancing
4	Self-controlling
5	Seeking social support
6	Accepting responsibility
7	Escape-avoidance
8	Positive reappraisal

Summary

The focus of this book is to consider controversies in psychology so the above brief section on personality testing does just that. I have only been able to give a flavour of the rigor of test development and extensive use that is made of psychometric tests. The enduring controversies include:

- labelling: is it appropriate to reduce a person’s personality to a simple numerical index?
- stability of personality characteristics: even if we can obtain an accurate measure of a person’s personality, does the result reflect their enduring personality characteristics or just how they were feeling when they took the test?
- validity: do the tests measure what they claim to measure, and in particular, has factor analysis given us descriptions of real personality variables or are they just statistical inventions?

- inheritance: how much of the differences in personality characteristics can be accounted for by genetics?
- Barnum Effect: are we gullible to the words of experts and can we put a lot of the effectiveness of personality tests down to this effect?

Design a psychometric test to measure stupidity (not necessarily the opposite of intelligence) or charmingness.

- Identify five items for the test
- Describe how you will assess the reliability of the test
- Suggest how you might assess the validity of the test

Review exercise

Further reading

- Gould, S.J. (1981) *The Mismeasure of Man*, Harmondsworth: Penguin. A classic text that goes through the ways that psychologists have attempted to measure intelligence.
- Kline, P. (1993) *Intelligence: The Psychometric View*, London: Routledge. A clear and concise account of the psychometric view of intelligence.

