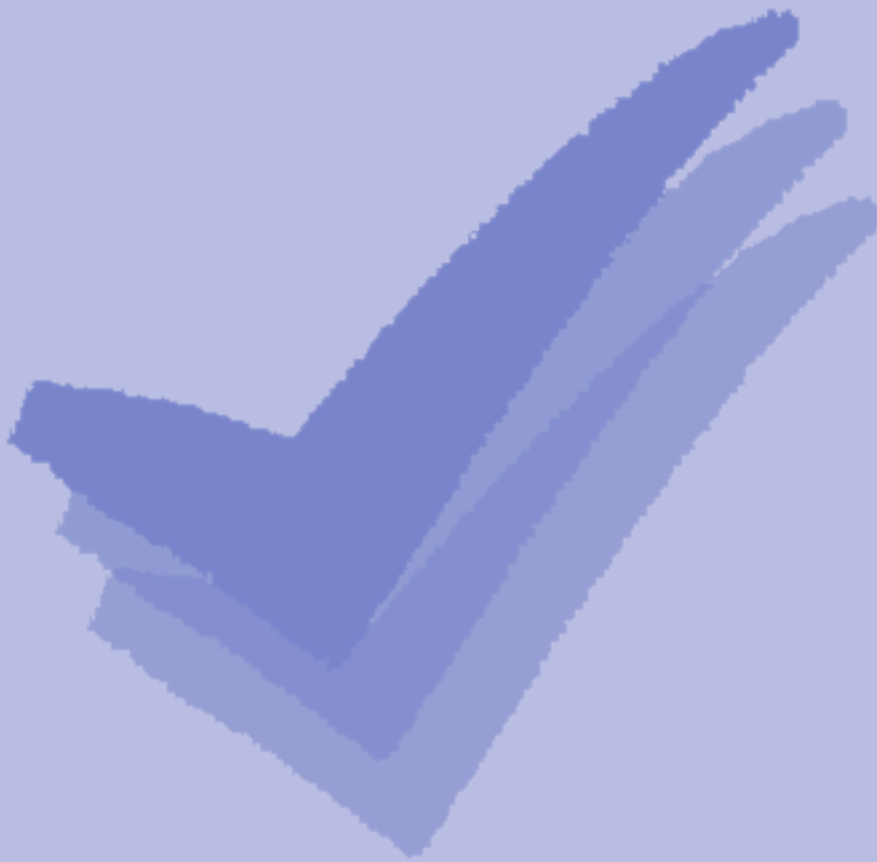


Equal Opportunities Guidelines

for Best Test Practice in the use of Personnel Selection Tests



SFTL

**These guidelines were produced in conjunction with the
Commission for Racial Equality and the Equal Opportunities Commission.**

Equal Opportunities Guidelines

for Best Test Practice in the use of Personnel Selection Tests

Contents	Page No
1. CHOICE OF TESTS	2
1.1 Psychometric qualities	2
1.2 Test content	3
1.3 Test level	3
1.4 Customisation	4
2. VALIDATION - TESTING THE TESTS	5
3. PREPARATION OF CANDIDATES AND ADMINISTRATION OF TESTS	6
3.1 Use of practice materials	6
3.2 Administration of tests	6
4. INTERPRETING SCORES	7
4.1 Choice of norm groups	7
4.2 Selection strategies	7
4.3 Group differences	8
4.4 Monitoring	9
5. USE OF PERSONALITY QUESTIONNAIRES	10
6. A FINAL WORD	10
Glossary	11
Notes	12
Useful Publications	12



© SHL Group plc, 2000

Email: uk@shlgroup.com • Internet: <http://www.shlgroup.com/uk>

The reproduction of these guidelines by duplicating machine, photocopying process or any other method, including computer installations, is breaking the copyright law.

Whilst SHL has used every effort to ensure that these guidelines reflect best practice, SHL does not accept liability for any loss of whatsoever nature suffered by any person or entity as a result of placing reliance on these guidelines. Users who have concerns are urged to seek professional advice before implementing tests.

Equal Opportunities Guidelines for Best Practice in the Use of Personnel Selection Tests

Research has shown that well-constructed psychometric tests predict job performance better than almost any other single selection measure. Relevant tests produce more accurate results than common selection measures, such as interviews and references¹. Tests give objective information about a candidate and have been shown in general to lead to better and fairer employment decisions. However, tests can sometimes have *disparate impact* on ethnic or gender groups, and it is therefore particularly important that improper usage is avoided. The following notes provide guidelines to ensure that tests are used appropriately.

1. Choice of tests

Choice of tests should follow from a careful *job analysis*. The choice of tests to be used, or the construction of new ones, should be based on the results of the analysis, which will identify the competencies, abilities and attributes required to perform the job.

1.1 Psychometric qualities

Tests should be psychometrically sound. A full description of what this entails is beyond these guidelines but may be found in textbooks on psychometrics.² The relevant information and statistics for judging the tests should appear in the test manual. These should include:

- Specification of the skill the test measures
- Description of groups for which the test is appropriate (educational background, work experience etc.)
- Details of the development process
- Test *reliability* statistics
- Evidence of test *validity*
- A report of steps taken to guard against bias (e.g. gender, ethnic, age) in the test
- A selection of relevant *norm* groups containing representative proportions of ethnic minorities, men and women.

Do not judge a test solely by how widely it is used; a test can become outdated, and companies sometimes use inappropriate or even bad tests.

1.2 Test content

Two aspects of test content should be considered. The most important is the actual skill or attribute being measured by the test, e.g. verbal comprehension, manual dexterity. All the tests used should measure skills or attributes identified as necessary to do the job. Skills not required on the job or in training should not be a necessary requirement in order to complete a particular test. For instance, the test should not require understanding of complex vocabulary or performance at speed, unless these are relevant to the job. Under the law, the employer may be required to show that tests used correspond to a real need, are appropriate with a view to achieving the objective and are necessary to that end.

The second aspect of test content is the context in which the skill is measured. This should, as far as possible, reflect the type of content found in the job. For example, a typing test should require the typing of material similar to that required on the job. However, care must be taken not to include material requiring knowledge specific to the organisation, that would put external applicants at an unfair disadvantage. Test content that is of a more general nature should be equally accessible to all applicant groups - men and women, ethnic minorities. For instance, in the typing test, if job relevant text is too technical for an external applicant to deal with before training, more general text should be used. This should be of content matter equally familiar to all groups.

1.3 Test level

The level of difficulty at which the skill is measured should be appropriate to the job. A test which is too easy will not differentiate between applicants with good and poor potential. One that is too difficult could lead to greater disparate impact. The level of the test should also be appropriate for the likely applicant pool. If the general level of applicants is below the standard required for the job, employers should consider what they can do to attract better applicants. Training or job redesign options also need to be considered.

Where there is a tendency for applicants from one particular racial group or sex to fail to meet the required standard, Section 38 of the Race Relations Act (1976) and Section 47 of the Sex Discrimination Act (1986) sometimes allow positive action targeted at this group, in the form of training programs, to be instituted.

1.4 Customisation

A *customised* test is one which is constructed especially for a particular user. Customisation enables the production of a test measuring a skill and in a context which are both tailored to organisational needs. Customised tests have the following advantages and disadvantages:

Advantages

- ✓ The design of the tests is likely to be well suited to the intended use. The skill tested, as well as the level and content of the test, have the best relationship to the job analysis. *Validity* is therefore enhanced.
- ✓ In designing the tests, account can be taken of the probable knowledge and background of likely applicants where this information is available.
- ✓ The tests are likely to seem most fair to candidates, in that they reflect the content of the job.
- ✓ The security of the test material is under the control of the commissioning organisation. Candidates will not have seen the tests before and therefore cannot benefit from these specific practice effects.
- ✓ Although the initial outlay is greater, where large numbers of people are tested, customised tests can be more economical in the longer term than off-the-shelf tests.

Disadvantages

- ✗ Customised tests require a minimum of six months to develop before being available for use.
- ✗ They require relevant trial groups of several hundreds to be available.
- ✗ Good test publishers ensure that the off-the-shelf tests they distribute are monitored and updated. The organisation would be responsible for commissioning this work with customised tests.
- ✗ Should the job, or applicant population, for which the tests were designed change, the tests may become obsolete.
- ✗ External *norm* groups will not be available so no comparisons can be made with other groups. Internal norms have to be developed.

2. Validation - Testing the Tests

Test performance must be related to job performance for the tests to be effective. Wherever practicable, test users should, by carrying out a *validation* study, establish that the tests *correlate* with job performance or other relevant criteria such as labour turnover; such a study shows whether higher scorers on the test tend to be more successful on the job. Validation is imperative if a test is to be used for purposes other than those for which it was designed.

The research evidence (mainly from the USA) shows that, where a type of test has been shown to be valid for a particular job, it is likely to be valid for all similar jobs.³ This finding of *validity generalisation* means that where a test is to be used to select small numbers of people (say less than around 50), evidence of *validity* from similar job/test combinations in other places may be relied on. When larger numbers of employees are to be selected, specific validation studies can be carried out. The performance of a robust validity study requires in-depth knowledge of practical issues and statistical procedures, and expert advice and/or assistance should be sought.

Ideally, validation studies should occur before tests are introduced as selection measures. There are two basic study designs. In a *concurrent* study, job incumbents are tested and their performance evaluated at the same time. The relation between current performance on the test and in the job can then be examined. The advantage of this type of study is that it is quick, but the disadvantage is that experience on the job sometimes affects test scores, making the results of the study less reliable. For example, on a test of numerical computation, people with substantial recent experience would be likely to perform better than those with no such experience.

A *predictive* study avoids this problem by testing applicants before they start work. However, some time must then elapse before performance ratings can be made. For this reason a predictive study will not yield results as quickly as a concurrent study. Ideally, the test scores should not be used in making selection decisions until the validation is complete. In practice, however, the scores can be used conservatively while awaiting results, where not using them is likely to be detrimental to selection procedures.

In both these cases, the test results should be monitored for any differences in scores by ethnic group, age or sex. Even where a test is not validated before use, administration to a pilot sample should be undertaken if a test is to be used with large numbers of applicants. Test manuals and norms should also be consulted to see whether group differences are common and what has been done to ensure the test is fair. Although differences do not necessarily mean the test is biased, where they are found further investigation should follow. (See section 4.4)

3. Preparation of Candidates and Administration of Tests

3.1 Use of practice materials

Some candidates may be unfamiliar with testing so that it is difficult for them to perform at their best. Others may find the testing situation very stressful. Ethnic minority candidates in particular might perhaps under perform, because of the effects of educational disadvantage or race discrimination. Older candidates and those with less educational experience are also likely to suffer these sorts of problems.

Practice items at the beginning of a test can reduce the bias that may arise from differential 'test sophistication', helping some people but not others. They can also reduce nervousness by allowing a candidate to gain confidence in his/her ability to answer the test questions.

If possible, candidates should be notified in advance that they will be tested. Examples of what the tests will be like (practice tests) should be provided, so that candidates can familiarise themselves with the type of tasks involved. Such practice tests increase the effectiveness of the main tests in giving an accurate assessment of a candidate's abilities. In addition, they allow candidates with disabilities to check whether they will need any adaptation to standard test administration procedures. See the SHL *Guidelines for Testing People with Disabilities* for more details on using tests with disabled candidates.

3.2 Administration of tests

The administration instructions are extremely important and must always be strictly followed. Only qualified persons should administer tests. Abuse of procedures described in the test manual can lead to bias and possible unlawful discrimination.

Special care should be taken with people whose first language is not English, to ensure that they have understood the administration instructions properly. Some tests which are fair for native English speakers will present problems for people with a lesser command of English. Tests requiring reading skills, when these are not an integral part of the job, are particularly likely to be unfair. Where possible, such candidates should be tested in their native language and given an additional test of their command of English, if necessary.

An encouraging attitude on the part of the test administrator is always desirable, but it is particularly important to establish rapport with individuals who might lack confidence or who feel anxious about testing. The introduction to the testing session is an important part of the administration procedure. It allows the establishment of this rapport and should be conducted in a friendly manner. Information should be provided on why the tests are being used and how they fit into the assessment procedure. There should be an opportunity for candidates to ask general questions before the formal testing procedure starts.

4. Interpreting Scores

Only properly trained persons should be responsible for interpreting test scores.

4.1 Choice of norm groups

An individual's test score is generally evaluated in relation to the performance of a comparison or '*norm*' group. The norm group should be as representative of the applicant group as possible. In large-scale testing, the previous applicant pool can form a suitable comparison group. Where numbers are too small, an appropriate norm group should be chosen from the test manual. This should be, as far as possible, a group applying for a similar type of position with a comparable composition with respect to educational background, age, ethnic and gender mix. Where a perfect fit is not available, the norm group chosen should have similar average scores to those of the applicant group as a whole.

4.2 Selection strategies

There are two basic approaches to selection on the basis of the test scores of a group of applicants. The '*top-down*' approach selects people from the highest scorer downwards until sufficient people have been selected. This approach maximises the benefit that can be gained from using the test, its '*utility*', and is likely to produce the highest calibre work force. Where the *selection ratio* is low, that is where there are many applicants for few jobs, this method quickly reduces the number of candidates to be considered. However, it tends to increase *disparate impact* where there are score differences between groups, that is this method selects fewer applicants from the lower scoring group.

The '*cut-off*' approach selects people who score above a designated cut-off. The cut-off score is generally chosen to represent the ability level consistent with a reasonable chance of successful performance on the job; it clearly depends both on the level of the test being used and the skill requirements of the job. This method is particularly appropriate where there is a high selection ratio (few applicants for many jobs) and the test is used to screen out applicants who are not up to standard. The lower the cut-off is set, the lower the utility of the test for the organisation, but the less disparate impact the test will have if there are score differences between ethnic or gender groups.

The choice between these approaches and the actual cut-off score used, depends, in each individual case, on the selection ratio, the level of the job and/or training, the calibre of the applicant pool, and whether there is any question of disparate impact. The greater the demands of the job and the less training available, the higher cut-off scores would need to be. Where less skills are required and more training is available, or where group score differences have been observed, lower cut-off scores are more appropriate.

Often a more qualitative strategy is used, whereby the individual's relative performance on the test compared to the norm is evaluated together with other information. Measures such as personality questionnaires, practical exercises and interviews, as well as details about previous work experience and biodata, should provide additional information about a candidate's strengths and weaknesses.

4.3 Group differences

A problem arises when a significant difference is found between the average test performance of different ethnic groups or men and women. In the absence of *validation* evidence, there is likely to be a presumption that the group with the lower average performance was being indirectly discriminated against. That is, if the same entry standard were demanded of all applicants, the lower scoring group would find it harder to comply with the requirement.

Positive validation evidence for the test generally justifies the use of the test and rules out the possibility of unfair discrimination. By showing that those who perform poorly on the test also perform poorly on the job, a positive validation result confirms that rejecting low scorers is reasonable. The greater the degree of *disparate impact* resulting from the use of a test, the higher the validity of the test should be to justify its use.

There remains the possibility that overall validity is masking cases where a test has poorer or no predictive validity for some groups, or that group differences in test scores are not reflected in job performance. Much research into these issues in the United States, covering many types of tests and a wide range of occupational fields, has indicated that such scenarios are extremely rare if they exist at all - when good test practice had been followed ^{4,5}. There is a lack of published studies in this area in the UK and more work still needs to be undertaken.

Some experts argue that, where ethnic or gender group differences on test scores exceed those in job performance, separate *norm* tables for each group should be used for evaluating scores ⁶. Use of separate norms in these circumstances has not been tested in British courts, but it certainly would not be justifiable in any other circumstances. In all cases, the availability of direct or relevant (from similar tests and jobs) validation data means that discriminatory practices can be avoided.

It should be remembered that group differences relate to average performance. Even where there are substantial group differences, there will be members of the lower scoring group who have higher scores than many people from the higher scoring group and vice versa. Furthermore, job success does not generally depend on a single ability and tests do not have perfect predictive power. Therefore, on occasion, low scorers on a test will do better on a job than a test score may suggest. For this reason, it is preferable to interpret test scores together with other available information.

4.4 Monitoring

Whether there is a possibility that ethnic or gender group differences exist or not, test use should be continually monitored. In small scale applications, this will amount to ensuring the test remains relevant to the job, following other validity studies on the test, and using updated versions and *norms* as they become available.

Where larger scale use occurs, scores should be monitored at regular intervals in order to update norms. Monitoring by ethnic group and sex is required to look at any changes in the relative scores of different groups. A *validity* study should be carried out every few years or whenever changes in the job or applicant group are such that initial validity could have been affected.

Where substantial *disparate impact* has been found, the issues listed below should be considered. Many are also relevant when tests have poor validity.

- Has the job changed in some way since the original job analysis?
- Are the skills measured really relevant to the job?
- Is the way the skills are measured (e.g. language used, speed) appropriate to all candidates?
- Would customisation of the context in which skills are measured help?
- Are the tests at the right level for the job?
- Are the tests at the right level for the applicant population?
- Are the tests predicting job performance for all groups?
- Is there some other test or type of measure (e.g. trainability tests, *work samples*) which would provide the same information with better validity/without disparate impact?
- Can the selection rule be designed to minimise disparate impact, despite score differences?
- Can the job or training be redesigned so that the entry level required for the relevant skill or ability is lower?
- Can disadvantaged groups be trained to offset differences in test taking behaviour or to enhance relevant skills?

5. Use of Personality Questionnaires

Many of the above guidelines are also pertinent to personality questionnaires. These types of questionnaires differ from ability tests in several ways. There are no right or wrong answers to the questions they contain; each person answers what is correct for him or herself. Each questionnaire generally measures a number of scales or attributes rather than just one. There is no perfect score on a personality questionnaire. In one situation, a high score on a particular scale may be desirable and in another a low score. This makes the interpretation of personality measures more complex than that of ability measures.

Similar psychometric standards apply to personality questionnaires as apply to ability tests. Initial research has shown that there are relatively small differences in the average response patterns of different ethnic groups. Gender differences, where they occur, tend to reflect known differences in style between men and women.

A particular job may require a certain personality style just as it needs a specific cognitive ability. It is often possible to use the information from personality questionnaires in a qualitative manner to help build up a general picture about the suitability of an applicant. The job requirements should be investigated through job analysis and validation studies as described for ability tests before using scores in a mechanistic way, e.g. rigid cut-offs on single scale scores or combinations of scores should not be used without adequate validation data.

6. A Final Word

While there is strong evidence that tests generally provide sound, objective data on which to base selection decisions, it is important to investigate the relevance of a test before including it in a selection procedure. Test results should always be interpreted within context and should never be used on their own. It is likely that they only measure some of the relevant attributes and other selection methods need to be used to assess the other characteristics. There may well be greater costs involved in ensuring that a selection process is fair, but these should be far outweighed by the benefits of a capable and representative workforce. Finally, these guidelines provide a useful structure for evaluating the impact of any assessment or selection method. Interviews, biodata etc. should all meet the same rigorous standards of fairness and relevance.

Glossary

Correlation	The extent to which two measures vary together. For example, a positive correlation occurs when people who score high on one measure tend to score high on the other.
Customised Test	A test designed for the sole use of a particular organisation. This allows the test items to be framed in the context of the work of that organisation.
Cut-off	The score on a test which separates those selected from those rejected (the 'pass' mark).
Disparate Impact	A selection criterion has disparate (or adverse) impact when proportionately fewer members of one ethnic or gender group can meet the criterion.
Direct Discrimination	Treating a person less favourably because of their ethnic or gender group.
Indirect Discrimination	An unjustifiable requirement or condition which has disparate impact on people from one ethnic or gender group, to their detriment.
Job Analysis	A structured examination of the tasks inherent in a job and the skills required to perform them.
Norms	Standard distributions of test scores based on the performance of a representative sample of a given group. Psychometric test scores are interpreted through comparison with relevant norm groups.
Psychometric Test	A measure of a psychological construct which produces scores which are reliable and valid.
Reliability	The extent to which a measure produces consistent scores, e.g. when a person is tested twice or when two people of the same ability are tested.
Selection Ratio	A ratio of the number of people selected to the total applicant pool size. A high ratio occurs when many people are selected, a low one when few are selected.
Top-down Selection	A selection strategy in which the (next) person chosen is the one with the highest available test score.
Utility	The average gain (per employee, per year of employment) from selecting applicants using a particular measure over applicants selected without that measure.

Validity	The extent to which an instrument measures what it is designed to measure.
- Concurrent Validity	The ability of a measure to predict present performance levels.
- Predictive Validity	The ability of a measure to predict future performance levels.
Validity Generalisation	Recent work has shown that if a test is valid for a particular job, then similar tests will be valid for similar jobs. That is, validity is generalisable from one situation to another.
Work-sample Test	A test in which one or more practical tasks drawn from the job itself are performed.

Notes

1. Smith, M & Robertson, I T. **The Theory and Practice of Systematic Staff Selection.** Macmillan, London, 1986.
2. Toplis, J., Dulewicz, Vic, Fletcher, Clive. **Psychological Testing: A Manager's Guide** 1997 Chartered Institute of Personnel and Development, Wimbledon.
3. Schmidt, Frank L.; Hunter, John E. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. **Psychological Bulletin.** 1998 Sep Vol 124(2) 262-274
4. Hartigan, J.A. & Wigdor, A.K. **Fairness in Employment Testing: Validity Generalisation, Minority Issues, and the General Aptitude Tests Battery.** National Academy Press, Washington D.C., 1989.
5. Hough, Leaetta M.; Oswald, Frederick L. Personnel selection: Looking Toward the Future-Remembering the Past. **Annual Review of Psychology.** 2000 Vol 51 631-664.
6. Borman, Walter C.; Hanson, Mary Ann; Hedge, Jerry W. Personnel Selection. **Annual Review of Psychology.** 1997 Vol 48 299-337.

Useful Publications

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

Guidelines for Testing People with Disabilities. SHL (UK) Ltd. Thames Ditton, 2000.

The IPM Guide on Psychological Testing. Chartered Institute of Personnel Management, Wimbledon, 1997.

Psychological Testing: Guidance for the User. Steering Committee on Test Standards. The British Psychological Society, Leicester, 1989.

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

Selection Tests and Sex Bias. Pearn M.A., Kandola R.S. and Mottram R.D. HMSO, London, 1987.

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

THE CLIENT SUPPORT CENTRE
Telephone: 0870 070 8000
Facsimile: 0870 070 7000
Email: uk@shlgroup.com
Internet: <http://www.shlgroup.com/uk>

HEAD OFFICE
SHL
The Pavilion, 1 Atwell Place
Thames Ditton
Surrey KT7 0NE
Tel: 020 8335 8000

MANCHESTER OFFICE
SHL
Fairbank House
27-29 Ashley Road, Altrincham
Cheshire WA14 2DP
Tel: 0161 929 8299

SCOTLAND OFFICE
SHL
51 Timber Bush
Leith
Edinburgh EH6 6QH
Tel: 0870 070 9000

LONDON ASSESSMENT CENTRE
SHL
3rd Floor
2 Caxton Street
London SW1H 0QE
Tel: 020 7799 3464

BELFAST OFFICE
SHL
4 Malone Road
Belfast BT9 5BN
Tel: 028 9066 1616

MIDLANDS OFFICE
SHL
Arden House
Stratford Court
Cranmore Boulevard
Solihull B90 4QT
Tel: 0870 070 6000

