Aptitude Testing and the Legal Profession

6 June 2011

Dr. Chris Dewberry

Birkbeck, University of London

Acknowledgement: I wish to thank Chris Hutton for his help in compiling information in the section of this report entitled *Aptitude Tests Currently Used in the Professional Services Sector*, Appendix 3, and Appendix 4.

Contents

Executive Summary

Traditional methods for selecting people who wish to be educated and trained to work in the legal profession are increasingly difficult to implement successfully. These traditional selection methods have to a considerable degree focussed on educational qualifications, including GCSE and A level results for entrance to first degree law courses, and first degree results for entrance to vocational training programmes such as the Bar Professional Training Course. Difficulties in the continued use of these methods include the substantial increase in the proportion of GCSE and A level students awarded the highest grades in recent years, making attempts to discriminate between candidates problematic; the increase in students with an overseas education applying for entry to legal training courses, for whom it is often difficult to reliably compare the quality of their educational qualifications with those obtained by students in the UK; and the concern of the government and other agencies with social mobility, and with this the importance of opening careers in the professions to the most able and suitable candidates irrespective of the social and educational advantages or disadvantages they have experienced in the past.

As a consequence, interest in the use of aptitude tests for selecting people who wish to be educated and trained in the legal and other professions has gained momentum in recent years. Such tests may be perceived to offer several advantages over traditional selection methods. These include the assumptions that aptitude tests are objective, fair, and provide a powerful way to identify candidates with the greatest potential to succeed in their chosen profession irrespective of their social and educational background or geographical location. In this report these assumptions and other issues relating to aptitudes and aptitude testing are examined, and a set of criteria for evaluating the effectiveness of such tests in the legal profession are proposed.

Aptitude tests are one example of a more general category of assessment methods referred to as psychometric tests. Psychometric tests, which also include tests of personality, interests, motivation and others, are systematic and standardised methods for assessing the psychological and behavioural characteristics of people. The term “aptitude” is not always used in the same way. It can refer to different types of cognitive ability (e.g. verbal and numerical ability), may be extended to include areas outside of cognitive ability (e.g. aspects of personality or physical coordination), and is sometimes used interchangeably with the concept of ability. Here the word aptitude will be used to refer to *the extent to which an individual has* *the psychological and behavioural characteristics necessary to perform at a high level in a particular environment (including task, job, training or educational programme) in the long term*. Test of aptitude are usually composed of two types of assessment. One is concerned with measuring two or more areas of intellectual ability (e.g. spatial ability and numerical ability). The other focuses on areas of attainment considered relevant to the job or training that a candidate is being considered for. Examples of sub-tests of attainment are knowledge of the physical sciences, knowledge of spelling, and motor-coordination.

The concept of aptitudes and the procedure of aptitude testing are based on critical assumptions about the structure of human ability. From the outset, the structure of human intellectual ability has been one of the most researched topics in the science of psychometrics. The fundamental question here is whether there is a general construct of cognitive ability, with some people more able than others in a wide range of intellectually demanding areas, or, conversely, whether people differ with respect to specific and relatively independent types of cognitive ability (e.g. verbal and numerical). This issue has very considerable practical as well as theoretical importance. Put simply, if people differ broadly in terms of the degree to which they are intellectually able in many areas, a notion captured by the term “general cognitive ability” and often abbreviated to *g,* there is little point in developing tests of intellectual aptitude to match the ability profiles of people to specific jobs, roles, or training programmes. However, if verbal, spatial, numerical, and other types of intellectual ability are independent of each other, a clear rationale is available for the development of aptitude tests to measure these particular dimensions and match them to the needs of different jobs, roles, and training programmes.

From the 1930’s, for a period of about 60 years, psychologists tended to assume that several quite independent areas of cognitive ability could be measured. This assumption underpinned the great growth in aptitude testing during this period. However, during the last 20 years psychologists have reviewed and analysed very large amounts of data collected over several decades, and the consensus is that there is little independence between different types of cognitive ability. One implication of this is that aptitude tests primarily measure *g,* and a second is that attempts to match intellectual profiles to job profiles (or the profiles required for types of training, education etc.) in an attempt to predict future performance is ultimately misguided. This position is reinforced by empirical research showing that if performance on a job or training programme is first predicted with a measure of peoples’ general cognitive ability, and then it is predicted again with measures of their general cognitive ability *plus* measures of their specific intellectual aptitudes, there is little or no increase in the accuracy with which the job performance or training performance of people can be predicted.

Although the practice of trying to discriminate between individual components of intellectual ability with aptitude tests is undermined by the finding that one general construct of cognitive ability accounts for much of the variation in more specific cognitive abilities, there is a wealth of evidence showing that *g* has a considerable impact on many areas of life, not only with respect to educational and work performance and variables associated with this such as adult income, but also with many other areas such as the probability of divorce, having an illegitimate child, and being on welfare benefits. Indeed a great deal of research shows that amongst the various methods available to select job candidates (e.g. interviews, assessment centres, work samples etc.) there is no better predictor of job performance than tests of general cognitive ability.

Considerable research indicates that measured *g* in adults is influenced by both genetic and environmental factors. Environmental factors include family background and education. As a consequence, an individual's measured *g* reflects not only his or her genetically inherited potential to perform well cognitively, but also the result of many environmental influences, including the nature of the individual's family and educational background.

Research carried out in North America shows that there are substantial sub-group differences in measured *g.* For example, Americans with a Black African ethnic background obtain scores which are on average one standard deviation below those of their white counterparts. This has very significant implications for the use of cognitive ability and aptitude tests in selection, with selection systems relying heavily on cognitive test results being less likely to admit Black Africans and Hispanics than Whites, and this has almost certainly inhibited the use of such tests in many settings. Furthermore, research also shows that coaching and practice can have a marked effect on peoples' results in cognitive ability tests, and therefore unless all those taking the tests either have no practice or coaching, or the same amount of practice and coaching, unfairness and bias will be present in the assessment of cognitive ability.

Despite these problems, aptitude tests (which for reasons explained above largely measure general cognitive ability) are widely used in North America, with tests such as the SAT, MCAT, LSAT, ACT and GATB used on a very large scale, particularly for the selection of people to universities at undergraduate and postgraduate level. Continuing research and development has been carried out for many years on some of these tests, and this shows that some, such as the SAT, predict outcomes such as first-year undergraduate degree examination results relatively well, and also provide incremental validity in predicting these results over and above the grade point average obtained by US students at school.

At present little is known about the incidence and nature of aptitude testing in the UK, which it is carried out almost exclusively by private test administrators and publishers. In the preparation of this report about half of the UK tests publishers were contacted and none were able and willing to disclose information about the results of validity studies carried out on such tests when they are used to select professionals, most citing client confidentiality and commercial sensitivity as the reasons for this. In the selection for vocational education in the UK, tests of aptitude are used for the selection of undergraduate medical (UKCAT and BMAT) and law (LNAT) students. There is little published research on the criterion-related and incremental validity of these tests, and at present the extent to which they predict short or long-term performance in education, professional training, and subsequent job performance is unclear.

In the development and validation of a test of aptitude to select people to be educated, trained, or to work in the legal profession the following recommendations are made:

1. The purpose of the test should be clarified. For example, is the test intended to predict performance in the long or the short term; in an initial training course or legal career, or both?
2. If the test is not simply designed to measure *g*, the evidence that it has sufficient content validity should be investigated.
3. Test developers should consider a range of different techniques, including situational judgement tests and personality questionnaires.
4. If a test is designed to measure specific psychological constructs, there should be evidence that it has acceptable construct validity.
5. The internal and test-retest reliability coefficients of the test should be established, with information about the nature of the people examined, the

situations in which data was collected from them, sample sizes, and in the case of test-retest reliability the 95% confidence interval for the estimated coefficient.

6. Careful consideration should be given to the criterion to be used in establishing the criterion-related validity of the test, bearing in mind the purpose of the test (see point 1 above).

7. One or more criterion-related validity estimates should be reported. For each such estimate, the size and nature of the sample used to estimate it should be reported together with the 95% confidence interval for all such validity estimates.

8. The incremental validity of the test over and above alternative information available on candidates from other potentially predictive variables such as GCSE, A level, and undergraduate degree results should be reported. This validity information should not focus solely on the issue of whether or not the incremental validity is statistically significant, but also on effect size.

9. The extent to which systematic sub-group differences on the test exist in relation to social class, educational background, gender, and ethnicity should be investigated. The mean and standard deviation of the test scores in each group should be reported as well as *d* scores. The extent to which the predictive validity of the test (and sub-components of the test) is relatively equal in relation to all sub-groups should be examined, as should the relationship between the size of sub-group differences on the test and the size of sub-group differences with respect to other predictor variables (such as GCSE, A level, and undergraduate degree results). The consequences of including and excluding the aptitude test results on sub-group selection ratios should be reported. If sub-group differences are apparent, the use of differential item functioning (DIF) to further develop the test in order to reduce these effects should be considered.

10. All candidates should be given access to a sufficient number and range of practice tests, and ideally test coaching opportunities also. These practice tests should be sufficient in relation to availability, length, clarity, and quality. All candidates should be aware of these practice tests and coaching opportunities and all should be able to make use of them.

11. The method used to combine test results with other information about the candidates in order to arrive at selection decisions should be carefully considered, including the advantages of combining information actuarially rather than clinically.

12. Annual reports on the reliability, validity, and sub-group differences of the test should be published. This information should be used to develop and improve the test.

The Background to this Report

Providing high quality processes and systems to select people who wish to train and work in the legal profession is clearly very important. For many years programmes offering education to undergraduate law students, and education and training to people who wish to work as barristers, solicitors and in other legal professions, have, when selecting candidates, placed considerable emphasis on academic qualifications such as GCSE, A level, and undergraduate degree results. This approach is increasingly problematic. In relation to selection for undergraduate courses in law, the proportion of candidates achieving three A's at A level has been steadily increasing (prompting the introduction of the A* grade in 2010). The larger proportion of law school candidates achieving the highest possible grades at A level makes it more difficult to discriminate between them effectively for selection purposes, particularly for the most popular courses at the most selective universities. Selecting the best candidates for postgraduate programmes such as the Bar Professional Training Course has become more difficult also. Here there has been a substantial increase in the number of applications from people educated and examined overseas. In many cases these applicants possess undergraduate degree qualifications which cannot easily be compared to UK first degree standards, and this presents a difficulty for those using undergraduate degree qualifications as a basis for candidate selection. In addition, the current government's concern with social mobility (Crawford, Johnson, Machin, & Vignoles, 2011; H.M.Government, April 2011), the disproportionate number of students from independent schools studying at the most selective universities, and the tendency of students from independent schools to underperform at university compared to state school children with the same educational qualifications (Kirkup, Wheater, Morrison, Durbin, & Pomati, 2010), have further complicated the process of fairly selecting the most suitable applicants for education and training in the legal professions.

These challenges have raised awareness of the possible benefits of introducing novel selection techniques, such as aptitude testing, in the legal and other professions. In the United States, aptitude tests are widely used for the selection of undergraduate and postgraduate students, and more specialized aptitude tests are used for the selection of medical and law students. As will be discussed later in this report, these tests have been developed and refined over many years, and regular research and publications on their validity generally indicate that they make a useful contribution to the prediction of success in, for example, the examinations that students and trainees take in the first year of study. Because the content of aptitude tests generally reflects the types of cognitive and other demands required in the profession for which they have been designed, they would appear, in the context of selection, to offer considerable advantages over almost all educational qualifications acquired in school and university where the content examined may bear little relation to these specific demands.

The principal aims of this report are to provide a critical introduction to aptitudes and aptitude testing, and to set out criteria for evaluating the effectiveness of an aptitude test. In so doing the report will describe the historical background to aptitudes and aptitude testing, discuss the nature of aptitude tests, outline the ways in which such tests can be evaluated, consider the current use of aptitude tests for selection in the professional services in the UK and elsewhere, discuss the outcomes of research on the reliability and validity of aptitude tests, and make recommendations for those who may develop aptitude tests for the UK legal profession and for those who may wish to evaluate the validity and usefulness of such tests.

An Introduction to Aptitude and Aptitude Tests

Psychological testing broadly refers to the systematic and standardised assessment and measurement of the psychological characteristics of people responsible for, or associated with, their mental life, behaviour, and achievements. The primary focus of this report is on aptitudes and their measurement, and this represents one approach to psychological testing. A challenge encountered by anyone writing about aptitudes is that the term does not have a single meaning: the terms “aptitude” and “aptitude test” are used in a variety of ways not only by lay people but also by psychologists. For example, someone might be said to have an aptitude for administration, and the person making this claim may or may not imply that the person has an *innate* ability to perform at a high standard in this field. Alternatively, aptitude may refer to an area such as verbal ability – someone might have “an aptitude for words and language”. Whereas administration refers to a particular form of work, verbal ability refers to a particular type of cognitive ability, and this is therefore a quite different use of the term aptitude. This use of aptitude to refer not only to primary cognitive abilities, but also to what it takes to perform well in a particular job or role, occurs not only in the lay use of aptitude but also in psychological measures of aptitude. For example, the well-known Differential Aptitude Test (Bennett, Seashore, & Wesman, 1962) seeks not only to measure primary cognitive abilities (e.g. "numerical reasoning") but also ability relevant to specific classes of job (e.g. "clerical speed"). To add even greater confusion, aptitude is sometimes used interchangeably with the word ability. Here it is being used not to refer to a particular area of the capacity to perform well, but to the capacity to perform well in all intellectually demanding areas.

Given the loose way in which aptitude and aptitude tests are used, it is essential to begin this report with a clear definition of these terms. Here the word aptitude will be used to refer to *the extent to which an individual has the psychological and behavioural characteristics necessary to perform at a high level in a particular environment (including task, job, training or educational programme) in the long term*. Construed in this way, an aptitude is not a feature of a person, but rather expresses the relationship between a person's characteristics and the demands of a specific environment. So someone with the aptitude to be a barrister has the psychological and behavioural characteristics, and therefore the potential, to perform well in this role.

Aptitude tests are typically comprised of several different sub-tests, and these are of two different types (this will be discussed in more detail in the section on "*The Nature of Aptitude Tests and the Influences on Aptitude Test Results"* which begins on page 38). One type of test measures various aspects of general cognitive ability, and the other measures domain-specific attainments. The tests of general cognitive ability usually examine specific content areas such "verbal ability", "numerical ability", and "spatial ability". Tests of attainment on the other hand measure someone's performance in a relatively specific area or domain, and examples are tests of general scientific knowledge, spelling, and manual dexterity. Tests of attainment are similar to tests of achievement. However, whereas tests of achievement are concerned with evaluating performance in relation to some formal programme of education or training, and focus on the content of this education or training (e.g. A level examinations), tests of attainment measure the standard attained by someone in an area which, whilst quite possibly influenced by formal learning and training, is also a product of more general, everyday, and informal learning.

An operational aptitude test has normally been through two or possibly three stages of development. In the first stage the psychological and behavioural characteristics required for high performance in a given performance arena (e.g. medical education) are investigated. For example, if an aptitude test for a particular job is required, researchers may carry out a formal job analysis using methods such as hierarchical task analysis and the critical incident technique. Once these psychological and behavioural characteristics have been specified, an aptitude test is purposely designed to measure them, or a suitable aptitude test already in existence is taken “off the shelf”. In the third stage, the validity and reliability of the test is examined, and if in the light of this modifications and improvements are considered necessary, these are implemented and validity and reliability are assessed again. This process is repeated until the test designer considers the test ready for use. In some cases the process of test examination and modification is repeated after the test is operational, and the test may go through many stages of evolution and refinement over several decades.

Aptitude and cognitive ability are often associated with the concept of intelligence. Because intelligence is, at least among lay people, often assumed to be innate, tests claiming to measure ability, aptitude, or intelligence can be misinterpreted as measuring the purely innate or entirely genetically inherited ability to perform well. In fact no psychological test can claim to measure purely innate abilities because all require the person completing them to draw on things they have learned, and the sum total of the things that people have learned, at the moment they do the test, depends not only on genetically inherited ability but also on the pattern of environments and the multitude of experiences they have lived through. For this reason, wherever possible, reference to the term intelligence has been avoided in this report, and the term general cognitive ability is used instead. General cognitive ability, which can be and is assessed with psychological tests, refers to an individual’s ability to perform well at a broad range of cognitively demanding tasks, and a wealth of research evidence demonstrates that it is influenced both by genetic and environmental factors (Jensen, 1998; McGue, Bouchard, Iacono, & Lykken, 1993).

Having clarified the meaning of general cognitive ability, aptitude, and aptitude tests, I will provide a brief history of the development of these concepts and of the tests designed to measure ability and aptitude.   Some familiarity with the history of testing is very helpful if the current status and meaning of aptitude, the utility of this concept, and the nature and usefulness of currently available aptitude tests, are to be properly understood.

A Brief History of Psychological Ability Testing

The Origins of Testing

Psychological testing is not a recent development. Its origins can be traced back to China over 3000 years ago when the Emperor had his officials assessed to examine the extent to which they were fit for office (Higgins & Sun, 2002). In this *kejue* examination candidates were assessed in the "6 Arts" of music, archery, horsemanship, writing, arithmetic, and ceremonial rites. By the Han Dynasty (202 B.C - 200 A.D.) the kejue had evolved into written examinations on the "5 Studies" of civil law, military affairs, agriculture, revenue, and geography; and during the 7th century AD it became a national selection system, testing the ability of candidates to remember and interpret Confucian classics by writing essays, composing poetry, completing classical sentences, and choosing antonyms and homonyms.

The kejue had some of the characteristics of modern psychological tests. For example, the assessments were based on the assumption that by evaluating the performance of individuals over a relatively short period of time in a particular sphere of knowledge (e.g. military affairs) it is possible to predict their future performance in work which requires or benefits from this knowledge. Modern aptitude tests work in the same way; a person's performance over a short period of time is assessed, and this is used to predict their ability over a broad range of activities in the future. However, despite such similarities, these early Chinese examinations were essentially assessing knowledge in a particular domain, and as such they have more in common with current methods for assessing educational achievement, such as GCSEs and A levels, than with modern psychological tests.

The development of modern psychological assessment began in the 17th century. Christian Thomasius, a German philosopher, used judges to assess the extent to which individuals possessed one of the four basic dimensions of personality he believed that he had identified: sensuousness, acquisitiveness, social ambition, and rational love. These assessors were asked to indicate, for each dimension, the score an individual should receive on a 12-point scale from 5 to 50 with 5-point increments. For example, if someone were given a score of 5 on the scale of social ambition this would indicate that they had little or no social ambition at all. If they were assigned a score of 50 this indicated that the assessor judged them to have the maximum amount of social ambition, and if they were assigned a score of 25 or 30 they were thought to have an average amount of social ambition. This is the first recorded use of rating scales in the assessment of psychological characteristics, and probably the first example of the collection and use of systematic quantitative data in the history of psychology.

Some three hundred yeas later, early psychologists such as Wilhelm Wundt in Germany, Sir Francis Galton in Great Britain, and James McKeen Cattell in the United States, began to carefully apply scientific principles and methods to measure psychological characteristics, and in particular to measure human abilities. In 1862 Wundt used his “thought meter”, a pendulum with needles attached, which struck bells on either side, to try and assess how swiftly participants could think. As the pendulum swung from side to side, the needles attached to it struck the bells and rang them. The task of the observer was to record the position of the pendulum when the bells rang. The idea was that by comparing the actual position of the pendulum when the bells rang, to the position the observer perceived the pendulum to be in at this moment, the swiftness of the observer’s thought could be estimated.

Several years later Galton drew on and further developed Wundt’s laboratory methods. With new approaches enabling him to test large numbers of people simultaneously, Galton, a cousin of Charles Darwin, investigated the use of reaction time measurements and sensory discrimination tasks in the assessment of human intellect. He measured reaction times objectively with accurate scientific devices under controlled and standardized conditions. In essence, the scientific approaches to measurement, developed in the successful physical sciences such as physics and chemistry, were being applied in the psychological laboratory. Although later research, and in particular a critical study by Wisler (1901), showed that Galton’s aim, to assess complex human abilities by measuring simple ones such as reaction times, was ultimately futile, his focus on clearly specifying a set of psychological dimensions to be measured, his care in standardising measurement procedures, and his use of objective measurement techniques, have had a lasting impact on the field of psychological testing.

The Birth of the Modern Ability Test

The failure to find significant correlations between reaction times and, for example, university performance (Wisler, 1901), led early psychologists to realize that if the intention was to predict the performance of individuals on tasks requiring complex cognitive processes such as reasoning and problem solving, it was necessary to assess their performance on tests which required use of the same mental faculties. This realization led to the development of the modern psychological test of intellectual ability. The first such test was invented by the French psychologist Alfred Binet and his collaborator Théodore Simon in 1905. Unlike Wundt, Galton, and Cattell, Binet was addressing a practical problem. Universal education for all children was introduced in France in 1881, but with the beginning of the 20th century France was lagging behind other countries in its education provision for children with what are now referred to as learning difficulties. Because the evaluation of children by their teachers was not entirely trusted, some way of establishing the ability of these children, particularly of distinguishing between those with learning difficulties and others, was required. Without such a method it was deemed impossible to effectively identify children in need of special educational resources who could be helpfully placed in a suitable educational environment. In 1896 Binet and his assistant Victor Henri published a paper in which they argued that Galton's attempts to assess mental ability with simple reaction times were mistaken. Instead, they suggested, it was necessary to measure higher cognitive processes. About a decade later, Binet and Simon (1905) published the first measure of general mental ability based on the assessment of these higher mental processes. The test consisted of 30 scales, some of them measuring elementary abilities (e.g. can a child follow a moving object with his eyes, or grasp a small object), some measuring more complex abilities (e.g. can a child repeat a sentence of 15 words, or explain how common objects such as paper and cardboard are different), and some measuring relatively abstract and complex abilities (e.g. can the child differentiate between the concepts of "boredom" and "weariness").

In 1908 Binet and Simon published a revised and enlarged scale consisting of 58 items. The most important feature of this new scale was that about 300 "normal" children between the ages of 3 and 13 completed it, and this made it possible to indicate the score of the average child of particular ages (e.g. the average score on the test for 9 year olds, 10 year olds etc.). This innovation enabled the mental level of a child to be determined in relation to these averages. For example, if a 10 year old child obtained a score on the test approximating to the average score obtained by 12 year olds, the mental level (a term which was soon modified by others to "mental age") of the child was said to be 12. Stern (1912) suggested that a useful index of intellectual ability, an "intelligence quotient", could be obtained if a child's measured mental age was divided by his or her chronological age. Shortly afterwards Terman (1916), a psychologist at Stanford University, further revised the test. This latest version became known as the Stanford-Binet test, and it was used as a standard measure of the ability of children for many decades afterwards. Terman also suggested that the intelligence quotient derived by dividing mental age by chronological age should be multiplied by 100 in order to remove fractions, and for the first time in history, referred to this intelligence quotient in the form of an acronym familiar to the public to this day: the IQ.

The Introduction of Group Testing

The original 1905 Binet-Simon test was a breakthrough in the measurement of human ability, and it provided the blueprint for the development of almost all subsequent tests of the cognitive ability of children and adults. However, the widespread use of this new technique for measuring mental ability was hindered because it could not be administered to large numbers of people simultaneously. The test required a trained assessor to evaluate one child at a time. This reliance on the one-to-one administration of tests changed radically in 1917 when the United States entered World War 1. Suddenly the US Army had 1.75 million recruits, and some effective way of assigning them to different roles had to be found. A Harvard University professor, Robert M. Yerkes, argued that this task could be achieved quickly, economically, and effectively by using cognitive ability tests. Yerkes assembled a “Committee on the Examination of Recruits” and they developed two “group tests”, or tests that could be administered to large numbers of people simultaneously. These tests, the Army Alpha and Army Beta, had a profound influence on the development of subsequent ability tests. One reason for this is that they used several clearly defined areas of focus, namely:

- Following oral directions
- Arithmetical reasoning
- Practical judgment
- Synonym-antonym pairs
- Disarranged sentences
- Number series completion
- Analogies

- Information

However, perhaps the most important legacy of the of the Army Alpha and Beta was that they demonstrated that if cognitive ability tests are designed appropriately, they can be administered simultaneously to people on a very large scale. No longer was it necessary to use a trained psychologist to administer a test, and interpret the test results, one person at a time. Instead tests could be administered to large numbers of people at the same sitting, and the ability of each person could be estimated by simply adding up the number of items he or she answered correctly.

The New Science of Psychological Testing

By the 1920's therefore, the template for tests of cognitive ability had been formed. The tests were developed by trained psychologists who set about measuring the general cognitive ability of people, their aptitudes, and/or their ability to do some task deemed important to high levels of performance in a particular task or job role (e.g. clerical work). These tests tended to examine the ability to perform tasks requiring levels of higher cognitive ability as introduced by Binet rather than the elementary tasks such as those measuring reaction times which had been favoured by Galton. Tests were standardized in that people were asked to complete them under very similar conditions and were given the same items or problems to respond to. And the responses to tests were often quantified on scales, making it possible to infer that person X was, for example, not only more able at arithmetic than person Y, but more able by a particular degree or amount. Indeed it became possible to plot the distribution of people on these scales, and to examine the nature of these distributions. By using quantification, standardization, and systematic approaches, psychological tests by the 1920's were mirroring scientific approaches found to be so successful in the longer established sciences. And when psychologists found that the distributions of the scores that people obtained on tests generally follow the normal or Gaussian distribution often found in those "hard" sciences, the scientific credibility of psychological testing was enhanced still further.

The Growth of the Ability and Aptitude Testing Industry

Although many recruits were tested with Army Alpha and Beta during the First World War, the US Army made little use of the results, partly because they were suspicious of the newly developing science of psychology. However, when the tests were released for general use they became extremely influential, and informed the development of a broad range of subsequent tests including “intelligence” tests, aptitude tests, and, in the United States, college entrance examinations and scholastic achievement tests.

For example, in 1916 the US government set up the National Research Council (NRC). The aim of the NRC was to respond to the need for scientific research on a variety of projects considered important after the United States entered World War 1. One project on which scientists were employed was the development of a new test of child mental ability. The resulting measure, the National Intelligence Test, was administered to over seven million children in the 1920’s.

Shortly afterwards steps were taken to apply the new science of psychological testing to the selection of college students. In 1925 the College Entrance Examination Board (CEEB), the body responsible for overseeing the selection of US college students, developed scholastic aptitude tests for use in college admissions. These tests consisted of a set of problems which are still familiar in tests used today such as completing analogies, unscrambling sentences, and deciding on the next number in a sequence. These early tests eventually evolved into the College Board tests, the most notable of which was the Scholastic Aptitude Test introduced in 1926. The Scholastic Aptitude Test was widely used for college selection for many years. In 1994 the Scholastic Aptitude Test was renamed the Scholastic Assessment Test, and in 2004 it was renamed the SAT. The SAT is currently used as part of the admissions process by over 2,000 colleges and universities in the United States.

In 1947 CEEB, the American Council on Education, and the Carnegie Foundation for the Advancement of Teaching contributed their testing programs, a share of their assets and several key employees to form an independent non-profit organization called the Educational Testing Service (ETS). Today the ETS currently administers and scores more than 50 million test uses annually in more than 180 countries. As well as overseeing the administration of tests, the ETS has also been involved in the development of new tests. These include the Graduate Record Examination widely used as part of the selection process to US graduate schools, and the well known TOEFL test of English language ability. In the year it was established the ETS set up the Law School Admissions Council (LSAC), and in 1948 the LSAC introduced the Law Schools Admissions Test (LSAT) now used in the selection of almost all law school students in the United States. The extraordinary growth in the development of psychological tests in the last 100 years is evidenced by the fact that the ETS currently has a database of over 25,000 individual tests and other measurement devices developed worldwide.

In the UK, the development of testing as an industry has been rather different. Whereas the US government provided a large amount of financial and other resources for the growth of testing, this did not occur in the UK. The closest to a non-profit equivalent of the ETS is the National Foundation for Educational Research (NFER) an organization formed in the 1947 with funding from the British government, local education authorities in England and Wales, and teachers' unions. Whilst the NFER has for many years carried out research on psychological testing, it has not been involved in the large scale development or administration of tests. Instead this organization carries out a broad range of research on education and child services in the UK, with research on testing only forming a limited part of this activity. In contrast to the United States, the development of tests in the UK has almost exclusively taken place in the private sector. There are currently some 30 different private-sector test providers in the UK, with each one typically offering a variety of tests, sometimes in combination with other professional services such as management consultancy. Some of the tests supplied by these organizations have been developed in-house in the UK, whilst others have been developed overseas (often in the United States) and standardized for use in this country.

Aptitudes and Ability in the 20th Century

When, at the beginning of the 20th century, Binet measured the vocabulary of different children, and also their ability to judge, attend, and engage successfully in other important psychological tasks, he found that children differed in the extent to which they were good at them. Over time, Binet developed a simple taxonomy of ability. He argued that the essential features of intelligent behaviour are “to take and maintain a definite direction”, to “make adaptations” in order to arrive at the desired goal, and the “power of autocriticism”. He believed that whilst these faculties of direction, flexibility and judgement were associated, it was nevertheless important to distinguish between them. In addition, Binet was prepared to accept that emotion and personality also contribute to an individual’s general functioning and ability. One implication of Binet’s assumptions was that the ability to perform a task or job effectively or to reach a particular level of educational achievement depended on a number of different psychological processes working together in an organized and coordinated fashion. A second implication, critical for the future development of psychological tests, was that intellectual ability consists of a variety of different components and facets, and that because people can be more or less able with respect to each of these facets it is necessary to measure all of them if a full and complete picture of someone's ability and potential is to be ascertained. In other words, tests must yield ability profiles rather than a single overarching index of general cognitive ability.

Based on this assumption, a large number of tests were developed by psychologists, each designed to measure a specific set of psychological abilities, with the abilities measured often overlapping across different tests. This raised a question of fundamental importance in ability testing. If these tests measured different sets of abilities, which abilities were the relatively important ones in determining an individual's overall intellectual capability, and which were relatively trivial? Research on this question had begun at the turn of the 20th century (Spearman, 1904), but its proper development depended on the development of a statistical technique capable of solving the problem. Such a technique, factor analysis, has been available in a very rudimentary form since about 1900, but it was not until the late 1930's that it had developed sufficiently to allow researchers to draw some tentative conclusions about the underlying structure of cognitive ability. This work was undertaken in the United States by Spearman (1927), Kelley (1928) and Thurstone (1938). Despite these psychologists using the same statistical technique, factor analysis, and similar data (the performance of large groups of people on tests measuring a variety of different types of cognitive ability), they drew radically different conclusions about the fundamental nature of intellectual ability. One conclusion, associated with Spearman, is that there is one underlying dimension of cognitive ability. In other words people differ in their overall ability rather than with respect to specific elements of ability. Spearman referred to this general ability as *g,* an abbreviation still widely used by psychologists. The other conclusion, associated with Thurstone and others, is that intellectual ability is made up of several quite different components, and that it is essential to measure people on each these different elements separately if we want to know about their overall ability and the areas in which they have (and do not have) talent.

A Single Factor of Cognitive Ability?

In the late 19th and early 20th centuries Spearman studied a large number of correlations between the performance of people on different measures of ability including their performance at Classics, French, English, mathematics, and music. After analysing this data, and in the process helping to develop an early version of the now widely used and highly influential statistical technique of factor analysis, Spearman (1904, 1923, 1927) concluded that "intelligence" is composed of two types. The first consists of a general factor which he named *g*, and the second consists of a large number of specific factors which he referred to as s^1, s^2, s^3 etc. He believed that of these two types of ability, *g* is by far the most important. That is, the intellectual ability of people varies primarily in the extent to which they have a general and all-round capacity to understand things (Spearman called this "apprehension of experience"), recognize relationships between concepts ("eduction of relations"), and solve problems by applying principles understood in one domain to another ("eduction of correlates"). Whilst he acknowledged that people also vary in the extent to which they have the mental capacity to perform well in specific domains (e.g. some people may be better at solving verbal problems whereas others have a greater ability to solve numerical problems), Spearman believed that the influence of variation between people in their ability to perform well in specific domains is far less influential than the variation between them in *g* because ability in these areas was highly correlated. If people tended to be good at some things they tended to be good at others. Put simply, Spearman concluded that easily the greatest influence on an individual's ability to perform well in a specific task was their general mental ability.

Multiple Components of Cognitive Ability?

Spearman's emphasis on the importance of *g* was not shared by all psychologists – particularly those in the United States. Many took the view that specific abilities were considerably more important than Spearman believed. Of these dissenting psychologists, Thurstone was particularly influential in the first half of the 20th century. When Thurstone (1931) applied factor analysis to the scores obtained by people on a range of different ability scales he concluded that there were seven primary factors of cognitive ability:

- verbal comprehension

- word fluency
- number facility
- spatial ability
- associative memory
- perceptual speed
- general reasoning

The underlying assumption of this alternative school of thought is captured in the following quotation:

> Evidence from biology, from genetics, from sociology, from education, from anthropology, and from common sense, as well as from psychology, persistently suggests……that what is called intelligence is a mixture of quite different attributes having different genetic and environmental determinants, different courses of development over the life span, and different implications for understanding human achievement, human failings, human creativity, and human happiness (Horn, in Sternberg, 1986, p. 36).

The Sixty Year Reign of Aptitudes

Between the 1930s and 1980s most psychologists in North America studying and researching intelligence and individual differences in ability, and designing tests to measure these differences, subscribed to Thurstone’s multi-component view of intellectual ability. Prominent amongst the psychologists who have developed theories in this tradition are Guildford (1967, 1985) who carried on Thurstone’s search for the primary elements of ability, Gardner (1983, 1992) who developed a theory of multiple intelligences (these separate areas of intelligence include linguistic, logical mathematical, spatial, musical, bodily kinesthetic, interpersonal and intrapersonal), and Sternberg (1985, 1986, 1996) who developed a “triarchic” model of ability made up of componential (or analytical) intelligence, experiential (creative) intelligence, and contextual (practical) intelligence.

During this period a large number of aptitude tests were developed (see Appendix 2). Prominent among the many tests devised are the Differential Aptitude Test (DAT), the General Aptitude Test Battery (GATB), the Armed Services Vocational Aptitude

Battery (ASVAB), the Scholastic Aptitude Test (SAT), the American College Test (ACT), the Graduate Record Exam (GRE), the Medical College Admissions Test (MCAT), and the Law School Admission Test (LSAT). All of these tests were developed in North America and are still used there on a large scale. The SAT and ACT are used for selection into higher education, the GRE is used for selecting postgraduate students, MCAT is used as part of the medical school selection process, and LSAT is used in the selection of law school students. The current version of the LSAT is used as part of the selection process by all law schools in the United States approved by the American Bar Association and administered to about 170,000 law school applicants annually in the United States and Canada.

Compared to North America, aptitude testing in the UK has been patchy. They are not used systematically, and on a widespread scale, for either undergraduate or postgraduate university admissions. In some organizations aptitude tests are used for specific purpose. For example, NATS Ltd (formerly National Air Traffic Services) use aptitude tests in the selection of air traffic controllers. However, as discussed on in the section on UK Professional Service Personnel beginning on page 59, the reticence of UK test publishers to disclose relevant information about their aptitude tests means that the nature and frequency of the use of these tests is very difficult to gauge.

In relation to the selection of professionals in the UK, it would appear that four tests are used on a significant scale. The UK Clinical Aptitude Test (UKCAT) is currently used as part of the selection process for undergraduate medical students by 26 UK medical schools, and the BioMedical Admissions Test (BMAT) by a further six of them; the LNAT test is used for selection by several UK law schools; and the National Recruitment Office which coordinates quality assurance in the UK for the recruitment and selection of general practitioners have developed Situational Judgments Tests (SJTs) (see Appendix 1) to do so. Whilst SJTs are not generally considered a form of aptitude test, they are sufficiently similar to warrant a mention here.

These UK aptitude or aptitude-like tests will be discussed in more detail later in the section *Aptitude Tests in the UK and Ireland*. However, before moving on to consider specific aptitude tests in more detail it is important to focus on changes in the academic status of aptitude and aptitude tests over the last 20 years as these have very significant implications not only for the concept of aptitudes but also for the development, use, and interpretation of aptitude tests.

Aptitudes and Ability in the 21st Century

During the last 20 years the data accumulated from hundreds of studies of cognitive ability carried out over several decades have been subjected to a variety of sophisticated statistical analyses. As a consequence it has been possible to draw several critical conclusions, conclusions which today are accepted by most psychologists working on cognitive ability. The three most important conclusions are probably as follows:

(a) Most of the variation in cognitive ability is associated with *g* rather than with specific cognitive aptitudes or abilities.

(b) Cognitive ability is a very important construct, and variation in cognitive ability has a substantial impact on a broad range of life experiences and outcomes.

(c) In North America there are substantial sub-group differences in measured *g*. The average measured *g* of people from Black African and Hispanic backgrounds is substantially lower than the average measured *g* of their white counterparts. These sub-group differences are larger than those found with other common selection methods. Consequently the use of tests of *g* in personnel selection will usually result in a lower proportion of candidates from Black African and Hispanic backgrounds being selected than will the use of other selection methods.

These issues are discussed in turn below.

The Fall of Aptitudes and the Rise of g

In the last 20 years a large body of data from aptitude and ability tests has been analysed by a variety of prominent researchers (e.g. Hunter, 1986; Olea & Ree, 1994; Ree, Earles, & Teachout, 1994; Schmidt, Ones, & Hunter, 1992). These analyses have produced two findings of critical importance when considering the viability of aptitude tests. First, factor analyses (for a brief introduction to factor analysis in the context of cognitive ability see Appendix 2) of large numbers of aptitude tests have revealed that one overarching ability factor, Spearman's *g*, accounts for most of the correlations between different tests (Brand, 1996; Gottfredson, 1997; Jensen, 1998). Second, this *g* factor emerges strongly across different test batteries, the method of factor extraction used in factor analysis, and racial, cultural, ethnic and nationality groups (Reeve & Hakel, 2002). As Gottfredson (2002, p26) puts it:

> “People who do well on one mental test tend to do well on all others. When the scores on a large, diverse battery of mental ability tests are factor analyzed, they yield a large common factor, labelled *g.* Pick any test of mental aptitude or achievement – say verbal aptitude, spatial visualization, the SAT, a standardized test of academic achievement in 8th grade, or the Block Design or Memory for Sentences subtests of the Stanford-Binet intelligence test - and you will find that it measures mostly *g.* All efforts to build meaningful mental tests that do not measure *g* have failed.

The second finding relates to the following question: if performance on a job or training programme is first predicted with a measure of peoples’ general cognitive ability (that is, *g*), and then it is predicted again with measures of their general cognitive ability *plus* measures of their specific aptitudes, does the latter prediction model outperform the former? That is, does the measurement of aptitudes add anything over and above general cognitive ability in predicting peoples’ performance? If aptitudes do provide a useful amount of predictiveness over and above general cognitive ability then their inclusion as a part of the selection process is justified. But if measuring peoples’ aptitudes does not result in more accurate prediction of performance than measuring their general cognitive ability alone, there is little justification for the use of specific aptitudes in personnel selection. The results of these analyses are quite clear. When information about intellectual aptitudes are added to information about general cognitive ability there is little or no increase in the accuracy with which the job performance or training performance of people can be predicted (Gottfredson, 2002; Jensen, 1998; Ree & Carretta, 2002; Ree & Earles, 1991; Ree, et al., 1994; Schmidt, 2002). Schmidt (2002) shows that the specific tasks that people are required to do in different jobs has little impact on the predictiveness of general cognitive ability. The implication of this research is that identifying the specific tasks that people need to do in a given job, and identifying the cognitive abilities required to do each of these tasks, and then selecting aptitude tests which measure these abilities, is unnecessary. All that is required is an ability test which provides a good measure of *g*.

The Importance of Cognitive Ability

Hunter and Hunter (1984) carried out a large scale meta-analysis of over 400 individual studies examining the extent to which cognitive ability predicts job performance. They estimated the average correlation between general cognitive ability and job performance to be .57 for high complexity jobs, .51 for medium complexity jobs, and .38 for low complexity jobs.   These validities increase further still when objective rather than subjective (supervisor rating) measures of job performance are used (Schmidt, 2002). A comparison of the results of meta-analyses carried out on a broad selection of personnel selection techniques shows that tests of general cognitive ability have a level of predictive validity which is as high, or higher, than any method for selecting personnel, including structured interviews, work samples, assessment centres, personality questionnaires, job knowledge tests, job tryout, and all other widely used techniques for selecting staff (Schmidt & Hunter, 1998). A comparison of the extent to which different selection techniques predict job performance is shown in Table 1. Table 1 shows that only structured interviews are able to predict job performance as well as cognitive ability tests.   It should be noted that although the Hunter and Hunter (1984) meta-analysis of the predictiveness of cognitive ability tests used in Schmidt and Hunter's paper is North American and rather dated, a more recent European meta-analysis (Salgado & Anderson, 2002) found the validity of cognitive ability tests in predicting job to be at least as high as Hunter and Hunter's estimates.

Table 1

A Comparison of the Extent to Which Different Selection Methods Predict Job Performance

Selection method	**Validity coefficient**
Cognitive ability tests	0.51
Employment interviews (structured)	0.51
Job knowledge tests	0.48
Training and experience evaluation (behavioural consistency method)	0.45
Job tryout procedure	0.44
Integrity tests	0.41
Employment interviews (unstructured)	0.38
Assessment centres	0.37
Biographical data	0.35
Work sample tests	0.33
Conscientiousness (Big Five personality factor)	0.31
Reference checks	0.26
Job experience (years)	0.18
Training and experience evaluation (point method)	0.11
Years of education	0.10
Interests	0.10
Graphology	0.02

Source: Schmidt & Hunter (1998) with the exception of the result for work sample tests which is taken from Roth et al. (2005).

As well as being a relatively good predictor of job performance, cognitive ability tests are also good at predicting training performance. A review by Hunter (1986) of military databases containing information about the measured cognitive ability and training performance of 82,000 trainees revealed the average validity to be .63. This figure is similar to that found by other researchers including Ree and Earles (1991), Thorndike (1986), Jensen (1986), and Hunter and Hunter (1984).

Furthermore, there is evidence that general cognitive ability is associated with a broad range of life events and outcomes. Schmidt (2002) provides the following list from the work of Brody (1992), Herrnstein and Murray (1994), and Jensen (1980, 1998). The life events and outcomes associated with *g* include:

- School performance and achievement through elementary school, high school, and college
- Ultimate education level attained
- Adult occupational level
- Adult income
- A wide variety of measures of “adjustment” at all ages
- Disciplinary problems from kindergarten to 12th grade (negative relation)
- Delinquency and criminal behaviour (negative relation)
- Accident rates at work (negative relation)
- Poverty (negative relation)
- Divorce (negative relation)
- Having an illegitimate child (negative relation for women)
- Being on welfare (negative relation)

If cognitive ability is an important predictor of a broad range of outcomes, including educational and job performance, how does the degree to which it predicts such variables change over time? Zyphur et al. (2008) investigated this issue in a study of the extent to which cognitive ability and the personality variable of conscientiousness predicted both initial performance and changes in performance over time. They found that whilst cognitive ability was a better predictor of initial performance than conscientiousness, after the third semester conscientiousness was a better predictor than cognitive ability. Furthermore, whilst conscientiousness predicted changes in performance over time, cognitive ability did not. This suggests that performance is a dynamic variable, influenced by both “can do” variables such as cognitive ability and “will do” motivationally relevant variables such as conscientiousness. The implication of this is that whilst cognitive ability may be a comparatively good predictor of performance when all people are highly motivated by the situation they are in, when situation-specific motivation is reduced, dispositional variables such as conscientiousness can have a marked influence on performance. In intellectually challenging environments it is not usually enough to be able - it is also necessary to be motivated. Tests of ability or aptitude which rely very heavily on cognitive ability variables may therefore be less predictive of long-term performance than those which also measure variables associated the tendency to be motivated, either generally, or in the context, environment or domain in which future performance is to take place.

Adverse Impact: Sub-Group Differences in g

Despite the strong associations found between measures of *g* on the one hand, and job performance, training performance, and a range of life outcomes on the other, the adoption of ability tests for personnel selection in the United States has for many years been controversial. The primary reason for this is the abundance of research showing that there are substantial differences in the average scores of people from different ethnic groups on cognitive ability measures (Ployhart, Schneider, & Schmitt, 2006). In general, the average score of African Americans is about 0.75 to 1 standard deviation units below the average score for Whites, and the average score of Hispanic Americans is about 0.75 of a standard deviation unit lower than Whites. Even though this difference between Whites and others is less pronounced for high complexity jobs (Hough, Oswald, & Ployhart, 2001), it is still substantial. Asians are found to usually perform better than Whites on numerical measures of cognitive ability, but worse on verbal measures.

These findings have a clear and important practical implication: if cognitive ability tests are used as the sole means of personnel selection there will be a very marked impact on the proportion of people selected from different ethnic groups. For example, let us assume that Black Americans have cognitive ability scores on average one standard deviation below Whites, and that the population of scores from which these are sampled are normally distributed. In these circumstances if 10% of

Whites are selected for a job solely using cognitive ability test results, only about 1% of Black Africans will be selected; if 50% of Whites are selected only about 16% of Black Africans will be selected, and if 90% of Whites are selected, only about 60% of Black Africans will be selected (Ployhart, et al., 2006).

The tendency for White people to outperform people from Black racial and ethic backgrounds on tests of cognitive ability has for a long time produced a great deal of controversy in psychological testing. A variety of explanations for the phenomena have been offered, including the way that performance is defined and measured, the way that the cognitive test is presented, and the differential motivation of test-takers from different ethnic and racial groups, though none have resulted in the development of a technique that can substantially reduce the adverse impact of cognitive ability tests on people from these groups. Combining cognitive ability tests with selection methods associated with smaller sub-group differences than cognitive ability tests (e.g. personality questionnaires) can reduce sub-group differences in selection ratios to some degree, but does not eliminate them. Indeed, for the use of alternative selection techniques to eliminate the average differences found in favour of whites with cognitive ability tests, these selection techniques would have to produce results in which black people outperform whites to a degree comparable to that by which whites outperform blacks on tests of cognitive ability.

So far the discussion of sub-group differences in cognitive ability test scores has focussed on North American research. What evidence is there for such differences outside North America? Unfortunately there is little published research on this issue. One study carried out in the Netherlands (Nijenhuis & vanderFlier, 1997) found that the difference on verbal and numerical ability tests between the majority white Dutch on the one hand, and immigrants from Surinam, the Antibes, North Africa, and Turkey on the other, were between one and two standard deviations. These results suggest that sub-group differences in cognitive ability tests in Holland at this time were even more pronounced than those between Whites and Blacks found in North America.

I have not been able to identify any peer-reviewed articles on sub-group differences with respect to ethnicity or racial groups carried out in the UK or on people from outside the UK applying for training or education in this country. However, I have been able to carry out a limited investigation of this issue by examining some relevant data available on the *LNAT*, the test used by some UK Law Schools as part of the process for selecting undergraduate law students. Although the results of this aptitude test are used in combination with other candidate information such as A level results by these schools when selection decisions are made, it is nevertheless possible to examine the degree to which the LNAT would produce differences in the proportion of people selected with different educational, ethnic, and social class backgrounds if this test was used as the sole means of candidate selection. To do so I have drawn upon data posted on the LNAT website http://www.lnat.ac.uk/ and relating to candidates tested in years 2007-2008. In Tables 2, 3, and 4 the proportions of people who would be selected with the LNAT if scores on this test were used as the sole means of selection for law undergraduates are set out. Table 2 shows the results for educational background, Table 3 for ethnic background, and Table 4 for parental occupation. In order to interpret these tables, take a particular LNAT score and regard it as the cut-off for selection purposes. For example, a decision might be taken to accept all candidates with a score of 17 or more. The figures to the right of a given LNAT score show the proportion of candidates from each category that would be selected. For example, Table 3 shows that if the LNAT cut-off score for candidate selection was set at 17, the consequence would be that 51% of White British candidates would be selected, whereas the proportion of people with Black African, Indian, and Pakistani, backgrounds selected would be 30%, 27% and 27% respectively.

The data presented on the LNAT in Tables 2 to 4 are limited to only some of the candidate categories about which published information is available, and only focus on the 2007-08 year of application. Nevertheless these data suggest that there are some substantial differences in the LNAT performance of undergraduate law candidates from different educational, ethnic, and social class backgrounds groups. This issue is clearly worthy of a detailed investigation using a more extensive data set to explore whether or not a similar pattern is revealed.

Table 2

The Percentage of Candidates who Would be Selected at LNAT Cut-off Scores

By Type of School Attended

Based on Data for UK candidates 2007-2008

	Percentage of Candidates Selected by Type of School Attended				
LNAT Score	**Grammar** N=663	**Independent** N=670	**Comprehensive** N=793	**Sixth form college** N=1,538	**College of further education** N=395
4	100	100	100	100	100
5	100	100	100	100	99
6	100	100	100	100	99
7	100	100	99	99	98
8	100	99	99	98	98
9	99	99	99	97	94
10	98	98	97	94	89
11	97	95	94	91	83
12	94	92	91	85	78
13	91	87	86	78	70
14	85	84	80	70	63
15	79	75	73	61	53
16	71	66	63	51	45
17	57	55	51	36	33
18	42	39	37	24	22
19	32	31	27	17	14
20	21	19	20	11	7
21	11	11	13	6	5
22	6	5	7	3	3
23	2	2	4	1	1
24	1	1	2	0	0
25	0	0	1	0	0
26	0	0	0	0	0
27	0	0	0	0	0

Table 3

Percentage of Candidates Who Would be Selected at LNAT Cut-off Scores by Candidate Ethnic Background

Based on Data for UK candidates 2007-2008

	Percentage of Candidates Selected by Ethnic Background			
LNAT Score	**White British** N=2,954	**Black African** N=199	**Indian** N=284	**Pakistani** N=141
4	100	99	100	99
5	100	99	100	99
6	100	99	100	98
7	100	98	98	95
8	99	96	97	94
9	99	93	93	88
10	98	85	88	81
11	96	80	82	79
12	93	72	74	71
13	88	65	67	61
14	82	59	58	52
15	74	49	49	42
16	65	40	39	34
17	51	30	27	27
18	36	23	19	14
19	27	14	14	9
20	17	9	9	4
21	12	5	6	2
22	6	3	2	1
23	2	2	2	0
24	1	0	1	0
25	0	0	0	0
26	0	0	0	0
27	0	0	0	0

Table 4

Percentage of Candidates Who Would be Selected at LNAT Cut-off Scores

by Householder Occupation

Based on Data for UK candidates 2007-2008

	Percentage of Candidates Selected by Householder Occupation			
LNAT Score	**Senior professional** N=819	**Senior manager or official** N=700	**Skilled tradesperson** N=170	**Manual worker** N=174
4	100	100	100	99
5	100	100	100	99
6	100	100	100	99
7	100	99	99	98
8	100	99	99	97
9	99	99	98	94
10	97	98	93	91
11	95	95	91	84
12	92	92	85	80
13	87	86	79	74
14	81	82	74	63
15	74	74	64	54
16	64	64	55	43
17	51	49	38	29
18	39	35	28	16
19	30	26	19	11
20	19	17	12	6
21	12	10	5	4
22	6	5	2	1
23	2	3	1	1
24	0	1	1	0
25	0	0	1	0
26	0	0	0	0
27	0	0	0	0

The Influence of Practice and Coaching on Aptitude Test Results

There is meta-analytic evidence that a combination of test practice and coaching increases candidate performance on aptitude tests such as the SAT by about .76 of a standard deviation (Bangertdrowns, Kulik, & Kulik, 1983a, 1983b; Kulik, Bangertdrowns, & Kulik, 1984). This is a considerable effect, equivalent to an increase of about 10 points on an IQ test. The consequence would be that if there were two candidates of equal ability, and one was exposed to a practice version of a test and then coaching on it, and the other had neither the practice version nor the coaching, the former would have a considerably greater chance of being selected than the latter, the exact degree of this advantage depending on the ratio of selected to unselected applicants overall. To the extent that the availability of practice tests and coaching differs across candidates for reasons of geographical location, financial resources, time availability, awareness of practice tests and coaching programmes etc., we might expect the potential predictiveness of aptitude tests to be attenuated. Indeed, unless (a) all candidates have access to equally effective practice tests and coaching, or (b) none do so, less able candidates will outperform more able candidates on the tests, and vice versa. This effect not only reduces the effectiveness of such tests for selection purposes, but also introduces unfairness into the selection process.

The Nature of Aptitude Tests and the Influences on Aptitude Test Results

In this section I will draw together some of the themes and findings discussed so far and comment on their implications for aptitude testing. At first sight the idea of aptitude testing is straightforward and appealing. If we want to know how suited someone is to a particular task, training programme, or job, we can assess the degree to which they have the necessary aptitude for it. By selecting people with the right aptitude profile it may appear that we will have identified the people most likely to perform well. If there are more applicants performing at ceiling in GCSE's, A levels, or degrees than there are places available, aptitude tests appear to provide an objective tool by which these applicants can be ranked. And it would seem that by using these tests men and women with the right aptitudes for a job who have failed to achieve their full educational potential perhaps because they have had relatively poor educational backgrounds, can be given the chance they deserve.

However, when aptitude tests and the way they are used are examined more closely, a more complex picture emerges. The descriptions of a range of widely used aptitude tests in Appendix 2 shows that almost all of them measure reasoning in some form, often reasoning that is linked to verbal, numerical, or spatial problems. These tests are therefore measuring *g*. Put slightly differently, whilst they appear to be measuring distinctive sets of aptitudes, they are all, or almost all, primarily assessing *g* or general cognitive ability. To the extent that these tests are measuring cognitive ability rather than, for example, specific areas of job knowledge, research indicates that they are unlikely to provide a better prediction of job or training performance than a test designed to measure *g*.

Furthermore, when aptitude tests are not concerned with measuring factors such as verbal ability, numerical ability, and spatial ability which load very highly on *g* (and indeed can be considered alternative ways of measuring *g*), they appear in most cases to measure specific areas of knowledge or, more generally, attainment. To take two examples, one of the three components of the BMAT aptitude test used by some UK medical schools is concerned with the respondent's scientific knowledge, and the well-known Differential Aptitude Test measures, amongst other things, spelling ability. As pointed out at the beginning of this report, tests of scientific knowledge, and tests of how words are correctly spelled, are tests of attainment. For attainment tests making high-level cognitive demands of people (as distinct from ones making physical demands, such as eyesight tests or manual dexterity tests) the results are likely to correlate quite strongly with *g*. Because measured *g* is associated with a variety of environmental factors such as peoples' family background and the nature of their education, scores on these attainment tests will almost certainly be associated with such environmental factors also.

Where does this place aptitude and aptitude tests? Kline (2000, p234) takes a forthright position. He argues that because aptitude tests are actually composed of a mixture of measures of general cognitive ability and measures of attainment, they are "not as valuable as their name suggests", and that from a scientific standpoint the very concept of aptitude "should be abandoned". An alternative position, still cognisant of the implications of the dominance of *g*, the questionable practice of trying to discriminate meaningfully between different types of cognitive ability in aptitude tests, and the observation that aptitude tests are composed of tests of *g* and of attainment, is to say that those developing and using tests of aptitude should be very aware of these observations, should dispute any claims that the results of aptitude tests are independent of environmental influences, and should very carefully scrutinize not only the reliability and validity of their aptitude tests but also the possibility that these tests systematically favour people who have had particular sorts of backgrounds, education, and experiences.

That people with very well educated parents who go to exclusive independent schools and highly selective universities will tend to do better on aptitude tests than those with less advantageous backgrounds is not *necessarily* a reason to avoid using these tests. In fact the accumulation of cognitive skills, knowledge, and other psychological and behavioural characteristics in such environments may be of value in performing certain valuable roles in society. However, any claim that aptitude test results are somehow immune from, and independent of, these influences is unlikely to withstand critical examination. Measured *g* is known to be influenced by both genetic and environmental factors (Jensen, 1998; McGue, et al., 1993), and because aptitude tests tap *g* they are not immune from the impact that the environment has had on the people tested with them.

It is also important to note that systematic differences in life experience are not randomly distributed in society, but are of course associated with various cultural groupings including social class and ethnicity. As discussed earlier, research indicates that in the United States very significant sub-group differences exist in ability test scores, and that these differences are particularly great between people of white and Black African origin (Ployhart, et al., 2006). Although little is known about the relationship between ethnicity and ability test scores in the UK (Cook, 2006), the possibility that the considerable differences found in average scores between different ethnic groups in North America will be replicated in this country, and in people from overseas applying for education and training in this country, is certainly worthy of close attention.

To summarise, aptitude tests are composed of direct measures of *g* such as verbal and numerical ability and of various measures of attainment - intellectually demanding examples of which will be associated with *g*. Aptitude tests would be maximally useful if cognitive abilities were made up of several independent factors. If this were the case, the set of psychological characteristics required in a particular job could be matched with the aptitude profiles of candidates. However, an abundance of research demonstrates that cognitive abilities do not fall into several independent factors. Instead they are highly correlated – they are all indicators of *g*. In circumstances in which people with a wide range of general cognitive ability levels are being selected for cognitively demanding training programmes or jobs, research suggests that *g* is likely to be a very useful predictor of future performance. However, when there is little range in the cognitive ability of candidates, aptitude tests and more direct measures of *g* will be considerably less effective (see the section on *Operational and Corrected Criterion-related Validity*). In addition, the results obtained by candidates on aptitude tests, like direct tests of *g,* are not immune from the long-term effects of environmental influences on candidates including their family background and education, nor are they immune from the effects of opportunities to practice the tests in advance, or from test coaching. Finally, at least in the United States, *g,* is strongly associated with ethnicity. As a consequence of the link between *g* and aptitude tests, the results people obtain on these tests are likely to be associated with ethnicity also. They are also likely to be associated with variables such as level of parental education and type of school attended.

Criteria for Evaluating Psychometric Tests: Reliability and Validity

Having discussed the nature of aptitude and aptitude tests in some detail attention will now be turned to ways in which they can be evaluated. The techniques for assessing an aptitude test are the same as those for doing so with other psychometric tests. Two of the most important features of an effective test are its reliability and validity. There are several ways of examining the reliability and validity of selection methods, and each of these will be described in turn.

Reliability

If someone obtains a score of 45 out of 50 on a psychological test, would they obtain a similar score if they completed the test again (assuming that their score on the second occasion was not affected in any way by the first testing)? An unreliable psychological test (or other personnel selection method) is like an unreliable tape measure – when repeatedly used to measure the same thing it will give a different result almost every time. All psychological tests are reliable to certain degree, and some are more reliable than others.

Reliability can be assessed in several ways. With *test-retest reliability* a group of people are given a test on two occasions and their scores on the first administration are correlated with their scores on the second administration. In the case *of parallel-forms relia*bility two versions of a test are created and the scores respondents obtain on one version are correlated with the scores they obtain on the other. With *split-half* *reliability* the test is split in half, and respondents' scores on one half of the test are correlated with their scores on the other half. In the case of split-half reliability there are many ways to split a test. A fourth reliability estimate, *coefficient alpha*, indicates the average correlation which would be obtained if a test was split in all possible ways and each pair of test items correlated every time.

Three further points should be made in relation to the reliability of psychological tests. First, the four measures of reliability described above are not interchangeable - they measure different things. In particular, only test-retest reliability is sensitive to the extent to which test scores are reproduced across time – that is the degree to which the score a respondent obtains on one occasion is likely to remain stable over repeated test administrations and over time. Second, reliability is not a feature of a test but rather a feature of the test used in a particular situation with a particular group of people. Therefore, a test which has a particular level of reliability in one setting may not have the same level of reliability in another setting. Third, when reliability figures are given for a test these figure are only *estimates* of the reliability of the test in a particular setting. For example, if a test's manual states that the test-retest reliability of an ability measure is .83, this does not indicate that the reliability of the test will be .83 in all settings, or even that the reliability of the test will always be exactly .83 in the setting in which the test-retest reliability study was undertaken. Estimates of reliability are prone to error, and if the reliability of a test is deemed very important in a particular setting it is strongly recommended that the tests' reliability is measured in that setting, that a large sample of respondents (well over 100) are used in the reliability study, and that in the case of split-half, parallel forms and test-retest reliability, confidence intervals for the reliability estimate are provided.

Face validity

A method of selection is said to have face validity if it appears to assess what it is supposed to assess. For example, a test of cognitive ability has face validity if someone looking at the test items would be happy to conclude that they probably do measure cognitive ability rather than some other characteristic such as conscientiousness. Of course, the appearance of such a test can be misleading, and it is not possible to claim that because a test appears to measure cognitive ability it actually does so. Nevertheless, face validity can be important because those taking the test may respond to it in critical ways. A test of cognitive ability which appears to have little or nothing to do with cognitive ability may not be taken seriously by those asked to do it. If those completing it do not take a test seriously, it is likely to produce spurious results.

Construct Validity

Psychological tests are generally designed to measure psychological constructs, such as extraversion and verbal ability. Constructs cannot be directly observed, and they cannot be operationally defined with respect to a single, directly observed, external referent (Cronbach & Meehl, 1955). There is no single behaviour that indicates the extent to which someone is extraverted or cognitively able. Therefore measures of psychological constructs must sample from different behaviours, or samples of behaviours in different types of situations, in order to measure the construct. In the context of cognitive ability, there is no single behaviour which indicates how cognitively able someone is, and it is therefore necessary to examine their responses to a range of situations, such as, with regard to a psychological test, a range of different test questions.   Construct validity is concerned the extent to which a psychological test actually does measure the construct is designed to measure.   A range of techniques have been devised to establish the construct validity of a test, including convergent and discriminant validation, factor analysis, and theory-consistent group differences.

Content Validity

The content validity of a test is a function of the extent to which “the questions, tasks, or items on a test are representative of the universe of behaviour a test is designed to sample” (Gregory, 2010, p.111). In principle, to establish complete content validity, it is necessary to specify all of the items or tasks which could be used to measure a construct, and then to sample from these in constructing a test. In practice, and particularly for broad constructs such as cognitive ability, this is impractical and instead test developers fall back on a technique in which several experts in the field are asked judge the extent to which the test has content validity.

Criterion-related validity

In the context of selection, criterion-related validity is sometimes considered the most important of all the forms of validity. It is said that the relationship between a job and a selection method has criterion-related validity if the performance of people on the selection method predicts their performance on some criterion or set of criteria. The most common criteria are educational performance, training performance and job performance, but other criteria such as job satisfaction, can be used instead.

The index of validity used for criterion-related validity is the “validity coefficient”. The validity coefficient is the correlation coefficient obtained when the scores obtained by a group of people on a selection method, and their performance on a criterion measure are correlated. Criterion-related validity can be broken into two main sub-types: concurrent validity and predictive validity.

Concurrent validity

In the case of concurrent validity, the predictor and job performance criterion are measured for an existing group of employees, and scores on the former are correlated with scores on the latter. An advantage of this approach is that it is often relatively simple to do because the necessary data can be obtained quite easily. For example, a sample of existing sales employees might be asked to complete a cognitive ability test, and their performance on this might then be correlated with their sales performance. If concurrent validity is present, the higher the cognitive ability scores that a salesperson has, the more products they will tend to sell. One disadvantage of the concurrent validity approach is that because the relationship between the selection method and the criterion is being examined on existing employees, the findings will not necessarily be applicable to new employees. These new employees may differ from existing ones in various critical ways (for example, they may be younger and less experienced). As a consequence, the level of concurrent validity identified through a study of existing employees is not necessarily applicable to new applicants.

A second disadvantage of concurrent validity is that the selection method is not being used in the context for which it is intended. For example, an interview is often used for new applicants who know little or nothing about the organization they are applying to work in. If existing employees were interviewed, it would clearly be impossible to replicate the circumstances in which new applicants were interviewed: existing employees cannot easily put themselves in the position of new applicants by ignoring their knowledge of the organization. Therefore concurrent validity studies can only be meaningfully carried out on selection methods such as cognitive ability tests where previous knowledge of the organization and the people working in it is not an advantage.

Predictive validity

In the case of predictive validity, selection measure scores are obtained from job applicants and these are then used to predict the performance of those applicants who are subsequently appointed after they have been in the job for some time. An advantage of predictive validity is that it properly reflects what those involved in selection want to do with a selection method: evaluate applicants when they apply for the job and then use this to predict how they will perform when they are actually working in the organization. Furthermore, unlike the case of concurrent validity, with predictive validity those involved in selection do not have to be concerned that the sample of people upon which they are validating the selection method may be different from the people to be recruited. A disadvantage of the use of predictive validity studies is that they normally take months or even years to carry out because of the time-lapse between the collection of the selection measure data and the availability of performance data.

Despite the clear strengths of predictive validity studies, research suggests that they do not provide a better indication of the criterion-related validity of a selection measure than concurrent validity studies. Barrett et al. (1981), and Schmitt et al. (1984) examined the validity coefficients found for certain types of selection test when both concurrent and predictive tests of criterion-related validity were carried out. They found little difference in the validity coefficients obtained using the two methods. Partly for this reason I will henceforth refer to both predictive and concurrent validity as criterion-related validity.

Operational and "Corrected" Criterion-Related Validity

It is now common for "meta-analytic" studies to be carried out to examine the relationship between psychological tests and personnel selection techniques such as interviewing on the one hand, and criterion measures such as job or training performance on the other. In these meta-analyses the criterion-related validity of a variety of predictors is estimated using data from several previously conducted studies. An advantage of these meta-analyses is that they tend to be less prone to "sampling error" than single studies. That is, the "true" correlation between a particular predictor, such as cognitive ability tests and a criterion such as job training performance may be biased and distorted in a particular study by such factors as the specific type of cognitive ability test used, the nature of the job performance predicted, the limited sample of people on whom both cognitive ability test and job performance data are available, and unreliability in the measurement of training performance. By statistically combining the results of several studies the effects of these artefacts can be considerably reduced.

Meta-analyses can be used to evaluate the criterion-related validity of all types of psychological tests and other personnel selection techniques. However, the results of these analyses should be used with caution. Often they include both operational and “corrected” validity figures. For example, it is common to correct the results of the analyses for artefacts such as criterion unreliability and restriction of range. These corrections can be very useful in evaluating the “true” rather than observed validity of psychological tests and selection measures. However, it should be noted that a decision about whether to use a particular test or selection technique in a specific setting should not ignore the factors that will affect its usefulness in that setting. To take an example, it may be that a meta-analysis carried out on cognitive ability tests and training performance suggests that these tests offer quite accurate prediction of performance if the ability and performance of a broad range of job candidates are evaluated. However, if the tests are used in circumstances in which the variation in cognitive ability between candidates is relatively slight, the resulting restriction of range will mean that the tests are considerably less predictive in this setting, and may even offer no predictiveness at all. In practice therefore it is the operational rather than the corrected or “true” validity which matters, and it is generally useful, if introducing a test for the first time, to examine this operational validity closely.

Incremental Criterion-Related Validity

Incremental validity is concerned with the degree to which a particular test or personnel selection method predicts a criterion (e.g. job or training performance) over and above one or more other selection methods. It is an important form of validity in practice because although a given method or test may predict a criterion such as training performance to a particular extent, there may be little point using it if it adds nothing to already existing and readily available predictors. To take an example, imagine that GCSE and A level results are available, and an organization uses these to select job candidates. They then wonder whether a measure of cognitive ability would also be worthwhile. To examine the effectiveness of this selection technique they might correlate the scores people obtain on the cognitive ability test with the job performance of these people. If a reasonably strong relationship between the test and job performance is found they might conclude that it is worthwhile considering applicant cognitive ability test results in addition to their GCSE and A level results when making appointments. However, this may be a mistake. The critical issue here is not simply whether the ability test predicts performance, but rather whether it predicts performance *over and above* GCSE and A level results. That is, can we predict someone's likely job performance any more accurately when using GCSE, A level, and ability test results than when we use GCSE and A level results alone? The statistical method used to evaluate the degree of this increased or incremental validity is usually multiple regression, and sometimes a particular form of this technique referred to as hierarchical (or sequential) regression. It is important to note that the issue is not simply whether the test in question adds a statistically significant increment in predicting the criterion, but also whether the size of this increase is practically worthwhile.

Other Criteria for Evaluating Selection Methods

In addition to reliability and validity, a variety of other criteria can be used to assess psychometric tests and other selection methods. The most prominent of these are discussed below.

Acceptability

For a selection method to be usable it is necessary that it is viewed as acceptable, not only from the point of view of the organization using it, but also from the perspective of other interested groups such as relevant professional bodies, the legal system, trade unions, and the job candidates themselves. It is essential that the selection method neither harms test candidates, nor violates personal or professional standards, and, on a more positive note, is judged by the candidate to be useful to them. An example of a selection method which may seem useful to a candidate is the personality questionnaire. If a candidate receives feedback on their personality profile as a result of their responses to a personality questionnaire they may view this as useful for personal development.

Usability

However good a selection method may be in relation to the various criteria being discussed here, it is essential that it is usable. A selection test which takes six months for a candidate to complete is, in most circumstances, of little use however good it is in other ways. A usable selection method is one in which those administering it can be trained without too much time or expense, and in which the time and resources required for preparation, administration, analysis, and interpretation are acceptable. What is considered acceptable in one organization may not be considered acceptable in another. Indeed, views about the usability of a particular method might vary within an organization according to the type of people being selected and the circumstances in which the selection is taking place.

Generality

A selection method which can only be used for one specific type of job is clearly less attractive to an organization than one which can be used for jobs of many different types. The more jobs to which a selection method can be applied, the more generality it is said to have. One advantage of a selection method having high generality is that those involved in the selection process need only to be trained in the use of this one method in order to be in a position to select people for a wide variety of jobs.

While for the most part generality is viewed as a good thing, it is more advantageous in some contexts than others. In an organization in which 95 per cent of the workforce have the same role it will not matter so much that a selection method can only be used for this one job role. However, in an organization in which people do many diverse types of job, a selection method with high generality is clearly attractive.

Fairness

The process of selection inevitably involves rejecting some candidates. It is important that this process is achieved fairly, and in this context fairness is usually interpreted as meaning that people are not discriminated against because they happen to belong to a particular social, ethnic, or racial group.

When discrimination occurs it can be deliberate: an employer may, for example, falsely tell an Asian applicant that a job has been taken and then offer it to a white applicant. This is technically known as direct discrimination, and it is contrasted with indirect discrimination. In the case of indirect discrimination, an organization requires applicants to possess some form of ability, qualification, knowledge, skill, or characteristic which (a) is not actually important for the proper performance of the job in question and (b) people in some social groups are more likely to have than others.

So an employer might use physical strength as one of the criteria in the selection of someone to drive a large vehicle. This will disadvantage women because on average women are less physically strong than men, and if physical strength is not really necessary for the performance of the job the organization will have engaged in indirect discrimination.

Adverse impact

It is difficult, if not impossible, to spot adverse impact as it is actually occurring during personnel selection. Consequently adverse impact is identified by looking for its effects on the composition of the people who are selected and rejected. More specifically, the numbers of people from different groups who apply for jobs in an organization are compared with the number from each group who are actually selected. For example, the number of male and female applicants might be compared with the number of males and females actually selected. If the proportion selected from one group is clearly larger than the number selected from another group, this is evidence that the selection process has adverse impact. This can be more clearly explained with an example. Consider the profile of job applicants and candidates selected shown in Table 5.

Table 5

Number of People Applying and Selected for a Job

	White	**Asian**	**Black African**	**Black Caribbean**
Applications	854	75	23	67
Selected	150	7	1	9

It is helpful to express this information as the number of people applying for the job and the number rejected. This is set out in Table 6.

Table 6

The Number of People Selected and Not Selected for a Job

	White	**Asian**	**Black African**	**Black Caribbean**
Not selected	704	68	22	58
Selected	150	7	1	9

To get a better picture of the relationship between the people selected and rejected, these figures can be expressed as percentages as in Table 7.

Table 7

Percentage of People Selected and Not Selected for a Job

	White	**Asian**	**Black African**	**Black Caribbean**
Not selected	82.4%	90.7%	95.7%	86.6%
Selected	17.6%	9.3%	4.3%	13.4%

Table 7 shows that the proportion of people selected and not selected for the job does seem to be related to ethnicity, with the proportion of applicants selected from the three ethnic minority groups being smaller than that for Whites. At first impression this does seem to indicate that the selection system has adverse impact. However, it is possible that these differences have arisen through chance, or that they might be too small to be of consequence.

To examine whether the difference is due to chance a statistical analysis of the figures in the previous table can be carried out. Because the data are categorical and are arranged in a single contingency table, the relevant statistical analyses are chi-square, or the Fisher exact test. Calculating chi-square on the data we obtain chi-square (3) = 6.45, $p > 0.05$, indicating that if there is no underlying relationship between selection and ethnicity (i.e. there is no adverse impact), differences in the proportion of people selected from these ethnic groups as large as those observed here (or even larger) would nevertheless occur by chance 1 time in 20 or more. It is the convention in circumstances in which the observed difference is calculated to be due to chance 1 time in 20 or more to refer to this result as not statistically significant and in this case to accept the hypothesis that there is no underlying relationship between selection and ethnicity. Here therefore, there is no statistically significant relationship between whether or not people are selected and their ethnicity.

However, there are two problems with using a statistical test such as chi-square in these circumstances. First, whether or not a test of statistical significance can identify an underlying association between the number of applicants selected and the social groups of which they are members depends in large part on the statistical power available, and this in turn depends upon the size of the sample of people involved. So if there is a small number of job applicants, researchers are unlikely to find a statistically significant difference in the proportions selected from different groups even if adverse impact is actually taking place. Second, with very large numbers of job applicants researchers will be likely to find a statistically significant association between the proportions of people selected from different social groups even when differences in these proportions are so small as to be considered by many to be too trivial to be concerned with.

The four-fifths rule

To deal with this problem, uniform guidelines in the United States introduced what has become known as the four-fifths rule. This rule is based on the selection ratio, which is the ratio of the number of applicants selected to the total number of applicants. If the selection ratio for a group protected by the law is less than four-fifths of the selection ratio for the majority group, this establishes a ‘presumption of discrimination’.

Table 8 shows the selected/applied ratio for the four groups in the fictitious organization described in Tables 5 to 7.

Table 8

Ratio of People Selected to Those Applying in Each Ethnic Group

	White	Asian	Black African	Black Caribbean
Ratio of applied to selected	0.18	0.09	0.04	0.13

The highest selection ratio is the 0.18 (or 18%) selection ratio for Whites. According to the four-fifths rule, if another group has a ratio less than four-fifths of 0.18 a presumption of discrimination will have been established. Four-fifths of 0.18 is 0.14. As all of the ethnic minority groups have selection ratios less than 0.14, the four-fifths rule indicates a presumption of discrimination for all of these groups in this case. Another way of representing this is with what is called the **adverse impact ratio**. The adverse impact ratio is simply the selection ratio for the minority group divided by the selection ratio for the majority group. So in the case of the data in Table 8, the adverse impact ratio for Asians is 0.09 divided by 0.18, which is 0.5. If the adverse impact ratio is less than 0.8 (four-fifths), there is evidence of adverse impact.

Finally, in addition to the four-fifths rule and the use of statistical significance tests, Morris and Lobsenz (2000) have suggested that confidence intervals represent another useful way of dealing with adverse impact. The adverse impact ratio is, of course, only based on the sample of White and minority people actually selected by an organization. However, when it is claimed that an organization is using a selection method which has adverse impact, the idea is that the use of this method will lead to people from ethnic minorities being systematically disadvantaged. This systematic disadvantage surfaces in the adverse impact ratio. In these circumstances it is possible to examine whether the degree of adverse impact shown in the adverse impact ratio provides sufficient evidence that the selection method is producing adverse impact. Confidence intervals for the adverse impact ratio tell us with a given degree of certainty (usually the 95 per cent level of certainty is chosen) the possible range of values that the ‘true’ adverse impact may have. For example, based on the data in Table 8, the 95 per cent confidence interval for the adverse impact ratio of 0.5 for Asians is 0.24 to 1.04. Because 1.04 is greater than 0.8 (four-fifths), we cannot say with 95 per cent confidence that there is evidence of adverse impact in this data.

Selection methods and differences in sub-group performance

Using meta-analyses carried out by other researchers in North America, Bobko *et al.* (1999) examined the degree to which cognitive ability, structured interviews, biodata, and conscientiousness (the personality factor most strongly associated with job performance) are associated with ethnic sub-group differences (Whites versus Blacks). Table 9 shows the relationships they report between these four selection methods on the one hand and ethnicity on the other. The variable *d* is a measure of the size of the association between ethnicity and scores obtained with each of the selection methods. It is obtained by subtracting the mean score for ‘Blacks’ from the mean for ‘Whites’ and then dividing the result by the standard deviation of the scores obtained by these groups. The greater the *d* value, the greater the extent to which Whites are measured as outperforming Blacks (in units of standard deviation). A *d* of 1.00 indicates that the average score obtained by Whites is on average, one standard deviation greater than that obtained by Blacks. The strength of the relationship between ethnicity and the scores obtained by candidates assessed with each selection technique is also expressed as a correlation coefficient (for example, the correlation between cognitive ability and whether people are Black or White). Also shown, again with a *d* value and a correlation coefficient, is the measured association between ethnicity and job performance.

Table 9 shows that there is very little relationship between ethnicity (Whites versus Blacks) and conscientiousness. There is a relatively small relationship between candidate ethnicity and the scores obtained when biodata or structured interviews are used in selection. However, there is a relatively strong association between ethnicity and measured cognitive ability.

Table 9

The Relationship between Selection Methods, Job Performance and Ethnicity (Whites versus Blacks)

Measure	*d**	Correlation with ethnicity	Meta-analysis
Cognitive ability	1.00	0.37	Hunter & Hunter (1984)
Structured interview	0.23	0.09	Huffcutt & Roth (1998)
Conscientiousness	0.09	0.04	Schmitt, Clause *et al.* (1996)
Biodata	0.33	0.13	Gandy, Dye *et al.* (1994) Pulakos & Schmitt (1996)
Job performance	0.45	0.18	Ford, Kraiger *et al.* (1986)

* Note: Ones and Viswesvaran (1998) report a *d* for Black–White differences of 0.04 for integrity tests, and Schmidt, Ones *et al.* (1992) report a *d* of 0.50 for work sample tests.

Source: based on Bobko et al.*, (1999)*

Differential Item Functioning

One approach to the problem of adverse impact in psychometric testing is to try and identify problematic test items and then replace them with alternative items. The rationale for this approach is as follows. In a test of cognitive ability the probability that a white person will give the correct response to a particular test item (i.e. question) might be the same as the probability that someone from an ethnic minority will give the correct response to that item. However, it might also be the case that for a small number of items, perhaps because of the language used in the question or for other reasons, the probability of answering the item correctly may be greater for a white person of a given level of ability than for someone from an ethnic minority *of that same level of ability*.

It is possible to identify such problematic items using a technique called "differential item functioning", often referred to with the acronym DIF. The application of DIF is best understood with an example. Consider Figure 1 below. This graph shows the probability that candidates will respond correctly to a particular item in a cognitive ability test as a function of how cognitively able they are. The Y axis, on the left, shows varying probabilities that they will respond correctly to the item, from the certainly that they will fail at the bottom (probability of responding correctly is zero), to the certainty that they will respond correctly (probability is 1.) at the top. On the Y axis, that is the horizontal one at the bottom, is shown the ability of candidates, from very low ability on the left to very high ability on the right. The graph line shows that the probability of candidates with "poor" cognitive ability responding correctly to the item is low, that this probability increases sharply around the middle ability level, and then levels off again as we reach the most cognitively able candidates.

The graph can be interpreted as indicating that this item does a good job of differentiating between candidates of lower-middle and upper-middle cognitive ability, but is poor at differentiating between those of low cognitive ability, and also poor at differentiating between people of high cognitive ability. Ideally the test would have items that differentiate well at all ability ranges – that is would have relatively easy items which differentiate well between the least able people, and relatively difficult items which differentiate well between the most able candidates.

Now consider Figure 2. This time a separate line has been plotted for white and ethnic minority candidates. It indicates that ethnic minority candidates of the same cognitive ability as white candidates are *less likely* to respond correctly to the test item. Here it would be concluded that there is evidence of differential item function, or DIF because the item is functioning differently for the white and ethnic minority groups. If this effect is found steps might be taken to replace the item with one which is not problematic in this way.

Finally, consider Figure 3. Here there are two DIF problems with the item. First there is the same problem as in Figure 2 in that ethnic minority candidates of a given ability level are less likely to respond correctly to the item than white candidates. But in addition to this the slope of the graph is clearly different for these two social groups. This means that for which candidates the item differentiates well between those in the middle ability area, whereas for ethnic minority candidates it does not differentiate between people of different ability as well.   DIF analysis is therefore useful for identifying two types of problems with items, and can play an important role in the development of fair psychometric tests, including tests of aptitude.

Figure 1

The Probability of Candidates Responding Correctly to a Particular Psychometric Test Item as a Function of their Cognitive Ability

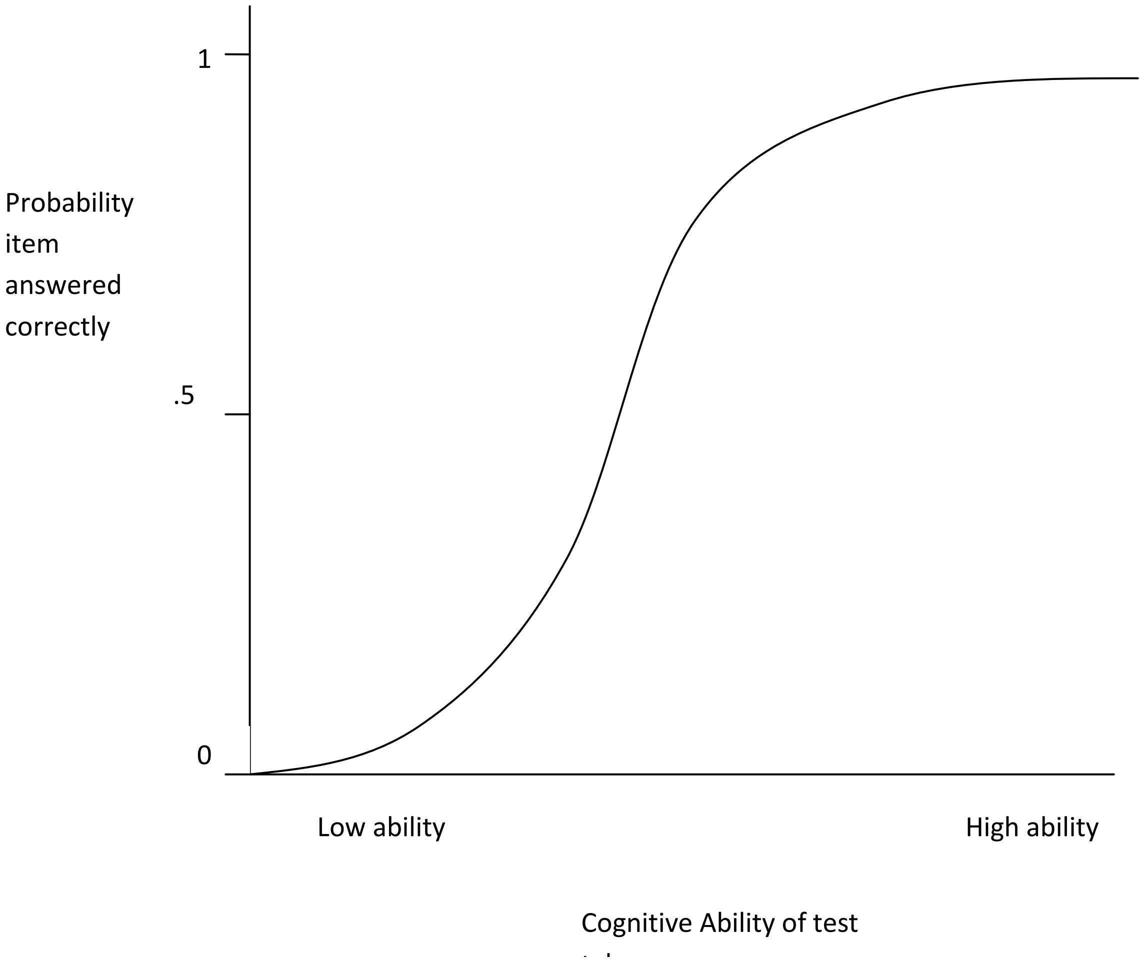


Figure 2

The Probability of White and Ethnic Minority Candidates Responding Correctly to a Particular Psychometric Test Item as a Function of their Cognitive Ability:

A Main Effect for Ethnicity

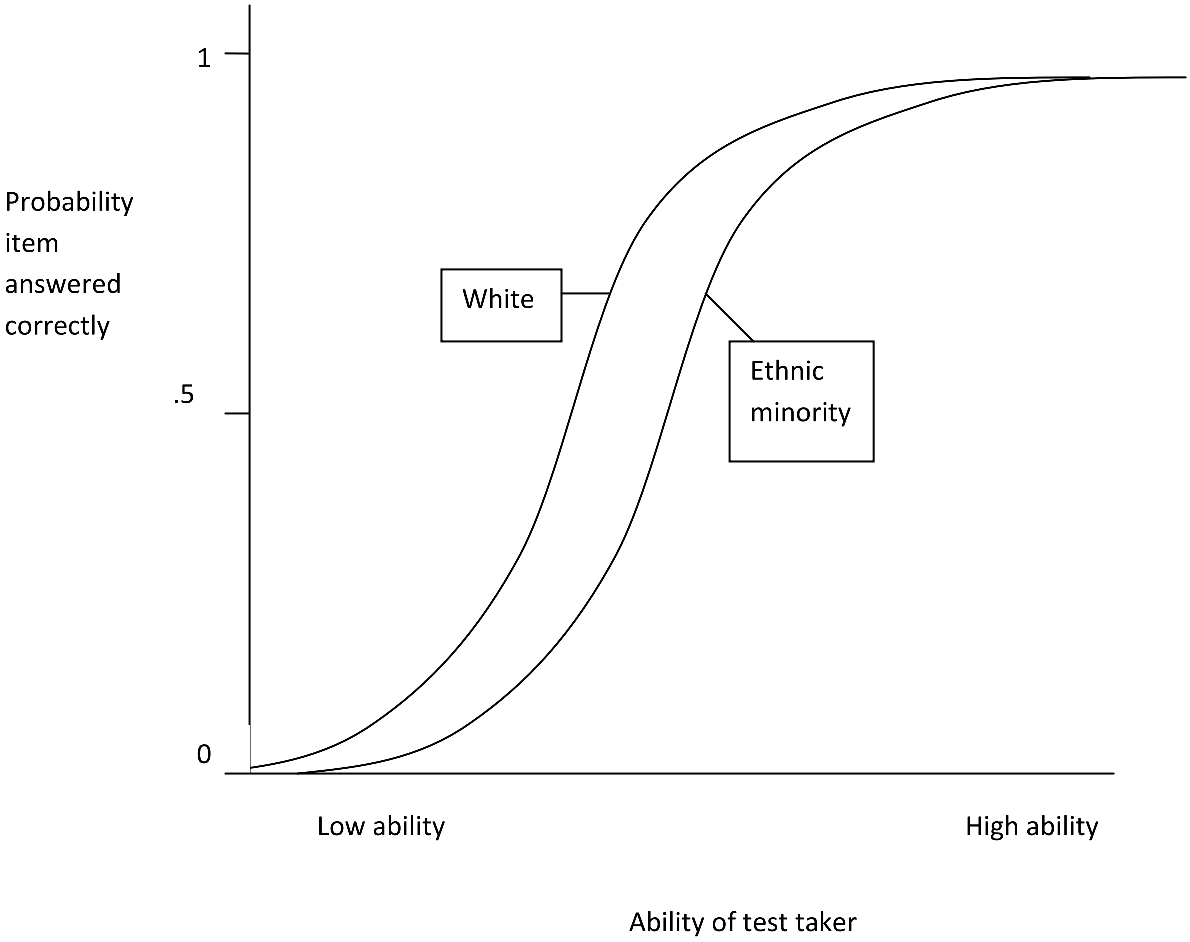
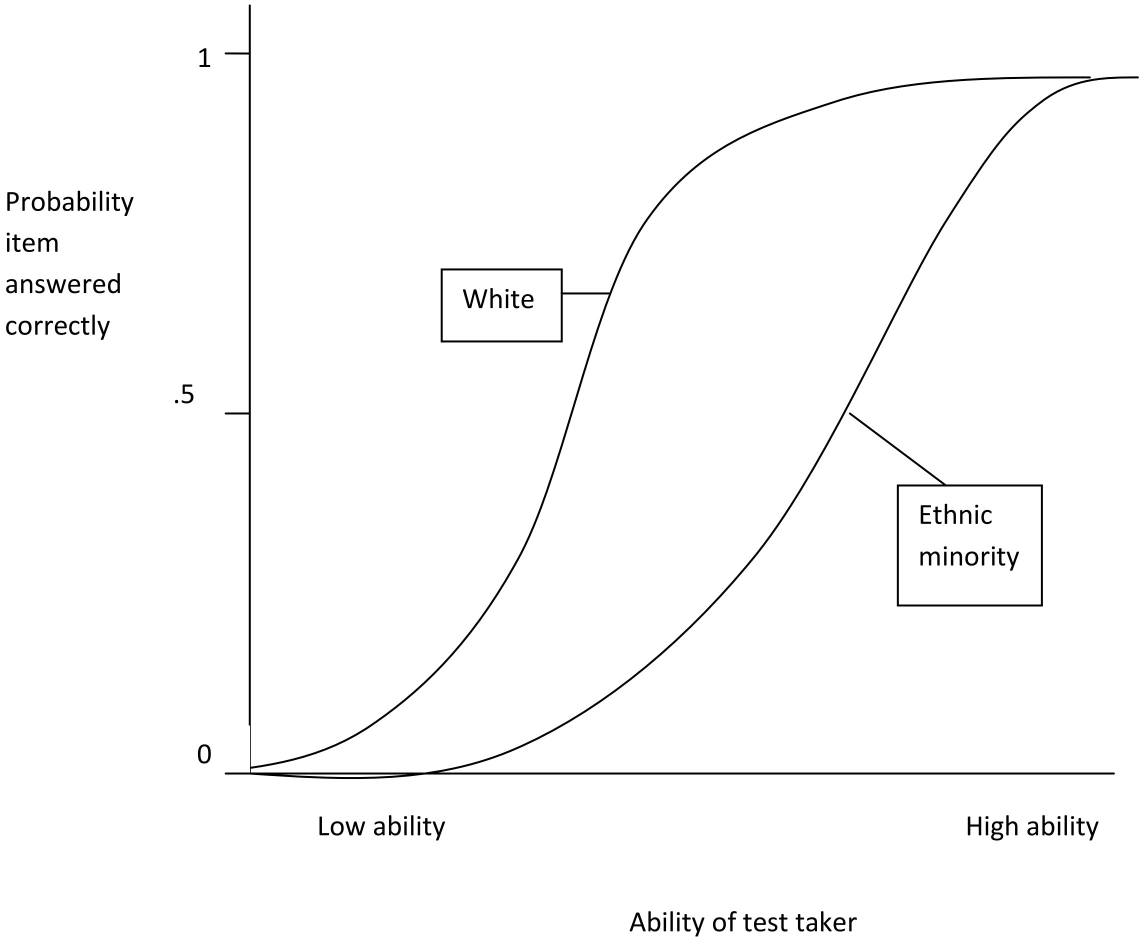


Figure 3

The Probability of White and Ethnic Minority Candidates Responding Correctly to a Particular Psychometric Test Item as a Function of their Cognitive Ability:

An Interaction Effect



Utility

The utility of a selection method is a function of the relationship between its benefits and its costs. These benefits and costs refer to any of the other criteria mentioned so far, including criterion-related validity, fairness, usability, reliability, the time taken to administer the method by personnel staff, the possible cost of consultants to design and oversee the system, the cost of training staff to use the system, the cost of running the system, etc. A method with high utility offers a substantial amount of benefits and a minimum of costs.

Weighing up all the possible benefits and costs is complex and difficult. However, for any organization which spends considerable amounts of financial and other resources on selection, and for which the selection of candidates with the best potential is very important, the relative utility of different selection methods is not a matter to be taken lightly.

Psychologists have developed a technique called **utility analysis** which is designed to make the process of evaluating the utility of different techniques more objective, and to express the relationship between benefits and costs financially. Given that financial arguments often carry substantial weight in organizations, utility analysis can give human resource personnel valuable ammunition in arguments about the type of selection method in which their organization should invest.

Aptitude Tests Currently Used in the Professional Service Sector

In this section the nature of some aptitude tests currently in use in the UK and North America will be described and wherever possible information on the reliability and validity of these tests provided. The section is divided into two parts, the first deals with aptitude tests in the UK and Ireland, and the second with aptitude tests in North America.

Aptitude Tests in the UK and Ireland

If aptitude or ability tests are to be used to select people for the UK legal profession it would be helpful to examine the existing use of such tests in other professional services in the UK. As discussed earlier, the BMAT and UKCAT aptitude tests are currently used for selecting medical students, but tests are not used on a large scale for university entrance to other professions. As a consequence it is difficult to ascertain how widely tests of aptitude tests, which professions, if any, outside medicine use them, and what their criterion-related validity is in these settings.

To address this issue a sample of UK test publishers were contacted on the telephone by my assistant and asked which professional service organizations they provided tests to, and what validity studies had been carried out in relation to the use of tests in these professional services. The sample of test publishers was obtained from a list of such publishers made available by the Psychological Testing Centre of the British Psychological Society (BPS). The Psychological Testing Centre lists 41 test publishers including subsidiaries, most based in the UK (see Appendix 3), and 20 of these were contacted by telephone. The contacted publishers are listed below.

Caliper UK

CDA Profile

GL Assessment

Hogrefe

Human Factors

JCA

Knight Chapman

Occupational Research

OPP

Pearson Assessment

Pearson Vue

Psychological Consultancy

Psytech

Quest Partnership

Saville Consulting

SHL Group

Talent Q

TalentLens

The Morrisby Organization

Thomas International

Each of these organizations has developed their own proprietary tests and/or supply tests that have been developed elsewhere. None of the organizations contacted claimed to provide an aptitude/ability test purposely designed for specific professions. Typically the norms for the tests used were said to be based on the general population or on particular occupational groupings such as graduate or managerial. Due to what was usually described as “commercial sensitivity” or “confidentially” representatives of the 20 contacted publishers were not willing or able to specify the organizations who used their tests.

A number of the representatives of test publishing organizations contacted by telephone said that they believed that organizations using their tests were generally unwilling to fund studies to investigate their validity. Despite this, several of these representatives indicated that the test publisher for whom they worked had carried out validation studies on aptitude or ability tests for particular occupational groups. However, in all but one case the results of these studies were described as “commercially sensitive” or words to that affect, and for this reason the publisher declined to make the results available. In one case a publisher indicated that a ‘validation’ had been carried out for a test used for solicitor selection (albeit on a small sample of 41) and indicated that the study could be made available. Unfortunately the publisher concerned was eventually unable to find the report on this 10 year old study.

The attempt to obtain information about the validity of ability and aptitude tests in the UK by approaching test publishers for relevant data was therefore unproductive. However, an investigation of some other organizations who are likely to develop and/or use psychological tests revealed that they had conducted some validation studies. Telephone conversations with representatives of the of the Civil Service and the Military confirmed that such validation had taken place, and that there is an ongoing validation study at the National Police Improvement Agency (NPIA). In addition, validation studies involving aptitude/ability tests for selection to professional roles are taking place in the Irish Aviation Authority, NATS Ltd, and the National Recruitment Office for General Practice training. Some work has also been undertaken on the validation of tests used in the UK for the selection of medical students: BMAT and UK CAT. These tests and information on their validity, where available, are described below. Also described are other aptitude tests currently used in the UK and Ireland.

LNAT

LNAT, or the “National Admissions Test for Law” is the only aptitude test currently used in the UK for the selection of people to the legal profession. It was established by a consortium of Universities and is operated by LNAT Consortium Ltd in partnership with Pearson VUE. The consortium is comprised of seven Universities: Birmingham, Bristol, Durham, Nottingham and Oxford, King's College London, and University College London. The test was developed Edexcel, an organization owned by Pearson.

The test consists of a multiple choice test and a written essay and is designed to measure the following verbal reasoning skills:

- comprehension
- interpretation
- analysis
- synthesis
- induction
- deduction

The test is use by participating UK law schools to aid in the selection of undergraduate law students.

There is currently no published research on the LNAT in peer reviewed journals, though some data is available on it at the LNAT website www.lnat.ac.uk. Some of this data was discussed earlier in this report in the section headed *Adverse Impact*.

Irish Aviation Authority

The Irish Aviation Authority uses aptitude tests in the process of the selection of air traffic controllers. To a large extent the selection process draws on FEAST selection tests (see below under NATS). There are eight stages in its selection process for Student Air Traffic Controllers which include:

Written tests:

Applicants who meet the minimum entry requirements are called to a paper and pencil aptitude test. Progression through the selection process is subject to verification of the minimum entry requirements.

Computer Based Aptitude Tests

Candidates who attain a high enough position in the paper and pencil aptitude test are called to a computer based aptitude test. The test battery consists of a set of tests which examine a candidate's ability in regard to a number of items including the following:

- Heading and Range Test
- English Listening and Comprehension
- Planning Ability
- Sort Task

- Alertness in Simple and Multi-Tasking Situations Test
- Visualisation Test

NATS

NATS Ltd. (formerly National Air Traffic Services Ltd.) is the main air navigation training supplier in the UK. There is a three-day assessment of candidates for the role of Air Traffic Controller. The focus in these notes is the assessment over the first two days as these involve the use of aptitude/ability tests. The third day involves group exercises, a personality questionnaire and a competency based interview.

Assessment Day One

Assessment day one is a half-day session consisting of a number of assessments designed to show if applicants have the key aptitudes and attributes of a successful Air Traffic Controller. One of these tests is the Air Traffic Control training sample test. This tests the knowledge and understanding of materials relevant to Air Traffic Control training.

The numerical test involves speed, time and distance calculations and compass directions. Candidates need to calculate speed (distance/time), time (distance/speed) and distance (speed x time). They also need to calculate directions and headings using the 360◦ of the compass. If applicants are successful in these initial stages they complete two aptitude tests; a spatial reasoning test, and a diagrammatic reasoning test.

Assessment Day Two

Assessment day two is a whole day session and involves a series of computer-based tests which provide us with more information on some of the key attributes and aptitudes of successful Air Traffic Controllers. These assessments include the First European Air Traffic Control Selection Test Battery (FEAST) and also some NATS assessments.

FEAST includes a number of sub-tests, all of which were specially developed to assess applicants for Air Traffic Control (ATC) training. The NATS computer assessments are a range of tests assessing abilities considered important for success in the role. These tests are designed to measure:

- Visualisation
- Sorting ability
- Planning
- Multi-tasking alertness
- English listening and comprehension

While it has not been possible to gain access to any of the studies my assistant was informed that the FEAST test package was developed and tested during 2000-2003. During 2004-2005, a pilot FEAST service, including delivery of the testing system and a database for storage of results and data, was the subject of a systematic evaluation. Apparently, there are now nearly 17,000 candidates in the FEAST database. In 2009 a predictive validation study was carried on FEAST but the results are unknown.

BioMedical Admissions Test (BMAT)

BMAT is used in the selection of students for certain medical and veterinary medical courses within five UK medical schools: the University of Oxford, the University of Cambridge, University College London, the Royal Veterinary College, Imperial College London, and the University of Bristol. It consists of two half-hour multiple-choice tests. The first, “Aptitudes and Skills”, is designed to evaluate a candidate’s skills in problem solving, understanding arguments, and data analysis and inference. The other “Scientific Knowledge and Application” is concerned with a candidate’s ability to apply scientific knowledge normally encountered in non-specialist school science and mathematics courses. There is also a one-hour written exercise.

The predictive validity of BMAT is discussed in several articles (Emery & Bell, 2009, 2011; Emery, Bell, & Rodeiro, 2011; McManus, Ferguson, Wakeford, Powis, & James, 2011), and the most thorough investigation is reported by Emery, Bell and Rodiero (2011). Emery and her colleagues examined the extent to which BMAT predicted the first year medical examination results obtained by 588 medical students at the University of Cambridge Medical School. They found that whilst the test of Scientific Knowledge and Application was a statistically significant predictor of first year examination performance, the test of Aptitudes and Skills were not.

United Kingdom Clinical Aptitude Test (UK CAT)

UKCAT is an aptitude test assessing verbal reasoning, abstract reasoning, quantitative reasoning, and decision analysis. It was introduced in 2006, and is currently used in the selection of medical and dental students by 26 medical and dental schools in the UK. In 2009 over 23,000 candidates completed the test online at remote testing centres operated by Pearson Vue. According to the UK CAT website www.ukcat.ac.uk the consortium operating the test is "committed to achieving the greater fairness in selection to medicine and dentistry and to the widening participation in medical and dental training of under-represented social groups".

Recently the predictive validity of UKCAT has been investigated in three studies. Lynch et al. (2009) examined the correlation between the UK CAT scores and first year examination results of 341 medical students at the University of Aberdeen and the University of Dundee. They found no statistically significant correlations between the examination results on the one hand and either the students' overall UK CAT score or any of the three subtest scores on the other. Nor did the UK CAT overall score or subscores predict withdrawals from the course. Yates and James (2010) examined the extent to which UK CAT scores predicted performance in the first two years of training at Nottingham University medical school. A sample of 204 students were included in the study, and the authors concluded that whilst two of the subtests were predictive of marks in two examinations, the total UK CAT score had little predictive value. However, Wright and Bradley (2010) examined the relationship between the UK CAT scores and examination results of 307 medical students at the University of Newcastle and found that the test was a statistically significant predictor of results in all but one of the knowledge-based examinations taken in the first two years of medical school training.

Evidence of the predictive validity of the UK CAT aptitude test in predicting the behaviour and performance of medical and dental students is therefore mixed. Given the large scale the UK CAT Consortium recently agreed to fund a scoping exercise to examine the feasibility of a large-scale longitudinal study into the effectiveness of this aptitude test. I will be conducing this exercise with Professor Chris Mc Manus of University College London, and it is anticipated that the initial findings from this study will be reported in the second half of 2011.

UK Aptitude Tests, Practice, and Coaching

The introduction of aptitude tests for medical school selection in the UK has been followed by the publication of several books designed to help prospective test candidates develop their skills in test-taking. These include *How to Master the UKCAT: Over 700 Practice Questions for the United Kingdom Clinical Aptitude Test* (Bryon, Clayden, & Tyreman, 2010), *How to Master the BMAT: Unbeatable Preparation for Success in the BioMedical Admissions Test* (Tyreman, 2009), and several others (Butterworth & Thwaites, 2010; Emery, et al., 2011; Green & Hawley, 2009; Hutton, Hutton, & Taylor, 2010; Picard, 2009). In addition, an organization called *Kaplan Test Prep and Admissions* offers classroom-based and online courses and private tuition in BMAT, UK CAT, and a range of other aptitude tests at prices currently ranging from £250 to £1,250. Another organization, Cataga, offers a book on designed to help candidates achieve good LNAT scores called *Ace the LNAT,* and also provides coaching on this aptitude test.

Given the meta-analytic evidence that a combination of test practice and test coaching increases candidate performance on aptitude tests such as the SAT by about .76 of a standard deviation (Bangertdrowns, et al., 1983a, 1983b; Kulik, et al., 1984), the availability of such books and coaching opportunities is worthy of note.

Clinical Problem Solving and Situational Judgement Tests

In recent years automated (machine-marked) tests have been introduced in the selection of General Practitioners (GPs') in the UK where they are used for the short-listing stage. Whilst the selection methods do not take the form of aptitude tests, they are sufficiently similar to warrant mention here. The first test is concerned with clinical problem solving (CPS). Here candidates are required to apply clinical knowledge in order to "solve problems reflecting diagnostic processes or develop management strategies for patients". (Patterson et al., 2009, p417). The second is a Situational Judgment Test (see Appendix 1) in which a variety of professional dilemmas that GPs may encounter at work are presented, and candidates are required to select an appropriative response from several alternatives. Preliminary validation of these techniques indicates that they are good predictors of the rated interview performance of candidates (*r* = .61 for the CPS and SJT combined) (Patterson, et al., 2009) though the extent to which they predict future job performance is currently unknown.

Aptitude Tests in North America

As explained in previous sections, aptitude testing is carried out on a very large scale in the United States. An exhaustive description of these tests and the reliability and validity data on them is beyond the scope of this report. However, details on three of the most widely used tests, the SAT, MCAT, and LSAT (the test used by almost all US law schools), is given below. A brief description of some additional North American tests, including ACT, ASVAB, and GATB are provided in Appendix 2.

SAT

People who wish to study for undergraduate degrees in the United States usually have to undertake one of two aptitude tests, the SAT or the American College Test (ACT).

The latest version of the SAT was released in 2005 and is made up of two components: the SAT Reasoning Test and the SAT Subject Tests. The Reasoning Test is used for general admissions to college whereas the Subject Tests are for specialist advanced areas of study.

The Reasoning Tests covers three areas: critical reading (extended reasoning, literal comprehension, and vocabulary in context), mathematics (numbers and operations, algebra and functions, geometry and measurement, data analysis, statistics, and probability), and writing (an essay, improving sentences, identifying sentence errors, and improving paragraphs).

The reliability of the SAT is very high. It has an internal consistency of .91 to .93, and with a small number of exceptions test-retest coefficients of between .87 and .89 (Gregory, 2010, p234). A great deal of validity information is available on the SAT and the latest studies can be found on the website of the College Entrance Examination Board at www.collegeboard.org. Donlon (1984, chap. 8) summarises data on the tests, and says that the SAT Verbal and Math scores correlated .42 on average with first year grade point average. The continuous work done to develop and validate the tests is impressive. The validity of the tests in predicting college performance is impressive also, and the test provides incremental validity over the grade point average students obtain in high school.

Medical College Admissions Test (MCAT)

The MCAT has to be taken by applicants to almost all medical schools in the United States. It consists of three multiple-choice sections (Verbal Reasoning, Physical Sciences, and Biological Sciences) and an essay (Writing Sample).

The split-half reliability of the component tests are acceptable and in the low .80s (Gregory, 1994). Julian (2005) examined the predictive validity of the MCAT in predicting medical school grade point average (GPA) during the first three years of medical school. A total of 14 medical schools were examined in 1992 and 1993. These schools were selected to be representative geographically, racially, and ethnically of United States medical schools. The outcome variables considered were (a) combined year 1 and year 2 GPA, and (b) year 3 GPA. The second outcome variable was performance at the United States Medical Licensing Examinations. The data considered here, which was obtained from all United States medical schools and hence involved a much larger sample of over 25,000 candidates, were from each of the three steps of these examinations. Step 1 is usually taken after the second year of training and is concerned with the understanding and application of basic science relevant to medical education, Step 2 is taken at the end of medical school and assesses clinical skills and knowledge and the extent to which these skills and knowledge can be applied under supervision, and Step 3 is taken a year after residency and is used to assess the degree to which graduates can independently apply their knowledge and skills. The predictive validity of the MCAT reported by Julian is shown in Table 10.

Table 10

MCAT Validity Coefficients for the 1992 and 1993 Cohorts

Reported by Julian (1995)

	Outcome variables predicted				
	GPA at Medical School		United States Medical Licensing Examinations		
	Years 1 and 2 combined	Year 3	Step 1	Step 2	Step 3
Sample size	4,706	4,706	27,406	26,752	25,170
Median validity coefficient	.59 (.44)	.46 (.32)	.70 (.61)	.60 (.49)	.62 (.49)
Range of medical school validity coefficients	.38 - .78 (.20 - .70)	.31 - .54 (.18 - .41)			

In Table 10 two types of validity coefficients are shown, first the validity corrected for restriction of range and second, in parentheses, the observed validity. These figures are impressive, and indicate that the MCAT is a strong predictor of performance in medical examinations, not only during initial medical training but also after graduation.

Law Schools Admission Test (LSAT)

Almost all applicants to law schools in the United States are required to take the Law Schools Admissions Test. The test, which was introduced in its original form in 1948, uses a multiple-choice format, and focuses on four areas: reading comprehension, analytical reasoning, and logical reasoning which is presented in two sections. In addition there is a thirty-minute writing test. The writing test is not scored but is sent to the laws schools to which the applicant applies.

The LSAT is administered by the Law Schools Admissions Council, and they provide regular reports on the validity of the tests to all law schools providing data. For example, in 2007 and 2008 a total of 165 law schools provided data on the degree to which LSAT predicted first year examination results. A report by Stilwell, Dalessandro and Reese (2009) reports a median validity coefficient across law schools of .32 for 2007 and .33 for 2008. When LSAT results are combined with undergraduate grade point average (UGPA) in a multiple regression model these coefficients rise to .46 for these two years. The median validity coefficient of UGPA alone was .28 in both 2007 and 2008, and if the two are combined using optimum weights derived from a regression model the LSAT therefore clearly makes a worthwhile contribution to the prediction of examination success in the first year of law school. In practice, law school admissions in the United States are influenced in large part by UGPA and LSAT scores (Schultz & Zedeck, 2008).

However, the LSAT has been criticised for failing to address the disproportionately low number of Black students attending law schools in the United States. Simien (1986) points out that in the 1980 US census 11.7% of people in the USA were Black, but only 2.7% were engaged in practice of law, and argues that the failure of the United States to increase the proportion of Black people working the legal profession largely to the use of the LSAT.

The Development of a New Selection Test for Law Students

Drawing attention to a number of weaknesses in the LSAT, Schultz and Zedeck have initiated the development of an alternative and more comprehensive test for selection to law school in the United States. The weaknesses in the LSAT to which they draw attention include its overreliance on the measurement of cognitive ability, the tendency for the test to reinforce racial and social class privileges because people from affluent white backgrounds on average obtain substantially higher scores on the test than their black people from poor backgrounds, and the almost exclusive focus on predicting first year grades at law school when validating the test (Schultz & Zedeck, 2008).

In the first phase of test development, Schultz and Zedeck carried out interviews and focus group discussions with a variety of people working in or connected with the legal profession – lawyers, law faculty, law students, judges and clients in order to establish the competencies required for effective work in the profession. From this 26 indicators of lawyer effectiveness were derived.   Some of these refer to competencies that are useful in many occupations (e.g. “problem solving”) and other which are specific to the legal profession (e.g. “researching the law”). In the second phase of the test development a number of tests were developed and several already existing “off the shelf” tests were identified also. In the third phase the predictive validity of the battery of tests were examined using a sample of 1,148 practising lawyers.

The results of the analysis indicated that there were few substantial gender or ethnic grouping differences in scores on the predictors, that the new predictors showed a degree of independence from UGPA and LSAT scores, and that whilst the new predictors did not show any incremental validity over UGPA and LSAT in predicting first year grades at law school, some of them showed a useful level of predictiveness of the rated job performance of the lawyers, whereas neither UGPA not LSAT did so (Schultz & Zedeck, 2008).

Conclusions and Recommendations

Compared to many other methods and techniques which can be used to select people for educational programmes, training courses, and jobs, cognitive ability tests and aptitude tests have a number of advantages. They can be taken quickly, online, and at remote locations, with large numbers of candidates assessed simultaneously. Once developed, cognitive ability and aptitude tests are relatively cheap to administer. They are objective, and candidate scores can be automatically computed and stored on remote servers via the Internet. Candidates can be ranked according to their performance on these aptitude tests. In addition, the results obtained by candidates on aptitude tests can be combined arithmetically with their results they obtain from other selection techniques, and an overall (weighted or unweighted) candidate performance score obtained.   There is evidence that both tests of general ability and tests of aptitude can yield useful amounts of predictive validity, indeed cognitive ability tests are amongst the techniques with the highest overall predictive validity in relation to both training performance and job performance.

However, there are also a number of risks and dangers associated with the use of ability and aptitude tests. First, the results from the tests can be construed as more precise and accurate than they actually are. Cognitive ability and aptitude tests, despite the scientific methods and techniques used in their development and scoring, do not measure either aptitude or cognitive ability exactly and precisely - both are unreliable to a degree and therefore have a margin of error. Second, aptitude tests may be incorrectly perceived to measure a variety of independent psychological characteristics which can be matched to the demands of a given educational programme, training course, or job. In fact, aptitude tests measure a combination of general cognitive ability, or *g*, and performance on one or more tests of attainment such as knowledge in a specific domain (themselves likely to be associated with *g*). The factors and dimensions measured in aptitude tests may be *presented* as if they are independent by those publishing aptitude tests, but they are almost certainly not independent – largely because to a greater or lesser degree they are actually measuring *g*. Third, aptitude tests may be incorrectly interpreted as measuring the extent to which someone has an innate suitability for a particular educational programme, training course, or job. In fact, partly because aptitude tests are measuring *g,* peoples' scores on measures of aptitude are influenced by the environment they have experienced (including family background and educational opportunities and experiences) as well as by innate potential. For this reason, people from privileged educational, family, and cultural backgrounds will tend to do better than others on aptitude tests. Fourth, because differences in environment are systematically associated with large social groupings, including social classes and ethnic groups, it is likely that there will be systematic differences in the average performance of people from these groups on aptitude tests. Fifth, there is evidence that practice and coaching on aptitude tests can increase peoples' scores significantly, and therefore unless practice and coaching is prevented, or is utilised by everyone taking a test, they will introduce bias and unfairness into the test results. Lastly, it is important to note that aptitude tests do not typically measure all of the characteristics necessary for high levels of educational, training, and job performance. An example of a variable known to influence performance over and above ability which is not usually considered in aptitude tests is the personality variable of conscientiousness.

In developing and evaluating aptitude tests designed for us in the UK legal profession a variety of considerations are worthy of close attention. In particular it is recommended that the following are considered carefully.

Recommended Criteria for Evaluating Ability or Aptitude Tests in the Legal Profession

1. The purpose of the test should be clarified. Is the test intended to predict performance in the long or the short term, in an initial training course or legal career, or both? Clarifying the purpose of the test in this way will be helpful in test development. It will also inform the process for validating the test.

2. Assuming that the test is not simply designed to measure *g*, what is the evidence that the test has sufficient content validity? In establishing content validity, to what extent has job analysis or a related technique successfully identified all of the psychological characteristics and behaviours required to perform at a high standard in the educational programme, training course, or job in question?

3. Has sufficient attention been paid to the range of different techniques that might be used in testing, including situational judgement tests, and personality questionnaires?

4. If a test is designed to measure specific psychological constructs, there should be evidence that it has acceptable construct validity.

5. What are the internal and test-retest reliability coefficients of the test, how large are the samples on which these reliabilities have been estimated, and in the case of test-retest reliability what is the confidence interval for this coefficient?

6. Careful consideration should be given to the criterion to be used in establishing the criterion-related validity of the test, bearing in mind the purpose of the test (see point 1 above).

7. What is the criterion-related validity of the test, how large is the sample that this validity is based on, and what is the confidence interval for this validity estimate?

8. What is the incremental validity of the test over and above alternative information available on candidates from other potentially predictive

variables such as GCSE, A level, and undergraduate degree results? For information about predictors of performance on the Bar Professional Training Course (formerly the Bar Vocational Course) and related courses it may be helpful to refer to research by Dewberry (2001).

9. To what extent do systematic sub-group differences on the test (and sub-components of the test) exist in relation to social class, gender, and ethnicity? In the event that sub-group differences are apparent, to what extent has differential item functioning (DIF) been used to minimize these effects? Is the predictive validity of the test relatively equal in relation to all sub-groups? What is the relationship between the size of sub-group differences on the test and the size of sub-group differences with respect to other predictor variables (such as GCSE, A level, and undergraduate degree results)? What are the consequences of including and excluding the aptitude test results in relation to various sub-group selection ratios?

10. Are all candidates given access to a sufficient number and range of practice test opportunities? Are the practice tests available sufficient in relation to availability, length, clarity, and quality? Are all candidates aware of these practice tests and coaching opportunities, and are they all able to make use of them?

11. How will the results of the test be combined with other information about the candidates in order to arrive at selection decisions, and what is the evidence that this is the optimum process for combining information? It is worthy of note that research on assessment centres in personnel selection suggests that combining candidate-related predictor information arithmetically is more effective than doing so through discussion (Dewberry, In press).

12. Are regular reports on the reliability, validity, and sub-group differences of the test to be published? If so, what steps will be taken in the light of this information?

Appendix 1: Personnel Selection Methods

In addition to tests of aptitude and general cognitive ability a variety of other methods for selecting candidates for training and jobs are available. The most widely used of these are listed below. Three techniques already well known to the reader - the application form, the reference, and the curriculum vitae (CV) - have been omitted.

Achievement tests

Achievement tests are generally designed to examine the extent to which someone has mastered the knowledge and skills in a particular area of expertise. In an educational context GCSE, A level, and undergraduate degree results are all examples of achievement tests. In the context of musical skill the Associated Boards of the Royal School of Music's (ABRSM's) assessment of piano playing competence are also examples of achievement tests.

Assessment centres

In an assessment centre, candidates are located in a single place (such as a hotel or training centre) for an extended period of time (usually one or two days). During this period they are exposed to a battery of selection methods (such as different types of work simulation, often including a leaderless group discussions and an in-tray exercise; one or more interviews; and psychological tests). In the case of the work simulations, referred to as "exercises" trained assessors monitor the performance of candidates score their performance with respect to several pre-determined competency dimensions. The data on each candidate derived from the various exercises and tests are then integrated, and the candidate is selected if their performance exceeds a pre-agreed threshold. Data integration is achieved either using a mechanical method such as summing a candidates (unweighted or weighted) scores across exercises or, more commonly, though a chaired consensus discussion involving the assessors who observed the candidate.

Biographical data

Biographical data (or biodata) consist of detailed information about the life experiences and background of job applicants. The use of biodata is based on the assumption that whether or not certain events and experiences have occurred in the lives of applicants are sometimes associated with whether or not they perform well at work. For example, if someone was given special responsibility at school (such as organizing a fund-raising event), this may help to predict whether or not they will perform well at work. There are various ways to collect this information including structured interviews, the coding of essays written by applicants about their life history, and, most commonly, self-report questionnaires.

Integrity tests

These are standardized tests designed to identify people who are likely to be dishonest or unscrupulous at work, and who may also be undependable and frequently absent from work. Integrity tests come in two forms: overt and personality-based. Overt tests have a section dealing with attitudes towards theft and dishonest behaviour, and a section in which the respondent indicates which dishonest activities he or she has engaged in the past. Personality-based integrity tests are more subtle, and obtain estimates of integrity by using responses to one or more relevant personality scales.

Personality questionnaires

Personality questionnaires (or inventories) generally consist of a series of multiple-choice questions. The candidate is asked about their attitudes, preferences, behaviours etc., and responses are scored in such a way that they indicate the person's position relative to others on specific personality dimensions such as extroversion and conscientiousness. Within the context of job selection, the assumption is that some personalities may be better suited to a job than others.

Job tryout

Candidates are tried in the job for a limited amount of time and their work performance is assessed during this period.

Job knowledge tests

Job knowledge tests involve an assessment of how much job-relevant knowledge an applicant has. This information is generally obtained by requiring the applicant to respond to questions on a form, and the answers may be presented in a multiple-choice format.

Selection interviews

Interviews can be categorized in various ways such as by the extent to which they are structured (structured, semi-structured, and unstructured), and who carries them out (for example, a single interviewer or an interview panel). Interviews are often given to those people who have been 'screened-in' at an earlier stage in the selection process (such as when application forms are sifted). Sometimes multiple interviews are used.

Situational judgement tests

In situational judgments tests (SJTs) candidates are presented with a variety of situations they might expect to encounter in a particular job role. This may take the form of a written description of the situation, or a video. They are then presented with several alternative courses of action and are required to select one or more appropriate of inappropriate responses to the situation they are faced with. For example, they might be required to select the most appropriate course of action, the most and least appropriate course of action, or to rank order the alternative from the most to the least appropriate. The instructions may ask respondents to select the action they think they should (or should not) engage in, or which they would (or would not) engage in. The tests are scored by adding together the number of correct responses.

Weighted application blanks

Weighted application blanks (or WABs), represent an extension of the standard application form. Weighted application blanks essentially consist of an application form for which a scoring key has been developed systematically by means of a research study. Information from this study is used to decide on the points that should be awarded for different answers to various questions on the form. For example, a certain number of points may be allocated to an applicant because he or she has a degree. People who submit the application forms which receive the highest scores are more likely to be accepted.

Work samples

Work samples involve the applicants performing certain aspects of the roles they will be expected to do in the job (such as a test of typing speed for clerical workers) during the selection process. One or more aspects of the work involved in a job may be simulated, and the performance of candidates is assessed on the simulation exercise. Those candidates who perform best are most likely to be selected.

Appendix 2: A Brief Introduction to Factor Analysis

Factor analysis is used when entities (usually people) are measured on several continuous variables and researchers wishes to know whether these variables can be reduced to a smaller set of latent or hidden variables, how much variation in all the data is accounted for by this smaller set, and what the nature of the smaller set of variables is.

Background

The classic uses of factor analysis in psychology have been in the study of intelligence and of personality. In the case of intelligence, psychologists were interested in whether human ability can be reduced to just one factor – *g*, or alternatively whether several independent types of ability exist. If ability is dominated by just one general factor, there would be a strong tendency for people who are relatively good at some indicators of ability (such as tests of numerical ability) to be good at most or all others (such as verbal and spatial ability). Here human ability is represented with a single construct, and someone who is relatively good at one thing will be probably relatively good at all others as well. However, if human ability is made up of several factors, people who are good at tests of numerical ability would not necessarily also be good at tests of verbal ability, or spatial ability. As a result it would be appropriate to conclude that ability is made up of several factors, and that to measure someone's overall ability, you would need to assess them separately on each factor.

Factor analysis is divided into two types: exploratory and confirmatory. Exploratory factor analysis, as its name suggests is used to explore the factor structure in a set of data. Confirmatory factor analysis is used to confirm or disconfirm the presence of a specific factor structure that is postulated to exist before data are collected. Although the mathematics underlying exploratory factors analyses are complex, with modern computer packages it is relatively straightforward to carry out this procedure. By comparison, confirmatory factor analysis is a considerably more complicated to undertake.

The first stage in a factor analysis is to obtain the correlations between all of the variables to be used in the analysis. These are set out in a “correlation matrix” such as that shown in Table 11. The correlation matrix in Table 11 shows (fictitious) correlations between the results of four ability tests: sequencing numbers, adding up numbers, completing sentences, and verbal comprehension.

Table 11

Fictitious Correlations Between the Ability to Sequence Numbers, Add Up Numbers, Complete Sentences, and Comprehend Verbal Information

	Sequencing numbers	**Adding up numbers**	**Completing sentences**	**Verbal comprehension**
Sequencing numbers	–	0.88	0.05	–0.11
Adding up numbers		–	0.13	0.03
Completing sentences			–	0.74
Verbal comprehension				–

This correlation matrix presented in Table 11 shows that the fictitious correlation between peoples’ scores on a test of sequencing numbers, and a test in which they are asked to add up numbers is 0.88, the correlation between sequencing numbers and verbal comprehension is 0.05, and so on.

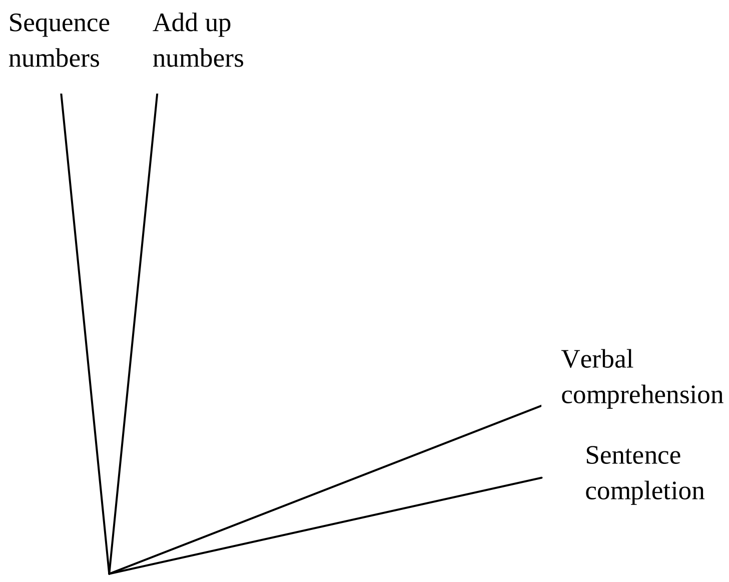
Factor extraction

The second stage of a factor analysis is referred to as “factor extraction”, and it is best explained using some simple geometry.   It is possible to represent the degree of correlation between any two variables as an angle. When two variables are not correlated at all (when the correlation coefficient is .00) they are represented as having an angle of 90 degrees to each other. As the degree to which they are positively correlated increases so the angle decreases – and in the extreme case of two variables which are perfectly positively correlated there would be an exact overlap between them. When the correlation between two variables is negative, the angle is greater than 90 degrees, and in the case of a perfect negative correlation it is 180 degrees.

If the correlations between the four variables shown in the correlation matrix above as angles, the strong correlation between sequencing numbers and adding up numbers will mean that the relationship between them is represented with a narrow angle, and similarly the strong correlation between the test of sentence completion and the test of verbal comprehension will result in a narrow angle also. However, the angles between sequencing numbers and adding up on the one hand and sentence completion and verbal comprehension on the other would be quite large because each pair is weakly correlated with the other pair. Such a geometrical relationship between the four tests is shown in Figure 4.

Figure 4

A Geometrical Representation of Fictitious Correlations Between the Ability to Sequence Numbers, Add Up numbers, Complete Sentences, and Comprehend Verbal Information



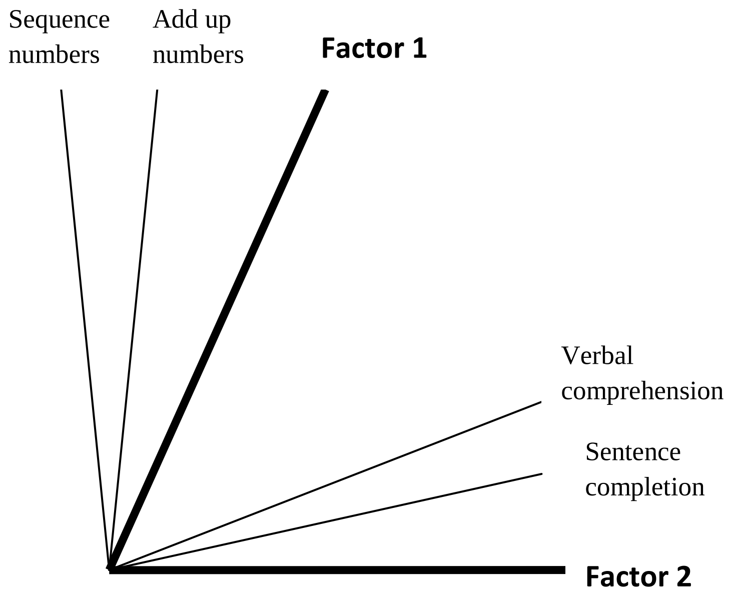
In factor extraction, *hypothetical* variables are placed in the best position to capture the pattern of inter-correlations in the correlation matrix. First all the people about whom data have been collected are given a score on a newly created variable, called a factor. The factor is selected in such a way that it correlates as highly as possible with all of the other variables, the ones that have actually been measured. Once this has been done, the correlation between this factor and all the other variables is eliminated (technically, 'partialled out') from the correlation matrix. This produces a new correlation matrix between the variables, one in which the variables do not correlate at all with the factor. A second factor is then created from this new correlation matrix, and the correlation between this second factor is then eliminated, producing a third correlation matrix, from which a third factor is created, and so on.

In Figure 5 two factors have been extracted, Factor 1 and Factor 2. The angle between each factor and the measured variables indicates the degree of correlation between them. So Factor 1 correlates relatively highly with adding up numbers because the angle between them is small, but relatively poorly with sentence completion as the angle this time is quite large.

The correlations between the test scores and the factors are called factor loadings. So the factor loadings of for adding up numbers and sequencing numbers are relatively high with respect to Factor 1, but low for Factor 2. Similarly the factor loadings for sentence completion and verbal comprehension are relatively high for Factor 2, but low for Factor 1.

Figure 5

Factor extraction



Factor rotation

The third step in factor analysis is called *factor rotation*. For mathematical reasons, when factors are extracted they are not placed in the ideal position with respect to the measured variables. That is, it is generally difficult to interpret the relationships between the factors and the measured variables from which they have been derived. To solve this problem the factors are *rotated* so that they are in the best possible position for interpretation purposes.

Figure 6

Factor Rotation

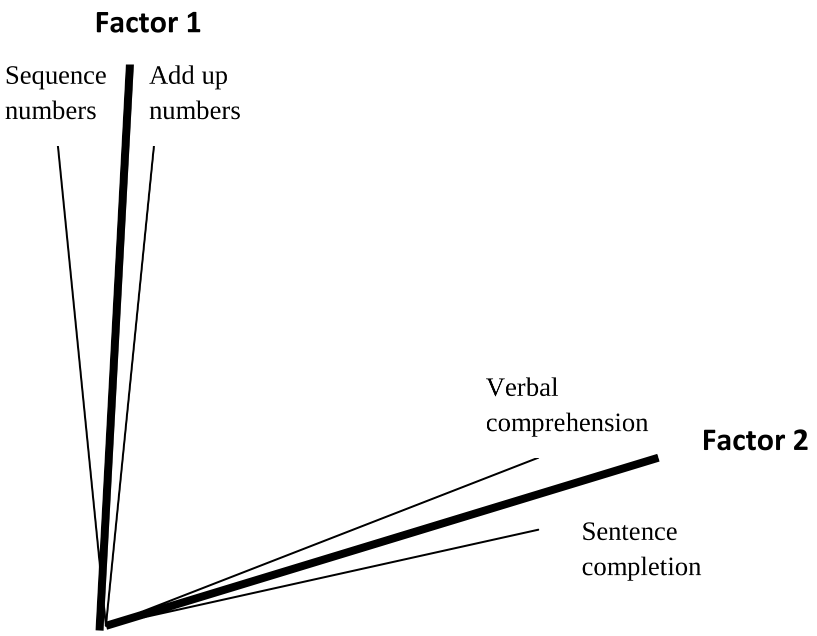


Figure 6 shows the position of the factors after rotation in this fictitious example. Here the rotation of the factors has maximized the loadings of the tests of number sequencing and adding up numbers on Factor 1, and minimized the loadings the tests of verbal comprehension and sentence completion on that factor. Similarly the loadings of verbal comprehension and sentence completion have been maximized on Factor 2, and minimized on Factor 1. This step makes it considerably easier to interpret the results of the factor analysis.

Interpreting the results of the analysis

The results of the factor analysis indicate the amount of the variance between the variables that each factor accounts for, and provides loadings of all the variables on each factor. A table like that shown in Table 12 is produced showing the loading of each variable on each factor.

Table 12

Fictitious Loadings of Two Factors

	Factor loading	
	Factor 1	**Factor 2**
Sequencing numbers	.79	.03
Adding up numbers	.68	.06
Sentence completion	.12	.77
Verbal comprehension	.14	.83

This indicates that the loadings for Factor 1 are 0.79 for sequencing numbers, 0.68 for adding up numbers, and so on. The convention is to take seriously any loading that is equal to or greater than 0.32. So here Factor 1 is highly associated with sequencing numbers and adding up numbers but not with the other two tests. In contrast, Factor 2 is highly associated with sentence completion and verbal comprehension but not with sequencing numbers of verbal comprehension.

The final step is to interpret the results and label the factors. In the example here this should be quite straightforward. As the sequencing numbers and adding up numbers tests are both clearly concerned with being numerate, Factor 1 might be given a label such as “numerical ability”. In the same way because the sentence completion and verbal comprehension tests are clearly associated with being verbally able, Factor 2 might be labelled “verbal ability”.

Although the sample above sets out the essential steps in factor analysis, it is very much a simplified account because in reality the number of variables is often 20 or more rather than just four. Nevertheless, even with 100 variables the steps are essentially the same: creating a correlation matrix, factor extraction, factor rotation, and interpreting the results of the analysis and labelling the factors.

In this example the fictitious results of the four tests resulted in a factor analytic solution in which two quite independent factors, verbal ability and numerical ability, emerged. However, if peoples' performance on the four tests were found to be highly correlated (see Figure 3) a factor analysis would produce a one-factor solution. This one factor might be labelled general ability, or *g.*   As explained on page 27, intensive research over the last 20 years indicates that the correlations between peoples' scores on different tests resembles the relationship depicted in Figure 7 rather than Figure 4, and that a single factor representing *g* explains much of the variance between ability test scores (see Figure 8).

Figure 7

The Geometrical Representation of Strong Correlations Between the Results of Four Ability Tests

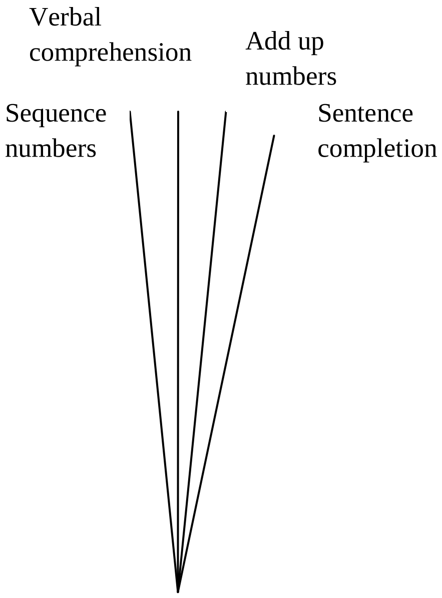
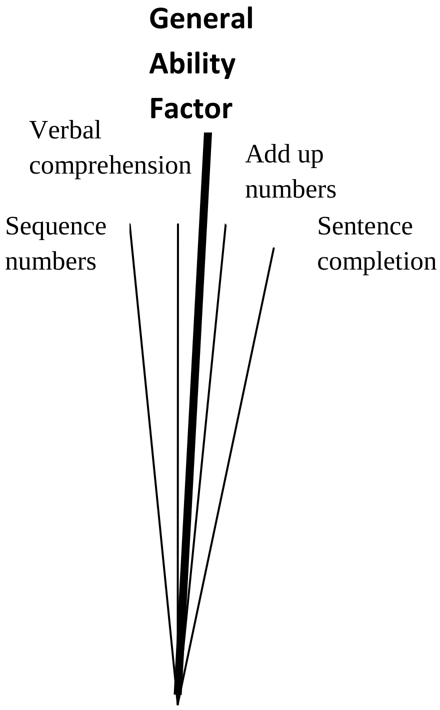


Figure 8

The Geometrical Representation of Strong Correlations Between Four Ability Tests: A Single Factor Solution



Appendix 3: Examples of Aptitude Tests

In the United States the *Buros Center for Mental Measurements* provides independent reviews of a large number of psychological tests. A search on the Buros website for tests of aptitude yielded the following 68 tests.

1. Academic Aptitude Test: Non-Verbal Intelligence: Acorn National Aptitude Tests
2. Academic Aptitude Test: Verbal Intelligence: Acorn National Aptitude Tests
3. Accounting Aptitude Test
4. Analytical Aptitude Skills Evaluation
5. Aptitude Assessment Battery: Programming
6. Aptitude Interest Inventory
7. Aptitude Profile Test Series
8. Aptitude Tests for School Beginners
9. Aptitudes Associates Test of Sales Aptitude: A Test for Measuring Knowledge of Basic Principles of Selling, Revised
10. Armed Services Vocational Aptitude Battery [Forms 18/19]
11. Ball Aptitude Battery, Form M
12. Canadian Dental Aptitude Test
13. Clerical Aptitude Test: Acorn National Aptitude Tests
14. Clerical Aptitudes
15. College Board Scholastic Aptitude Test and Test of Standard Written English
16. Computer Operator Aptitude Battery
17. Computer Programmer Aptitude Battery
18. Detroit Tests of Learning Aptitude, Fourth Edition
19. Detroit Tests of Learning Aptitude-Adult
20. Detroit Tests of Learning Aptitude-Primary, Third Edition
21. Differential Aptitude Tests for Personnel and Career Assessment
22. Differential Aptitude Tests, Fifth Edition
23. Differential Aptitude Tests-Australian and New Zealand Editions [Forms V and W]
24. Differential Aptitude Tests-Computerized Adaptive Edition
25. Electrical Aptitude Test (Form EA-R-C)
26. Employee Aptitude Survey, Second Edition
27. Eosys Word Processing Aptitude Battery

28. Evaluation Aptitude Test
29. Flanagan Aptitude Classification Tests
30. Group Diagnostic Reading Aptitude and Achievement Tests
31. Group Diagnostic Reading Aptitude and Achievement Tests, Intermediate Form
32. Hay Aptitude Test Battery [Revised]
33. Inventory of Vocational Interests: Acorn National Aptitude Tests
34. Iowa Algebra Aptitude Test(tm), Fifth Edition
35. Management Aptitude Test
36. Mechanical Aptitude Test (Form MAT-AR2-C)
37. Mechanical Aptitude Test: Acorn National Aptitude Tests
38. Mechanical Aptitudes
39. Multi-Craft Aptitude Test
40. Multidimensional Aptitude Battery-II
41. Musical Aptitude Profile [1995 Revision]
42. NSight Aptitude/Personality Questionnaire
43. Occupational Aptitude Survey and Interest Schedule-Third Edition
44. P.C. User Aptitude Test
45. PASAT 2000 [Poppleton Allen Sales Aptitude Test]
46. Profile of Aptitude for Leadership
47. Programmer Analyst Aptitude Test [One-Hour Version]
48. Programmer Analyst Aptitude Test [Two-Hour Version]
49. Programmer Aptitude Battery
50. Programmer Aptitude Series
51. PSB Aptitude for Practical Nursing Examination
52. PSB Health Occupations Aptitude Examination
53. PSB Registered Nursing School Aptitude Examination
54. Roeder Manipulative Aptitude Test
55. Sales Aptitude Test
56. Scholastic Aptitude Scale
57. Senior Aptitude Tests
58. Systems Analysis Aptitude Test
59. Systems Programming Aptitude Test
60. Trade Aptitude Test Battery
61. USES General Aptitude Test Battery
62. USES General Aptitude Test Battery for the Deaf
63. USES Nonreading Aptitude Test Battery, 1982 Edition

64. W-APT Programming Aptitude Test
65. Wiesen Test of Mechanical Aptitude (The)
66. Wolfe Computer Operator Aptitude Test (The)
67. Wolfe-Spence Programming Aptitude Test
68. Work Aptitude: Profile and Practice Set 1

A brief outline of some of the most commonly used tests of ability/aptitude is set out here. The description of each test covers its nature and objectives, what it seeks to measure, and who it is intended for. Where it has been possible to obtain the relevant information, the number of items in each test and the time required to complete the test have been included also. The tests are listed below.

- American College Test (ACT)
- Armed Forces Vocational Aptitude Battery (ASVAB)
- BioMedical Admissions Test (BMAT)
- Dental Admission Test (DAT)
- Differential Aptitude Test (DAT) for Personnel and Career Assessment
- Employee Aptitude Survey (EAS)
- Flanagan Aptitude Classification Tests (FACT)
- General Aptitude Test Battery (GATB)
- Law School Admission Test (LSAT)
- Medical College Admission Test (MCAT)
- Multidimensional Aptitude Battery-II
- Professional Aptitudes
- Raven’s Progressive Matrices
- SAT Reasoning Test

- UK Clinical Aptitude Test (UKCAT)
- Wonderlic Cognitive Ability Test

American College Test (ACT)

The ACT is curriculum-based. The designers state that the ACT is not an aptitude or an IQ test. Instead, the questions on the ACT are directly related to what students have learned in high school courses in English, mathematics, and science.

Sub-scores:

- English
- Math
- Reading
- Science
- Writing (optional)

Time: 2hrs 55mins + 30mins for Writing option

No. of items: 215

Armed Forces Vocational Aptitude Battery (ASVAB)

The ASVAB is the entrance test to enlist in the US Military but is also used by employer organizations at entry levels. The military use of ASVAB has two primary purposes: First, it determines whether applicants have the mental aptitude to enlist in the military branch of their choice, and second, the results help the service(s) determine which military job(s) applicants have the mental aptitude for.

- Word Knowledge
- Paragraph Comprehension

- Mathematics Knowledge
- Arithmetic Reasoning
- General Science
- Auto and Shop Information
- Mechanical Comprehension Test
- Electronics Information
- Numerical Operations
- Coding Speed

BioMedical Admissions Test (BMAT)

BMAT is a subject-specific admissions test taken by applicants to certain medicine, veterinary medicine and related courses at particular institutions. The test is in three parts and is designed to assess skills in problem solving, understanding argument and data analysis and inference; a candidate's ability to apply scientific knowledge normally encountered in non-specialist school science and maths courses; the ability to select, develop and organise ideas and communicate them in writing in a concise and effective way.

Sub-scores:

- Aptitude and Skills
- Scientific Knowledge and Applications
- Writing Task

Time: 2hrs

No. of items: 35, 27 + 1 essay question from a choice of 4

Dental Admission Test (DAT)

DAT is a computer based standardized exam taken by potential dental school students in the United States and Canada. It is designed to help students assess their aptitude for a career in dentistry and to assist dental schools in selecting first-year-students. Although there is a separate Canadian version with differing sections, both American and Canadian versions are usually interchangeably accepted in both countries' dental schools. This outline describes the American DAT.

Sub-scores:

- Survey of the Natural Sciences
- Perceptual Ability
- Reading Comprehension
- Quantitative Reasoning

Time: 4hrs 15min

No. of items: 30 - 40

Differential Aptitude Test (DAT) for Personnel and Career Assessment

The DAT assesses eight different types of ability, or aptitude, which are related to success in different areas of employment. It is essentially a profiling instrument. Its co-standardised tests provide an eight point profile which portrays relative strengths and weaknesses in an individual's key aptitudes.

Sub-scores:

- Verbal reasoning
- Numerical reasoning
- Abstract reasoning
- Clerical Speed and Accuracy

- Mechanical Reasoning
- Space Relations
- Spelling
- Language Usage

Time: 2hrs (6-20 minutes per test)

Employee Aptitude Survey (EAS)

The EAS is a series of tests designed to assess the cognitive, perceptual, and psychomotor abilities that are required for successful job performance in a wide variety of occupations. Each of the EAS tests can be used individually or as part of a battery to assist in employee selection and career counselling.

Sub-scores:

- Verbal Comprehension
- Numerical ability
- Space Visualisation
- Numerical reasoning
- Verbal Reasoning
- Symbolic Reasoning

Time: 5-10mins per test

No. of items: 30 - 75

Flanagan Aptitude Classification Tests (FACT)

The FACT assesses aptitudes that are important for successful performance of particular job-related tasks. An individual’s aptitude can then be matched to the job tasks. The FACT helps to determine the tasks in which a person has proficiency. Each test measures a specific skill that is important for particular occupations. The

FACT battery is designed to provide measures of an individual's aptitude for each of 16 job elements. In developing the FACT battery, Dr. Flanagan did not copy types of items found in other research; rather he devised new item types designed to measure the specific job elements.

The tests provide a broad basis for predicting success in various occupational fields. All are paper and pencil tests that can be given to an individual or group by a single examiner. Each of the 16 tests is printed in a separate booklet. This allows the tests to be administered individually or as a complete battery. The tests include:

Sub-scores:

FACT 1 - Inspection
FACT 2 - Coding
FACT 3 - Memory
FACT 4 - Precision
FACT 5 - Assembly
FACT 6 - Scales
FACT 7 - Coordination
FACT 8 - Judgment and Comprehension*
FACT 9 - Arithmetic
FACT 10 - Patterns
FACT 11 - Components
FACT 12 - Tables
FACT 13 - Mechanics
FACT 14 - Expression*
FACT 15 - Reasoning
FACT 16 – Ingenuity

* Untimed

Time: 30mins per timed test

General Aptitude Test Battery (GATB)

The *GATB* measures nine distinct aptitudes using 12 separate tests (eight pencil and paper tests, and four performance tests):

- General Learning Ability
- Verbal Aptitude
- Numerical Aptitude
- Spatial Aptitude
- Form Perception
- Clerical Perception
- Motor Co-ordination
- Finger Dexterity
- Manual Dexterity

LNAT

LNAT was developed by a consortium of UK universities. The stated purpose of the test as stated on the LNAT website is to provide "a fair way to assess a candidate's potential to study law at undergraduate level, regardless of their education or personal background".

The test is comprised of a multiple choice test and a written essay. It is designed to measure the following verbal reasoning skills:

- comprehension
- interpretation
- analysis
- synthesis
- induction
- deduction

A candidate's LNAT score and written essay are forwarded to the Law School to which he or she has applied. In making selection decisions participating UK Law Schools use LNAT results in combination with traditional methods of selection such as A Level results and interviews. There is no standardised format for combing

LNAT scores with other information about a candidate's potential, and this is left to the discretion of the Law Schools using the test.

Time: 135 mins. (recently extended from 120 mins.) with multiple choice test 95 mins. and essay 40 mins.

No of items: 42 (recently increased from 30)

Law School Admission Test (LSAT)

The LSAT has been designed as part of the admission process for all American Bar Association approved law school applicants. It is also widely used elsewhere, for example by Canadian and Australian law schools. The test consists of five sections (one unmarked) and is designed to measure the reading and comprehension of complex texts with accuracy and insight; the organization and management of information and the ability to draw reasonable inferences from it; the ability to think critically; and the analysis and evaluation of the reasoning and arguments of others.

Sub-scores:

- Reading Comprehension
- Analytical Reasoning
- Logical Reasoning

Reading Comprehension Questions

These are designed to measure the ability to read, with understanding and insight, examples of lengthy and complex materials similar to those commonly encountered in law school. The Reading Comprehension section contains four sets of reading questions, each consisting of a selection of reading material, followed by five to eight questions that test reading and reasoning abilities.

Analytical Reasoning Questions

The analytical reasoning question are designed to measure the ability to understand a structure of relationships and to draw logical conclusions about that structure.

Respondents are required to reason deductively from a set of statements and rules or principles that describe relationships among persons, things, or events. The purpose of the Analytical Reasoning questions is to reflect the complex analyses a law student performs when solving legal problems.

Logical Reasoning Questions

The Logical Reasoning questions are designed to assess the ability to analyze, critically evaluate, and complete arguments as they occur in ordinary language. Each Logical Reasoning question requires the respondent to read and comprehend a short passage, then answer a question about it. The questions are designed to assess a wide range of skills involved in thinking critically, and emphasize key skills for legal reasoning. These skills include determining how additional evidence affects an argument, reasoning by analogy, identifying argument flaws drawing well-supported conclusions, and applying principles or rules.

Writing Sample

Applicants have 35 minutes to write an essay on the topic provided. Although the writing sample is not graded, Law Schools will receive a copy of it. They are said to pay attention to clarity, grammar, and word usage when evaluating the applicants work.

Logical Reasoning Section I

Time: 35 minutes
Format: 24-26 questions
Topics Tested: Analyzing Arguments and Evaluating Arguments

Logical Reasoning Section II

Time: 35 minutes
Format: 24-26 questions
Topics Tested: Analyzing Arguments and Evaluating Arguments

Logic Games Section

Time: 35 minutes
Format: 22-24 questions
Topics Tested: Basic Logic, Systems of Order, and Outcomes

Reading Comprehension Section

Time: 35 minutes
Format: 26-28 questions
Topics Tested: Identifying Purpose, Identifying Structure, and Ascertaining Main Idea

Experimental Section

Time: 35 minutes
Format: 22-28 unscored, experimental questions
Topics Tested: Any material tested in other LSAT sections
Question Types: Could be any from other LSAT sections

Writing Sample

Time: 35 minutes
Format: Two-page written response to a prompt
Topics Tested: Writing Ability, Ability to Argue a Position, and Ability to Analyze an Argument

Time: 6x35mins

No. of items: 23 - 28

Medical College Admission Test (MCAT)

MCAT, is a computer based standardized examination for prospective medical students in the United States and Canada. It is designed to assess problem solving, critical thinking, written analysis, and writing skills in addition to knowledge of scientific concepts and principles. Prior to August 19, 2006, the exam was a paper-and-pencil test; since January 27, 2007, however, all administrations of the exam have been computer-based.

Sub scores:

- Physical Sciences
- Verbal Reasoning
- Writing Sample
- Biological Sciences

Time: 4hrs 20mins

No. of items: 40 – 52 (2 prompts for the Writing Sample)

Multidimensional Aptitude Battery-II

The Multidimensional Aptitude Battery-II (MAB-II) assesses aptitudes and intelligence. It yields a profile of ten subtest scores, and scores for Verbal, Performance and Full Scale. Scores can be expressed as standard scores, percentiles, or IQ's. The 10 domains of intellectual functioning are grouped into two broad categories of ‘Verbal’ and ‘Performance’.

Sub scores:

- Information
- Comprehension
- Arithmetic
- Similarities
- Vocabulary
- Digit Symbol
- Spatial
- Picture Arrangement
- Object assembly

Time: 10x7mins

Professional Aptitudes

The Professional Aptitudes tests have been designed by Saville Consulting for use with managers, directors and professionals. The test publishers claim that each test is shorter than the industry standard whilst maintaining robust reliability and validity. There are three types of test, "verbal analysis", "numerical analysis" and "diagrammatic analysis".

Professional Verbal Analysis

This assesses the ability to evaluate complex written information. The assessment contains a series of single and dual passages followed by questions. Respondents base their answer to the questions on the information presented.

Sub-scores:

- Understanding Word Meaning
- Comprehending Text
- Making Verbal Inferences
- Evaluating Written Materials
- Comparing Arguments

Time: 20 mins

No of items: 28

Professional Numerical Analysis

This assesses the ability to comprehend, evaluate, and process numerical data. The assessment contains a series of single or dual data sets, followed by questions which need to be answered using the data presented.

Sub-scores:

- Understanding Tables
- Comprehending Graphs
- Making Numerical Inferences
- Evaluating Quantities
- Comparing Data

Time: 20mins

No of items**:** 28

Professional Diagrammatic Analysis

This assesses the ability to evaluate processes represented diagramatically. Diagrams in the form of panels and illustrations that define logical processes are presented. Respondents answer questions based on these diagrams.

Sub-scores:

- Understanding Tables
- Comprehending Graphs
- Making Numerical Inferences
- Evaluating Quantities
- Comparing Data

Time: 20mins

No of items**:** 28

SAT

This was formerly known as the Scholastic Aptitude Test (SAT). The SAT is designed to measure literacy and writing skills that are needed for academic success in college. It also assesses how well the test takers analyze and solve problems.

Sub-scores:

- Critical reading
- Writing
- Mathematics

Time: 3hrs 45mins

UK Clinical Aptitude Test (UKCAT)

The test assesses a range of mental abilities and behavioural attributes identified by university medical and dental Schools as important.

Sub-scores:

- Verbal Reasoning
- Quantitative reasoning
- Abstract Reasoning
- Decision Analysis
- Non-cognitive analysis (robustness, empathy, integrity)

Time: 2hrs (excl. non-cognitive analysis where timing varies)

No. of items: 28 - 65

Appendix 4: The British Psychological Society's Psychological Testing Centre - Directory of Test Publishers

- Assess Systems (Bigby Havis & Associates)
- Caliper UK
- CDA Profile Ltd
  Centre for Corporate Culture
- Consulting Tools Ltd
- Criterion Partnership Ltd
- Eras Ltd
- Eysenck Cripps Cook Occupational Scales
- GL Assessment
- Hogrefe
- Hudson
- Human Factors UK Ltd
- JCA (Occupational Psychologists) Ltd
- Kenexa
- Knight Chapman Psychological Ltd
- Lafayette Instrument Company
- Master Management International A/S
- MHS (UK)
- Mindmill
- OPP Ltd
- Pario Innovations Ltd
- Personal Consultancy Solutions Ltd
- Previsor (ASE)
- Profiles International Inc

- Psychological Consultancy Ltd
- Psytech International Ltd
- Quest Partnership Ltd
- Saville Consulting UK Ltd
- Schuhfried
- Selby & Mills Ltd
- SHL Group Ltd
- Stuart Robertson and Associates
- TalentQ
- Team Focus Ltd
- The Holst Group
- The Morrisby Organisation
- Thomas International Ltd
- TMS Development International Ltd
- View Assessments International Inc

References

Bangertdrowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1983a). Effects of coaching performance on achievment - test performance. *Review of Educational Research, 53*(4), 571-585.

Bangertdrowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1983b). Synthesis of research on the effects of coaching for aptitude and admissions tests. *Educational Leadership, 41*(4), 80-82.

Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive-validity designs – a critical reanalysis. *Journal of Applied Psychology, 66*(1), 1-6.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1962). *Differential aptitude tests*. New York: Psychological Corporation.

Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux. *L'Année psychologique, 11*, 191-224.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*(3), 561-589.

Brand, C. (1996). *The g factor: General intelligence and its implications*. New York: Wiley.

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego, CA: Academic Press.

Bryon, M., Clayden, J., & Tyreman, C. J. (2010). *How to master the UKCAT : Over 750 practice questions for the United Kingdom clinical aptitude test* (2nd ed. ed.). London: Kogan Page.

Butterworth, J., & Thwaites, G. (2010). *Preparing for the BMAT: The official guide to the Biomedical Admissions Test* Portsmouth, NH: Heinman.

Cook, M. (2006). *Personnel selection: Adding value through people*. Chichester: Wiley.

Crawford, C., Johnson, P., Machin, S., & Vignoles, A. (2011). Social Mobility: A Literature Review: Department for Business. Innovation and Skills.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(281-302).

Dewberry, C. (2001). Performance disparities between whites and ethnic minorities: Real differences or assessment bias? *Journal of Occupational and Organizational Psychology, 74*, 659-673.

Dewberry, C. (In press). Integrating Candidate Data: Consensus or Arithmetic? . In N. Povah & C. Thornton (Eds.), *Assessment and Development Centres: Strategies for Global Talent Management*. Farnham, Surrey: Gower.

Donlon, T. F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.

Emery, J. L., & Bell, J. F. (2009). The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. *Medical Education, 43*(6), 557-564.

Emery, J. L., & Bell, J. F. (2011). Comment on I. C. McManus, Eamonn Ferguson, Richard Wakeford, David Powis and David James (2011). Predictive validity of the BioMedical Admissions Test (BMAT): An evaluation and case study. *Medical Teacher, 33*(1), 58-59.

Emery, J. L., Bell, J. F., & Rodeiro, C. L. V. (2011). The BioMedical Admissions Test for medical student selection: Issues of fairness and bias. *Medical Teacher, 33*(1), 62-71.

Gardner, H. (1983). *Frames of Mind: The Theory of Mulltiple Intelligences*. New York: Basic Books.

Gardner, H. (1992). *Multiple intelligencies: The theory in practice*. New York: Basic Books.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*(1), 79-132.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance, 15*(1-2), 25-46.

Green, M., & Hawley, N. (2009). *Succeeding in the Bio Medical Admissions Test 2009 (BMAT) (Entry to Medical School)*. London: Apply2 Ltd.

Gregory, R. J. (1994). Aptitude tests. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Gregory, R. J. (2010). *Psychological testing : history, principles, and applications* (6th ed.). Boston, Mass.: Allyn and Bacon.

Guildford, J. P. (1967). *The Nature of Human Intelligence*. New York: McGraw Hill.

Guildford, J. P. (1985). The structure-of-intellect model. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements and applications*. New York: Wiley.

H.M.Government. (April 2011). *Opening Doors, Breaking Barriers: A Strategy for Social Mobility*.

Herrnstein, R., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American Life*. New York: Free Press.

Higgins, L. T., & Sun, C. H. (2002). The development of psychological testing in China. *International Journal of Psychology, 37*(4), 246-254.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection And Assessment, 9*(1-2), 152-194.

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83*(2), 179-189.

Hunter, J. E. (1986). Cognitive-ability, cognitive aptitudes, job knowledge, and job-performance. *Journal of Vocational Behavior, 29*(3), 340-362.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Hutton, G., Hutton, R., & Taylor, F. (2010). *Passing the UKCAT and BMAT 2010*. London: Learning Matters.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jensen, A. R. (1986). G - Artifact or reality. *Journal of Vocational Behavior, 29*(3), 301-331.

Jensen, A. R. (1998). *The G factor - The science of mental ability*. Westport, CT: Praeger.

Julian, E. R. (2005). Validity of the Medical College Admission Test for predicting medical school performance. *Academic Medicine, 80*(10), 910-917.

Kelley, T. L. (1928). Crossroads in the mind of man: A study of differentiable mental abilities. Stanford, CA: Stanford University Press.

Kirkup, C., Wheater, R., Morrison, J., Durbin, B., & Pomati, M. (2010). Use of an aptitude test in university entrance: a validity study: Final Report: National Foundation for Educational Research.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). New York: Routledge.

Kulik, J. A., Bangertdrowns, R. L., & Kulik, C. L. C. (1984). Effectiveness of coaching for aptitude-tests. *Psychological Bulletin, 95*(2), 179-188.

Lynch, B., MacKenzie, R., Dowell, J., Cleland, J., & Prescott, G. (2009). Does the UKCAT predict Year 1 performance in medical school? *Medical Education, 43*(12), 1203-1209.

McGue, M., Bouchard, T., Iacono, W., & Lykken, D. (1993). Behavior genetics of cognitive ability: A lifespan perspective. In R. Plomin & G. McClearn (Eds.), *Nature, nurture, and psychology*. Washington, D.C.: American Psychological Association.

McManus, I. C., Ferguson, E., Wakeford, R., Powis, D., & James, D. (2011). Predictive validity of the Biomedical Admissions Test: An evaluation and case study. *Medical Teacher, 33*(1), 53-57.

Morris, S. B., & Lobsenz, R. E. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53*(1), 89-111.

Nijenhuis, J. T., & vanderFlier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal Of Applied Psychology, 82*(5), 675-687.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria - not much more than g. *Journal of Applied Psychology, 79*(6), 845-851.

Patterson, F., Carr, V., Zibarras, L., Burr, B., Berkin, L., Plint, S., . . . Gregory, S. (2009). New machine-marked tests for selection into core medical training: evidence from two validation studies. *Clinical Medicine, 9*(5), 417-420.

Picard, O. (2009). *Get into medical school : 600 UKCAT practice questions : includes full mock exam, comprehensive tips, techniques and explanations*. London: ISC Medical.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and theory* (3rd ed.). Mahwah, N.J.: Lawrence Erlbaum.

Ree, M. J., & Carretta, T. R. (2002). g2K. *Human Performance, 15*(1-2), 3-23.

Ree, M. J., & Earles, J. A. (1991). Predicting training success - Not much more than g. *Personnel Psychology, 44*(2), 321-332.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance - not much more than g. *Journal of Applied Psychology, 79*(4), 518-524.

Reeve, C. L., & Hakel, M. D. (2002). Asking the right questions about g. *Human Performance, 15*(1-2), 47-74.

Salgado, J. F., & Anderson, N. (2002). Cognitive and GMA testing in the European community: Issues and evidence. *Human Performance, 15*(1-2), 75-96.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*(1-2), 187-210.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel-selection. *Annual Review of Psychology, 43*, 627-670.

Schmitt, N., Gooding, R.Z. Noe, R.A. and Kirsch, M. (1984). Metaanlyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Schultz, M. S., & Zedeck, S. (2008). Idenfication, development, and validation of predictors for successful lawyering. Retrieved from http://ssrn.com/abstract=1442118 website:

Simien, E. (1986). Law School Admission Test as a Barrier to Almost Twenty Years of Affirmative Action. *Thurgood Marshall Law Review, 12*, 359-393.

Spearman, C. (1904). 'General intelligence': objectively determined and measured. *American Journal of Psychology, 15*, 201-292.

Spearman, C. (1923). *The nature of 'intelligence' and the principles of cognition*. London: Macmillan.

Spearman, C. (1927). *The Abilities of Man*. New York: Macmillan.

Stern, W. L. (1912). Uber die psychologischen methodern der intelligenzprufung. [The psychological methods of intelligence testing]. *Educational Psychology Monographs, 13*.

Sternberg, R. J. (1985). *Beyond IQ: A Triarchic Theory of Human Intelligence*. New York: Cambridge University Press.

Sternberg, R. J. (1986). *Intelligence applied: Understanding and increasing you intellectual skills*. San Diego, CA: Harcourt Brace Javanovich.

Sternberg, R. J. (1996). *Successful intelligence*. New York: Simon Schuster.

Stilwell, L. A., Dalessandro, S. P., & Reese, L. M. (2009). Predictive Validity of the LSAT: A National Summary of the 2007 and 2008 LSAT Correlation Studies: Law School Admission Council.

Terman, L. M. (1916). *The meaurement of intelligence*. Boston: Houghton Mifflin.

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior, 29*(3), 332-339.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38*(406-427).

Thurstone, L. L. (1938). *Primary Mental Abilities*. Chicago: University of Chicago Press.

Tyreman, C. J. (2009). *How to master the BMAT: Unbeatable preparation for success in the Biomedical Admissions Test*. London: Kogan Page.

Wisler, C. (1901). The correlation of mental and physical tests. [Monograph supplement]. *The Psychological Review, 3*(6).

Wright, S. R., & Bradley, P. M. (2010). Has the UK Clinical Aptitude Test improved medical student selection? *Medical Education, 44*(11), 1069-1076.

Yates, J., & James, D. (2010). The value of the UK Clinical Aptitude Test in predicting pre-clinical performance: a prospective cohort study at Nottingham Medical School. *BMC Medical Education, 10*.

Zyphur, M. J., Bradley, J. C., Landis, R. S., & Thoresen, C. J. (2008). The effects of cognitive ability and conscientiousness on performance over time: A censored latent growth model. *Human Performance, 21*(1), 1-27.