PSYCHOMETRIC TESTING IN THE WORKPLACE

- *Extent of psychological testing at work*
- *How soundly based are the techniques?*
- *Are they being misused?*

Employers are making increasing use of psychological tests for assessment and recruitment purposes. These tests can provide much useful information when used properly, but there are increasing concerns that this is not always the case. For instance, some uses of tests in recruitment procedures have been shown to be discriminatory, and there have been allegations that psychological tests have also played a part in selecting candidates for redundancy.

This report examines the scientific basis of the different types of psychological tests commonly used by employers, their strengths and weaknesses and the issues that arise.

POST TECHNICAL REPORT 59 April 1995

POSTreports are intended to give Members an overview of issues arising from science and technology. Members can obtain further details from the PARLIAMENTARY OFFICE OF SCIENCE AND TECHNOLOGY (extension 2840).

CONTENTS

THE TESTS

Psychometric (or psychological) tests are designed to assess and quantify a wide range of individual characteristics such as intellectual ability, natural aptitudes, personality traits and personal interests. Hundreds, if not thousands of such tests have been devised to date and a detailed description of all these is beyond the scope of this short report. Nevertheless, most of the important tests can be assigned to the broad types summarised in **Box 1** (page 2).

These include:

- **occupational tests** such as cognitive tests, work sample and other (aptitude and ability) tests;
- **self-assessment techniques** such as personality questionnaires and interest inventories;
- other techniques commonly used by employers (e.g. job analysis);
- tests more often found in educational or clinical environments.

When used as selection tools, employers commonly combine tests assessing different characteristics to provide a more complete picture of the individual and his/her suitability for the job. A typical combination might consist of a personality assessment and a general ability/aptitude test covering numeracy, literacy and reasoning. These might be supplemented by a specific work sample test to evaluate performance on the most important parts of the job, already identified through a job analysis exercise. Some (usually larger) organisations have special **assessment centres** where the tests are conducted, and where the results of the tests can be set alongside performance in interviews, group discussions, etc.

Although most of the tests outlined in Box 1 originated as 'pencil and paper' exercises, and are still available in that form, an increasing number of computer-based versions are being developed. These have the advantage of allowing **adaptive** tests to be designed (i.e. where the next question is selected according to the answer given to the previous one), as well as allowing the results to be worked out and interpreted much more swiftly. Other tests involve complex materials or specialised apparatus - for instance a manual dexterity test might require individuals to insert thin steel rods into a specially designed steel frame.

Tests differ substantially in the amount of theory and interpretation involved. Some yield quantitative scores (of literacy, numeracy, etc.) that may be used to compare an individual's performance against other groups of people who have also taken the test. Others (e.g. personality questionnaires, interest inventories) are not really 'tests' at all, but are rather **self assessment** exercises which are used to build up a profile of an individual's character, behaviour, interests, etc., and are often expressed as a qualitative and somewhat subjective 'pen-portrait'.

Organisations wishing to use psychological tests have to decide whether to develop (or commission) their own (**customised**) tests, or whether to buy them '**off-the-shelf**' from one of the many **test publishers**. In general, developing customised tests is expensive and

Box 1 TYPES OF PSYCHOLOGICAL TESTS

There are many different types of psychometric tests, and selecting the most appropriate one requires employers to conduct a **job analysis** exercise. One way of conducting such an exercise is to use **job analysis questionnaires**, which allow the job to be broken down into its component parts and the skills needed to be identified. These are useful in their own right (e.g. for setting up salary structures), but the skills so identified can then guide the choice of relevant psychometric tests to be used during the selection process. In general, no single test (of whatever type) is likely to give all the information required, and decisions (on recruitment, redundancy, counselling, etc.) are usually based on the results of more than one test taken in conjunction with information from other sources (e.g. interviews).

Occupational Tests

These measure various mental and physical skills and aptitudes. Tests measuring mental skills are often called **cognitive tests**, and these are generally designed to assess one or more of 7 main skills (**verbal, numerical, spatial, memory, reasoning, word fluency** and **perceptual speed**) arising from psychological theory. Because they measure 'core' skills required for most types of job, cognitive tests are often regarded as being applicable across a wide range of occupations. Other widely used occupational tests include **work sample tests**. These are derived in a more empirical fashion from job analysis exercises, and are designed to mimic key aspects of a job under test conditions. For instance, office workers may be given 'in-tray' exercises, firemen map-reading tests, computer operators may be tested for accuracy and speed of data input, etc. More general tests of physical aptitudes such as manual dexterity may also be used where job analysis has shown this to be appropriate.

Self-assessment Questionnaires

Personality assessments are self-assessment exercises, where candidates are asked to report on their attitudes, beliefs, emotions, feelings and behaviour. There are many different '**dimensions**' to personality - a person may be described as creative, innovative, decisive, adaptable, resilient, etc. - and a personality questionnaire may assess any combination of these. Some tests assess the full range of dimensions, others focus on dimensions that are thought to be particularly relevant to the workplace (e.g. the Occupational Personality Questionnaire includes dimensions such as 'forward planning', 'change orientated' and 'conceptual'). Recent personality theory suggests that all the possible different dimensions may be boiled down to just five fundamental ones. Although there is some dispute as to exactly what these '**Big 5**' should be called, the following terms are widely used; **extraversion, agreeableness, conscientiousness, emotional stability** (as opposed to neuroticism) and **openess** to experience or culture. Some personality tests focus on assigning people to fundamental **personality types** by assessing some combination of these Big 5 dimensions.

Interest inventories are questionnaires that provide information on a person's interests and preferences. These are then matched against responses from people in a wide range of different occupations, and areas of strong similiarity or dissimiliarity may be used as the basis of advice concerning career decisions.

Other methods

Biodata analysis is a statistical scoring system that uses biographical data as the basis for predicting work performance. A range of other tests may also be used in the workplace to assess factors such as **creativity, trainability** and **stress / anxiety**. Psychological tests are also widely used in **educational** (attainment tests assessing grammar, spelling, arithmetic etc.) and **clinical** (e.g. diagnostic tests assessing mental health) settings. Techniques such as **graphology** (handwriting analysis) and **astrology** (horoscopes) are not recognised by professional bodies such as the Institute of Personnel and Development and British Psychological Society.

time-consuming, and thus an option available to only the larger organisations. Most thus use off-the-shelf products, selecting from the wide range available to build up a combination of tests designed to meet the employer's perceived needs.

WHO USES TESTS AND WHY?

Occasional surveys of employers in both public and private sectors by academic researchers suggest that most large companies and around half of local authorities use psychological tests in the workplace, and that their use has increased slightly in recent years (**Figure 1**). For instance, the proportion of local authorities using tests increased from around 42% in 1986 to just over 50% in 1991. In the business sector, 63-68% of large companies (Figure 1) use cognitive tests and an increasing proportion are also using personality tests (57% in 1991 compared with 47% in 1988). Smaller organisations (not shown in Figure 1) tend to rely less on such tests, with one survey in 1988 showing around 16% and 22% of companies using cognitive and personality tests respectively.

The reasons for using psychometric tests are shown in **Figure 2**, which shows recruitment to be the main area of application (e.g. of the local authorities using cognitive or personality tests in 1991, virtually all (98%) used them for recruitment and selection). Employers also use psychological tests for a wide range of other purposes within their organisations, including internal selection, team building, re-organisations, performance appraisal, personal development and career counselling. Most of these uses are not contentious, but where tests are used to inform decisions on redundancy or recruitment, they can be subject to close scrutiny and challenge.

PROS AND CONS OF TEST USE

Publishers claim a number of advantages in using tests rather than relying just on interviews and other conventional selection methods. Tests are seen as:

- more objective and unbiased;
- better at predicting future performance;
- easier to monitor;
- more cost-effective.

Figure 1 EXTENT OF USE OF PSYCHOMETRIC TESTS

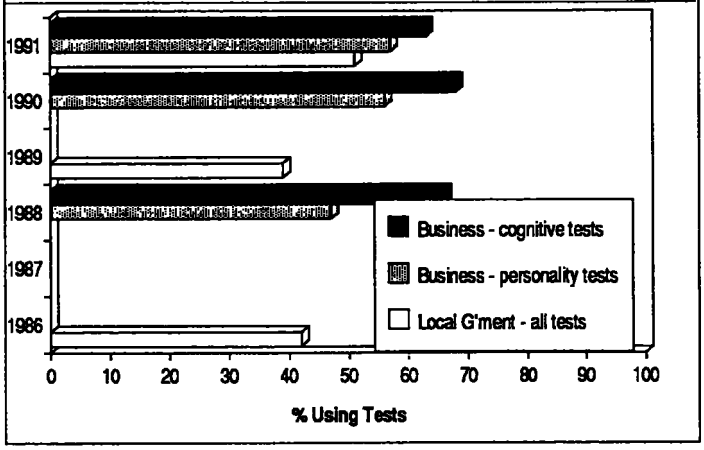


Figure 2 REASONS FOR USING PSYCHOMETRIC TESTS

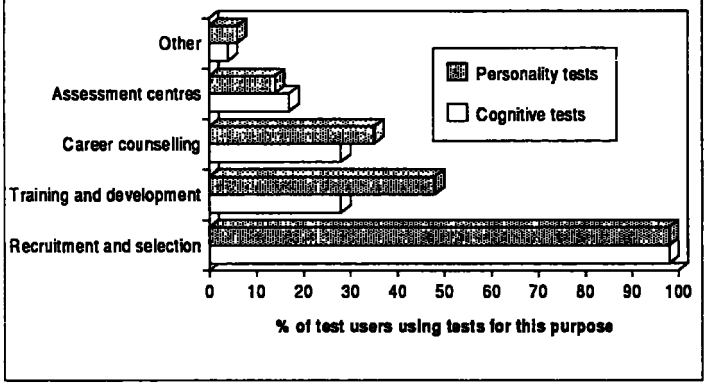
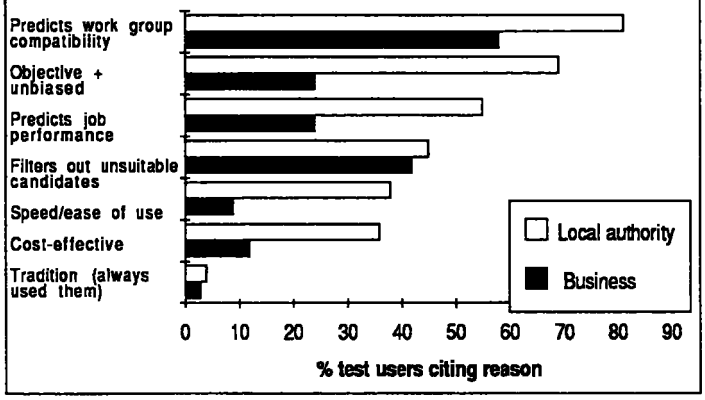


Figure 3 PERCEIVED BENEFITS IN USING PSYCHOMETRIC TESTS



In terms of the advantages actually desired and expected by the test users, their **predictive potential** is one of the most frequently cited reasons for using them (**Figure 3**). Thus 81% of local government users of personality tests expected them to predict work group compatibility, and 55% to predict future job performance. Other reasons given for using personality tests include because they are seen as being objective and unbiased (by 69% of local authority users) and as an effective way of screening out unsuitable candidates (by 45% of local authority and 42% of business users). In general, surveys suggest that local authorities have higher expectations of the tests than their colleagues in the business sector.

In seeking to achieve these positive outcomes however, both test publishers and users need to be alert to the potential pitfalls awaiting the unwary user of psychological tests. A poorly-designed or poorly applied test may not reveal the important characteristics required or, worse still could discriminate between candidates on a false basis.

The most obvious consequence of using inappropriate tests (or using tests wrongly), is that the organisation may end up selecting the wrong people or making personnel development decisions on a misleading assessment of their needs. At best, this is merely a waste of time and money, but at worst it may lead to bad decisions from both the employer's and employee's point of view, or even legal action.

Legal complications follow in two sets of circumstances. Tests are not often used to select people for **redundancy** because employers are generally already familiar with the strengths and weaknesses of existing employees, and thus usually have a basis on which to make the necessary choices. Nevertheless, employers using psychometric tests for such purposes may face legal challenge before an Industrial Tribunal through a claim of unfair dismissal.

The second legal danger is if the tests can be shown to **discriminate unfairly** on the basis of race, gender, religion, etc., thereby infringing UK legislation (outlined in **Box 2**) outlawing such discrimination. Since tests are designed to discriminate (or at least differentiate) between different people, they can provide fertile ground for disagreements over whether such discrimination is 'fair' or 'unfair'. Indeed, the **potential** for unfair discrimination exists wherever consistent differences are found in average test scores between men and women, or between members of two different ethnic groups, and the onus is then very much on the tester to demonstrate fairness. In practice, this requires that the test be shown to be both **relevant** (i.e. it measures a specific skill or attribute that is needed for the job) and **predictive** (i.e. that people scoring highest in the test will perform best in the job). Where no such justification is available, there is likely to be a presumption of unfairness, and the test user may end up in court facing claims of indirect race or sex discrimination.

HOW WELL DO TESTS PERFORM?

Assessing a test's performance is largely a matter of establishing its usefulness as a predictive tool (its **validity**). In practice this involves investigating the strength of the link between test scores and some measure of subsequent job performance (annual review rating, productivity, attendance, etc.). In the case of off-the-

BOX 2 UK LEGISLATION RELEVANT TO SELECTION AND ASSESSMENT

The **Race Relations Act (1976)** makes racial discrimination unlawful in Great Britain. Two types of discrimination are defined, **direct** (where a person is treated less favourably on racial grounds) and **indirect** (where the same treatment is applied to all racial groups, but the effect is discriminatory). In practice, psychometric tests may be a cause of indirect discrimination if they produce differences in average scores between racial groups. The Act established a Commission for Racial Equality to help enforce the legislation.

The **Sex Discrimination Acts (1975 and 1986)** also applies to Great Britain and makes it unlawful to treat anyone, on the grounds of sex, less favourably than a person of the opposite sex would be treated in the same circumstances. These provisions apply to employment, education, advertising, and the provision of housing, goods, facilities and services. In employment and related advertisements, it is also unlawful to discriminate because a person is married. The Equal Opportunities Commission was created to ensure effective enforcement of the Act.

The **Fair Employment Act (NI), 1989** applies only to Northern Ireland and contains specific provisions aimed at the prevention of religious discrimination. It requires employers to follow a recommended recruitment system, and to monitor the 'religious composition' of its workforce.

The **Disabled Persons (Employment) Acts (1944 and 1958)** require employers of 20 or more people to employ a quota (currently 3%) of registered disabled people, and larger organisations (250 or more employees) to publish information on their policy of recruitment of disabled people. Although it is not illegal to fail to meet the 3% quota, where this is the case it is illegal to recruit a person who is not registered disabled until the quota has been met.

shelf tests (which account for most testing in the workplace), the publisher should have conducted such studies as part of the test's development, and this validation should be part of the test documentation. Nevertheless only a limited number of job types, groups and other circumstances can be covered in pre-sale evaluation, and these will not necessarily match the circumstances in which the test is applied. Consequently it is important to know how far validity in one application is indicative of a more general validity, or whether every conceivable application needs to be separately validated.

Opinion on this question has shifted considerably in recent years. Most of the studies of test performance conducted by occupational psychologists up to the late 1970s gave inconsistent results, so that a given test might prove to be highly predictive in some studies, but much less so in others. Overall, researchers and personnel managers concluded that the performance of psychometric tests varied too much from one application to another to allow general conclusions to be drawn on their relative value.

This view changed during the 1980s, when various studies were re-examined using **meta-analysis,** a statistical tool for analysing several smaller studies as a single, larger, group, thus minimising the effects of imperfections in the original study designs. Meta-analysis showed that much of the inconsistency observed in the earlier studies was a result of the small sample sizes used, and once this was taken into account some of the tests were found to predict job performance well across a wide range of different occupations.

Meta-analysis and other statistical methods have now been applied to a wide range of different selection methods, which has allowed professional bodies such as the British Psychological Society (BPS) to draw conclusions on their relative effectiveness. Overall, all the commonly used psychological tests show some degree of predictive value, in contrast to selection methods such as handwriting analysis (**graphology**) and horoscopes (**astrology**) which show zero validity. Thus a recent BPS review concluded that graphology was not a viable means of assessing a person's character or abilities.

One way of measuring the predictivity and value of a test is to see how much of the future variation in job performance between candidates is predicted by the test results (in statistical terms, known as the variance). Although figures vary according to the exact nature of the statistical tools used, tests which attempt to sample important elements of the job (i.e. **job simulations)** are among the most highly predictive, accounting for up to 30% of variation in job performance. Close behind are the more general tests of mental ability (**cognitive tests),** which can account for up to 28% of variation in job performance, and which have been shown to be applicable to a wide range of different occupations. Interviews where the questions have been derived from job analysis exercises (**structured interviews**) are fairly predictive (11-12%), and much more so than traditional (**unstructured**) job interviews. **Personality questionnaires** do show some predictive value, but are in general among the least predictive of the commonly used psychological tests, typically accounting for no more than 4-5% of variation in overall job performance. More recent work however has shown that personality questionnaires are better when measured against specific competencies rather than ratings of overall job performance, when the best can account for some 10-11% of variation in a specific competency.

Such general conclusions may prove useful in guiding users in their choice of tests, and the tests can be shown to significantly improve the objectivity and increase the ultimate success rate of a selection procedure. However, the tests only measure certain characteristics, and the moderate 'variances' encountered mean that there are many other factors not revealed by the tests which

contribute to future performance in the job. Thus, good performers in the test may still fail, just as some of those doing poorly might well succeed. For this reason, the results need to be seen in appropriate perspective and selection decisions should never be made on the basis of a single test, but should consider information from as wide a combination of sources as possible (e.g. tests, interviews, references).

In addition, even the most predictive test can produce meaningless scores if it is not properly administered by suitably qualified personnel, and considerations of best practice are discussed in the next Section.

CURRENT PRACTICE

Companies wishing to deploy psychometric testing methods can supplement the advice given by the test publishers with guidance from professional bodies such as the British Psychological Society (BPS), Commission for Racial Equality (CRE), Equal Opportunities Commission (EOC), and the Institute of Personnel and Development (IPD). These provide a wealth of advice on the appropriate tests to use and how they should be applied, and some of the main points are summarised in **Box 3**, covering:-

- Deciding what to measure,
- Choosing the right test(s),
- Making sure they work,
- Applying them properly,
- Interpreting the results,
- Monitoring their effect,
- Training needs.

Practice does not always follow the ideal recommended by the professionals however. For instance, although careful procedures are laid down for deciding on the test appropriate to the job, research commissioned by the EOC on 13 organisations using tests in a variety of situations showed that most tests were chosen on the basis of trust, with only two organisations having conducted formal job analyses prior to deciding which tests to use.

Companies also need to be aware of potential limitations in the information provided by the test publishers. Providing information on **content and level** (Box 3) is relatively straightforward. Most commercially available tests are also **reliable** in the sense that the same person will register similar scores in repeat sittings of the same test. However, some of the personality questionnaires are more variable in this respect especially if there is a longer time between taking tests.

Demonstrating **fairness** however, is a greater challenge. In order to do so, test publishers need to collect data on the performance of large numbers of individuals from different educational, social, ethnic, etc. backgrounds to show that performance is not biased in favour of one group or another. With some (e.g. ability) tests, such studies are an essential part of the test's development and validation, and the test documentation includes tables of 'norms' to allow individual scores to be ranked in relation to the population at large, as well as various groups within the population. In order to reduce real or perceived bias however, the composition of the group used to obtain these norms should match as closely as possible (in terms of ethnic origin, gender, educational background, etc.) the mix of people who will actually be tested. Ranking test scores by comparing them with wholly inappropriate norms is one of the most common causes of unfair discrimination.

However, simply knowing (say) that a candidate is in the top (or bottom) 10% of scores for a particular ability test, or has been assigned to a particular personality category, is of little use to an employer unless this information predicts future job performance. As discussed in the previous Section, publishers do **validate** tests by linking test scores to relevant aspects of job performance, and the best test publishers will supply quite detailed and comprehensive information on validity. Even so, when it comes to justifying a particular test in a particular set of circumstances, the burden of proof lies with the **test user** and they will have to provide the necessary evidence of validation in the event of any complaints (e.g. of unfair dismissal or sex or race discrimination) arising from use of a test. For this reason, most guidance on best test practice strongly recommends users to conduct their own validation studies before using test scores as the basis of selection decisions.

Conducting such studies however, can present test users with serious practical problems. For instance, the CRE guidelines state that tests should be validated on all the main ethnic groups likely to take the test, and that group sizes of at least 100 people are required for statistically meaningful results. Mustering groups of this size is beyond the means of all but the largest employers and there is some disagreement among occupational psychologists whether such measures are needed; this is discussed further in the final Section.

Given the practical problems involved, it is perhaps not surprising that - despite the guidance to the contrary - many organisations do not conduct further validation studies on the tests they use and rely solely on evidence of validity supplied by test publishers. Even where users attempt validation studies, they are often abandoned as being too time-consuming and impractical, often requiring a level of statistical analysis far above that available in most personnel departments.

Box 3 USING TESTS PROPERLY

1 DECIDING WHAT TO MEASURE

This essential first step involves an analysis of what the job actually entails, and what attributes are required to do it well. Large organisations may have in-house experts to carry out such job analyses, but smaller firms use outside consultants or test publishers. Conducting a thorough **job analysis** not only assists companies in selecting individuals with the right abilities to do the job, but also helps them to justify the tests they have used in the event of any claims of unfair rejection from disappointed candidates.

2 CHOOSING THE RIGHT TEST

Once the necessary skills or attributes have been identified, factors determining which tests to use include:

- **content** (tests should measure specific attributes or skills identified in the job analysis);
- **level** (is the level of difficulty at which a skill is measured appropriate to that encountered in the job?);
- **reliability** (does the test consistently measure the same thing throughout its duration, and are the results reproduceable?);
- **fairness** (what measures have the publishers taken to ensure that the test does not unfairly disadvantage women, ethnic minorities, etc?);
- **validity** (does the test measure what it is supposed to, and in what way are the results useful?).

Most tests are bought off-the-shelf from test publishers, and reputable suppliers provide users with comprehensive manuals on the above. Where suitable off-the-shelf tests are not available however, some organisations (particularly larger ones) may design (or commission) their own. Since the content and level can be tailored to suit the exact needs of the organisation, such ('bespoke') tests are often more relevant to the job in question. However, the necessary development studies may take 6 months or more.

3 MAKING SURE THE TESTS WORK

Although test publishers go to considerable lengths to demonstrate the validity of a test before putting it on the market, it cannot be assumed that a test that works well in one situation will automatically work well in all others. Professional (e.g. BPS, IPD, EOC and CRE) guidance thus strongly recommends test users to conduct their own validation studies. This may be done in one of two ways:

- tests may be given to existing employees and scores checked against against measures of job performance, or;
- new recruits can be given the test without using the results for selection, and the scores compared against subsequent performance.

4 ADMINISTRATION AND INTERPRETATION OF TESTS

The way in which tests are presented, administered and interpreted also influence the outcome. Candidates who have never taken tests before often find the experience daunting and it may help to allow candidates to have a practice session. When selecting the candidates on the basis of their test scores, one method is the **'top-down'** approach, where candidates are chosen from the top score downwards until the required number have been obtained. An alternative is to set a **'cut-off'** - chosen to represent a reasonable level of job performance - and accept all candidates who exceed this minimum score. In general, guidelines discourage users from placing undue emphasis on the results of any single test, recommending instead that test scores can be placed into context against tables of 'norms', and then considered in conjunction with other information (scores from other tests, performance in interviews etc.).

5 TRAINING

Training is very important . In practice, there are two levels of control over test users.

Professional accreditation. The BPS launched an accreditation scheme in 1991 which awards certificates of competency to administer ability tests. A **'Level A'** certificate requires a tester to demonstrate competency in these main areas:-

- **defining assessment needs** (e.g. job analysis);
- basic principles of **scaling** and **standardisation** (e.g. using tables of norms);
- **reliability** and **validity**;
- **deciding when** tests should be used;
- **administering** and **scoring** tests;
- making appropriate **use of test results**;
- maintaining **security** and **confidentiality**.

These main areas are subdivided into a total of 97 different elements, and testers wishing to hold a Level A certificate must prove their competence in each of these to a suitably qualified Chartered Psychologist.

More recently (1994), the BPS has launched another scheme (**Level B**) covering the competencies required to use personality tests. Applicants must have already passed Level A, and demonstrate (again to a suitably qualified chartered psychologist) competency in each of 9 additional main areas:

- **personality theory**;
- **personality assessment**;
- **administration** of tests;
- **interpretation** of tests;
- providing **feedback** on test outcome;
- **understanding** of different main approaches to personality testing;
- **validity** and **reliability**;
- use of **computer-based** assessments and computer-generated reports;
- **when and how** to use personality and interest assessments.

Test Publishers also operate a policy of only selling their products to suitably trained and qualified individuals. Most publishers accept the BPS certificates as sufficient evidence of appropriate training to supply users with at least some of their tests, but there is an increasing trend to require potential users to attend training courses run by the publishers themselves.

6 MONITORING

Guidance issued by the EOC and the CRE emphasises the importance of continuously monitoring scores in practice, so as to supply further data about the predictive qualities of a test, and identify any (race or sex) discriminatory problems that were not apparent in the validation studies. Accumulating information in this way means that the test user can take early and appropriate action (e.g. discard or modify the test) if and when such problems are detected. On the other hand, if monitoring continues to give a test a 'clean bill of health', then the users justification for using the test is progressively enhanced.

Table 1 RULE-OF-THUMB GUIDE TO VALIDITY / FAIRNESS OF VARIOUS SELECTION / ASSESSMENT METHODS

POOR	VARIABLE	GOOD
Interviews	Interest inventories	Job simulations
References checks	Personality questionnaires	Peer assessment
Handwriting analysis		Bio-data
		Assessment centre
		Ability tests

Source: EOC, 1988. 'Avoiding Sex Bias in Selection Testing'.

In order to assist users in their choice of tests, the EOC has published a rule-of-thumb guide to the combined validity and fairness of commonly-used methods, as shown in **Table 1**. In general, tests are fairer and more valid than some conventional selection methods such as (unstructured) interviews and reference checks, with those that assess specific skills (job simulations, aptitude tests) outperforming those assessing attributes such as personality. Guidance is also available through a regularly updated "Review of Psychometric Tests for Assessment in Vocational Training", a publication originated by the Department of Employment but now maintained by the BPS. The BPS also encourages users to seek professional advice by publishing a Register of some 7,000 Chartered Psychologists.

A particular problem in administering tests is where tests are given to people for whom English is a second language. The CRE guidance for employers points out that tests containing unnecessarily complex written instructions may unlawfully discriminate against ethnic minority candidates, and thus recommends them to be kept simple and brief, and be fully supplemented with oral instructions. Research on ethnic minority candidates commissioned by the CRE has also highlighted the importance of allowing questions to be asked during the test preamble in order to make sure that instructions have been fully understood. In the absence of precise instructions, the CRE research found quite fundamental uncertainties - for instance some candidates were unsure whether they were allowed to turn the pages of the test paper, and if so whether they could turn backwards as well as forwards!

Given the vast range of tests available, their complexity, the importance of using them correctly and the potential pitfalls awaiting unwary users, organisations such as the IPD and CRE recommend that all people involved in commissioning, evaluating or interpreting tests should hold the appropriate BPS certificate(s) (Box 3). Those involved solely in administering the tests (but not in any other aspect of their use) need not be certificate holders, but should have received the appropriate training recommended by the test publisher.

CASE STUDIES

Some of the potential weaknesses in the tests or the way in which they were applied or interpreted have led to actions in the Courts or before Industrial Tribunals.

Box 4 CRE ENQUIRY INTO SELECTION TESTS IN LONDON UNDERGROUND

Following the King's Cross disaster, and acting on the recommendations of the subsequent (Fennell) Enquiry, LUL undertook a large-scale reorganisation of its existing management structures. Part of this involved the creation of 250 new middle management posts, which were advertised in December 1988 and filled by February 1989. Despite the fact that some 29% of all applicants (and 28% of those selected for interview) were of ethnic minority origin, such groups received only 11% of job offers, and this led to allegations that the selection procedure used had indirectly discriminated against these applicants. These allegations were susequently investigated by the CRE, which published a report on the matter in October 1990 (Lines of Progress - An Enquiry into Selection Tests and Equal Opportunities in London Underground).

Two aptitude tests (of written communication and numeracy), a personality questionnaire and an interview were used to assess applicants for these jobs. Analysis of the interview and test results (Table 2) showed that ethnic minority applicants fared considerably worse than their white colleagues in both the aptitude tests and the interview (but not the personality test). An independent consultants review of the selection procedure included in the CRE report suggested that although the aptitude tests were of some predictive value, they were inappropriate in terms of:-

- **content** (the form of written communication tested did not match that required by a LUL manager);
- **level** (the tests were unnecessarily complex);
- **appearance** to the candidate (the content of the numeracy test could have been more relevant to the work at LUL);
- **administration** (insufficient time was given to allow candidates to familiarise themselves with the tests, no practice examples were given, too little time was allowed to actually complete the tests).

Table 2 SCORES ACHIEVED BY DIFFERENT ETHNIC GROUPS IN LUL's SELECTION PROCEDURE

ETHNIC GROUP	AVERAGE MEAN SCORE		
	Written	Numeracy	Interview
White	37.2	19.6	18.1
Asian	30.5	14.1	14.2
Afro-Caribbean	25.4	9.3	15.7

Source CRE, 1990. 'Lines of progress; An enquiry into selection tests and equal opportunities in London Underground'

One such example is the case of London Underground Ltd. (LUL), described in **Box 4**. Here, when challenged to demonstrate that selection procedures had not inadvertently discriminated against ethnic minorities, LUL were unable to defend their use on several grounds. They could not demonstrate that the tests were relevant since they had not been chosen on the basis of a thorough job analysis (the tests recommended by the external consultants were originally chosen in relation to an entirely separate exercise). Nor could the tests be shown to be highly predictive and fair, since no specific validation studies had been conducted because of pressures of time. Overall, the CRE concluded that the

selection procedure *"must have resulted in unlawful discrimination against some ethnic minority job applicants"*. This conclusion was accepted by LUL, which has since reviewed its selection procedures and reached out-of-court settlements with a number of the applicants.

Another, similar, case occurred in 1990, when 8 (ethnic minority) guards at Paddington Station took British Rail (BR) to court (supported by the CRE) alleging racial discrimination after failing train driver selection tests. The assessment process included various aptitude tests, a personality questionnaire, an interview, and tests for vigilance and attention (which had been developed by another European rail operator). A review of test performance subsequently showed that ethnic minority groups achieved lower average scores than white groups on the aptitude tests, and that this was particularly pronounced in the case of one (off-the-shelf) test of verbal comprehension. Like LUL, BR had not accumulated sufficient information from validation studies to justify their use of these tests under challenge, and eventually settled out of court. As part of this settlement, BR agreed to review its assessment process. Several points emerged from this review:

- the content of the verbal aptitude test was judged to be inappropriate to the job, and the test was dropped;
- the technical manuals supplied with published tests provided insufficient evidence on fairness and validity;
- the system by which scores were interpreted (using cut-off points but allowing a slightly low score in one test to be compensated for by a higher score in another) may have disadvantaged ethnic minority candidates, and the system was changed;
- ethnic minority candidates place more emphasis on accuracy than speed when taking tests, and their test performance may be improved by receiving guidance on test-taking strategy;
- access to a training programme improved the aptitude test scores of candidates who had previously failed the assessment process.

These cases illustrate the potential dangers of using tests straight off the shelf. In both cases, the difference in average scores between ethnic minority and white groups meant that the tests had the **potential** to be unfair. Under these circumstances, the onus is very much on the test user to show that they are fair and to provide strong evidence of a link between test scores and job performance. Neither of the companies in the cases above had been able to validate the tests in this way.

Some of the other pitfalls that may await the unwary test user are illustrated by the cases of **Southwark Council** and **Anglian Water**. Both these organisations have recently hit the headlines, attracting considerable comment in the media for allegedly using personality tests in the process of selecting people for redundancy. Over-reliance on personality tests for such purposes would be particularly controversial since, as discussed earlier, they are not a good indicator of likely future job performance (see Table 1). Yet in both these cases, some of the individuals earmarked for redundancy felt that their personality profile had played a significant role in the decision. Their complaints have been taken up by the public services union Unison, and both organisations may have to defend their selection procedures before Industrial Tribunals in the near future.

However, neither case is as straightforward as much of the media coverage suggested. Faced with significant overstaffing in its Revenues and Benefits Section, Southwark Council decided to undertake a staff assessment exercise. Although the Council had considerable in-house experience of staff assessment, it decided to use external consultants, and not to allow these consultants access to existing personnel records. In this way it was hoped that the exercise would be viewed by the staff affected as being independent and fair. As agreed with the Council, the consultants assessed all staff using occupational tests and interviews, with more senior staff also receiving personality questionnaires. However, the consultant mistakenly used a type of questionnaire (the California Personality Inventory) which was different from that originally agreed, and which the Council considered to be inappropriate for this type of exercise. The Council stresses that the results of this questionnaire were ignored in their final decision-making process, and thus played no part in the eventual shedding of some 50 jobs.

Anglian Water points out that its exercise was not solely concerned with redundancy, but part of a broader programme to focus on the competencies necessary to meet a changing business environment. Some candidates (364) for senior posts took personality and aptitude tests at an assessment centre, and the results were made available in a report to selection panels, and to each candidate prior to interviews. 2,700 employees applying for other posts completed a personality questionnaire to inform a personal report given to each individual at a career development workshop, and which was used later in discussions on competencies, change and self development. Final placement decisions also took into account relevant qualifications, skills and experience. The outcome of this process will be that around 900 fewer people will be employed, involving around 80 compulsory redundancies.

Both Southwark Council and Anglian Water thus argue that the selection processes used were fair, and can be defended before Industrial Tribunals should the need arise. But whatever the outcome of any such actions,

these cases highlight the importance of the **appearance** of the test to those people who have to take it. In general, employees more readily accept (and are thus less likely to question) tests that are obviously relevant to the job in question than those where the relevance is less apparent. In both these cases, the use of personality tests alienated some employees - who failed to see what possible relevance their answers to general questions on attitudes, beliefs, etc. could have on whether they kept their jobs - even though their employers argue that the test results played little or no role in the actual redundancy decisions.

ISSUES

The examples above raise a number of more general issues of potential interest to Parliamentarians concerning the tests and the way in which they are used.

Tests and Discrimination

One of the biggest problems facing test users is that ethnic minority groups can achieve lower average scores than white candidates in some occupational tests. As the case studies discussed earlier illustrated, under these circumstances it can be extremely difficult even for committed equal opportunities employers to defend their selection methods. This raises a number of issues relating to the adequacy of guidance on best test practice, as well as the circumstances of test use.

One issue is that there is no objective definition of **fairness** - in other words, just how much lower do average test scores for ethnic minority groups have to be for a test to be considered potentially unfair. At present, the only real guidance on this matter comes from the so-called **'4/5 ths rule'**, which arose from experience in the USA and states that a test has the potential to be unfair if the proportion of ethnic minority applicants achieving the cut-off score is less than 4/5 ths of the proportion of majority (i.e. white) applicants 'passing' the test. However, this rule has not been tested in a UK court, and has no formal standing in this country, so that some test users (e.g. British Rail) have called for a legally recognised standard of fairness to be agreed for the UK.

Quite what form such a 'fairness' standard should take however, is by no means clear. Tests which 'fail' simple threshold standards such as the 4/5 ths rule are not **necessarily** unfair, but merely possess the **potential** to be so. Demonstrating **actual** fairness or unfairness in these situations depends whether the test user can show that the test is relevant (i.e. it assesses characteristics required for the job and at an appropriate level) and valid (e.g. it predicts job performance). However, what level of proof is adequate in practice is far from clear, as there is little or no case law in this area and professional guidance on validation studies is inconsistent.

Professional guidance differs even to the extent to which test users need to conduct validation studies in the first place. The CRE guidance (Psychometric Tests and Racial Equality, 1992) places the onus firmly on employers to conduct their own validation studies for all the tests they use, stating that *"reliance on a test supplier's assurances that their tests are fair and unbiased is not, in itself, an adequate defence"* and pointing out that *"it should not be assumed that a test that has been proved valid for one job will also be valid for another"*. But guidance issued by publishers (e.g. SHL Equal Opportunity Guidelines for Best Practice in the Use of Personnel Selection Tests) recommends only that test users should carry out a validation study *"wherever practicable"* and that in some circumstances (e.g. where a test is to be used to select relatively few people), *"evidence of validity from similar job / test combinations in other places may be relied on"*.

There is also some difference in opinion over the exact nature of the studies required to validate a test. CRE guidance is quite explicit that *"the relationship between test scores and job performance should be examined separately for each ethnic group"*, and recommends that at least 100 people are required in each group to produce statistically meaningful results. While the practical difficulties of complying with such advice are considerable, groups such as the CRE do not see them as insurmountable. For instance, larger employers who have been using tests for some time may be in a position to compile sufficiently large groups of ethnic minority employees to conduct separate validation studies. Alternatively, tests users could be encouraged to feed back results from small samples to test publishers, who could use meta-analysis to reach useful conclusions by combining the information from the different sources.

However, the need for studies of this type has been questioned by some occupational psychologists and test publishers, who argue that if a test has been shown to be valid for one group, then experience shows that it is also likely to be valid for another (provided suitable care has been taken in its administration). Such claims are based on American studies that purport to have dispelled the idea that the link between test scores and job performance might vary from one ethnic group to another (so-called **differential validity**). However, this is an area of considerable scientific debate, and it is widely recognised that there is a need for more studies (particularly in the UK) before the whole issue of differential validity can be clarified.

But even these difficulties pale into insignificance before the fundamental question of what happens when a relevant and valid test provides consistently lower scores for certain groups. This has presented professionals with serious dilemmas, particularly in the USA where extensive studies showed that black (and to a lesser extent hispanic) candidates consistently underperformed in general cognitive tests widely held to be good general indicators of future job performance. Some occupational psychologists point out these tests do provide companies with an objective basis on which to select staff, and that failure to use such tests could reduce company performance. On the other hand, using such tests increases the risk of legal challenge and also makes it difficult to provide equal employment opportunities for all groups.

In practice, many employers have moved away from any test which gives consistently different scores for different groups, irrespective of whether they believe the differences are objective or a result of test bias. For instance, the US Department of Labor suspended the use of a cognitive ability test (the 'GATB') due to a lack of scientific unanimity over its fairness. (One review by the US National Research Council concluded that scores from ethnic minority applicants should be "adjusted" to make allowance for typically lower scores, but this did not meet with the complete agreement of the wider scientific community.)

Overall, current guidance on validation appears to be inconsistent, and test users have called for clarification of what constitutes best practice in this area. Such clarification is unlikely to arise through case law, since experience to date suggests that organisations are reluctant to defend their use of potentially discriminatory tests before the Courts or Industrial Tribunals. This has led to suggestions that the way forward lies in closer collaboration between all the various interested parties - test users and publishers, academia and professional (e.g. BPS, IPD) and statutory (e.g. CRE, EOC) bodies - to develop standards of professional best practice. Indeed, the CRE have suggested that the various parties might consider forming a working group to examine issues such as differential validity, and its likely impact on school leavers, graduates, etc.

Quality Assurance and Regulation

As the number of tests on the market continues to grow, there have been increasing calls for greater regulation of their quality. At one extreme, there have been suggestions that the use of some assessment methods should be banned entirely, particularly where there is evidence to show that they lack value as predictive tools. Hand writing analysis (graphology) is a case in point, where, despite a consensus among professionals that it has zero predictive value (apart from legibility of course!), there is anecdotal evidence that it is used by a number of large employers in this country, and may be more widespread elsewhere in Europe (particularly in France, Switzerland and Germany). This raises questions whether the continued use of such methods can be seen as fair employment practice - indeed, the use of handwriting analysis for recruitment purposes is banned in some U.S. States for precisely these reasons. However, the Minister of State for Employment has made it clear that there are no such plans for the UK, stating that *"provided ...they do not discriminate unlawfully, employers should be free to use whatever selection methods they consider best suit their needs"*. (Hansard, 233, Col 271-272W, 1993).

In addition to these general issues of validity and fairness, a number of other concerns have been voiced relating to the growth in computer-based test interpretation systems. These systems can be easy to obtain (given the ease of dissemination of computer software) and are readily used by unqualified people, who merely have to arrange for answers to be entered into the computer. The software itself then generates reports which may take a variety of different forms (statistical tables, qualitative computer-generated personality descriptions, 'pen-portraits', etc). Because all the test assumptions are embedded in the computer software, there are concerns that the basis for computing the results may not be appreciated by the user. Another weakness is that computer software inevitably relies on hard and fast rules, whereas an expert would be able to rely on personal experience to tailor the reports according to the context of the test. Many conclude therefore that a computer's interpretation of a given test will often be inferior to that of a suitably qualified expert. This weakness may be compounded by the high quality 'finish' of some computer generated reports, which may imbue certain tests with an authority that is not merited on psychological grounds.

The above concerns have led some test users to suggest that professional bodies such as the BPS should develop performance criteria (e.g. of fairness) and operate some kind of accreditation scheme, so that organisations could buy tests carrying a BPS 'kitemark' or use publishers approved by the Society. The BPS however, have never operated any such scheme, and have no plans to do so in the future. Accrediting tests would be a massive undertaking, since it would involve systematically assessing the many thousands of different tests currently on the market. Rather than attempting such an approach, the BPS have instead concentrated on competency schemes (outlined in Box 3) to ensure that only suitably qualified and trained people are involved in choosing, piloting, administering and interpreting tests.

USEFUL SOURCES OF INFORMATION

ASE, 1993. 'Responsible Test Use, Guidelines for Test Publishers and Test Users', ASE, Windsor.

British Psychological Society, 1992. 'Psychological Testing: A Guide', BPS, Leicester.

British Psychological Society, 1993. 'Graphology in Personnel Assessment', BPS, Leicester.

British Psychological Society (updated regularly). 'Review of Psychometric tests for Assessment in Vocational Training'.

Commission for Racial Equality, 1990. 'Lines of Progress: An Enquiry into Selection Tests and Equal Opportunities in London Underground', CRE, London.

Commission for Racial Equality, 1992. 'Psychometric Tests and Racial Equality', CRE, London.

Commission for Racial Equality, 1993. 'Towards Fairer Selection: A Survey of Test Practice, and Thirteen Case Studies', CRE, London.

Department of Employment. 'Assessment through Psychological Testing', 1988.

Equal Opportunities Commission, 1988. 'Avoiding Sex Bias in Selection Testing: Guidance for Employers', EOC, Manchester.

Equal Opportunities Commission. 'A Short Guide to the Sex Discrimination Acts', EOC, Manchester.

Home Office. 'Racial Discrimination: A Guide to the Race Relations Act 1976', Home Office, London.

Institute of Personnel Management, 1993. 'IPM Code on Psychological Testing', IPM, London.

Saville and Holdsworth Limited, 1992. 'Equal Opportunities Guidelines for Best Practice in the Use of Personnel Selection Tests', SHL, Thames Ditton, Surrey.

Saville and Holdsworth Limited, 1992. 'Guidelines for Best Practice in the Use of Assessment and Development Centres', SHL, Thames Ditton, Surrey.

Saville and Holdsworth Limited, 1992. 'Guidelines for Testing People with Disabilities', SHL, Thames Ditton, Surrey.

The Psychologist, 1994. 'Testing in the Workplace', The Psychologist, 7, January 1994.

GLOSSARY

Biodata - a statistical scoring technique for assessing biographical data.

BPS - British Psychological Society

Cognitive tests - general term for any test assessing mental (e.g. verbal, numerical, reasoning) skills.

CRE - Commission for Racial Equality

EOC - Equal Opportunities Commission

Indirect discrimination - where the same treatment is applied to all groups but the effect is discriminatory.

Interest inventories - questionnaires that assess peoples interests and preferences.

IPD - Institute of Personnel and Development (formerly the Institute of Personnel Management)

Meta-analysis - statistical tools for analysing a number of small studies as a single, larger group.

Personality questionnaires - self assessment exercises assessing attitudes, beliefs, feelings and behaviour.

Psychometric (or psychological) tests - 'an instrument designed to produce a quantittative assessment of some psychological attribute or attributes' (BPS).

Validity - the accumulated evidence on predictivity, content, fairness, etc. which justifies the use of a test in a given application.

Work samples (job simulations) - tests designed to replicate key functions of a job.