Aptitude Testing for University Entrance:

A Literature Review

Angus S. McDonald

Paul E. Newton

Chris Whetton

Pauline Benefield

Contents

List of Tables

List of Figures

Note: The SAT I: Reasoning Test has two elements, a verbal section and a 'math' section. Since this American English term is that used in the test itself, it has been referred to in this way throughout this report.

The authors would like to thank Liz Gibson for her careful work in producing the final manuscript.

Executive Summary

1. This report presents the results of a literature review on aptitude testing for university entrance. The work was commissioned by The Sutton Trust and undertaken by the National Foundation for Educational Research.

2. Evidence for the review was obtained by searching a number of electronic databases and the websites of the main organisations involved in admissions testing in the United States. The central areas covered in this review are an overview of admissions procedures in countries which use aptitude tests, the extent to which aptitude tests are fair predictors of university performance, and the effects aptitude testing may have on the wider education system. Research on the prediction of university performance in Britain is also reviewed, and the findings reported from interviews conducted at three universities currently developing additional ways of assessing applicants.

3. Probably the best known admissions test is the Scholastic Aptitude Test (SAT), which was introduced by the College Board in 1926 with the goal of standardising the admissions process to universities and colleges in the United States. In its current form, the SAT assesses verbal and math ability, mainly through multiple-choice items. Admissions tests in countries such as Israel and Sweden follow a similar format, and evidence suggests that they assess the same underlying constructs as the SAT.

4. Many colleges in the United States use SAT scores in conjunction with high school record as part of their selection process. However, there is considerable variation in practices, with some colleges not using admissions test scores for some or even any of their students.

5. High school record and admissions test scores are able to predict college grades to a modest degree. Of the two, high school record is generally the better predictor, and once this has been taken into account the additional ability of the SAT to predict performance is limited.

6. There has been considerable controversy over the fairness of tests such as the SAT, particularly in relation to sex and ethnic score differences. This review considers a range of evidence to reach a judgement on test fairness. Evidence suggests that both the verbal and math sections of the SAT are somewhat biased in favour of males, although it cannot be ruled out that this is due to male SAT takers being a more highly selected group. Evidence for students from ethnic minorities is somewhat less conclusive. Although groups such as African Americans score about one standard deviation lower on the SAT than Whites, this pattern is repeated on virtually all tests which measure aspects of intelligence. Even when overt attempts have been made to equalise the scores of Blacks and Whites, Blacks are still seen to score lower. The reasons behind this are not fully understood.

7. The question of test fairness is probably most adequately addressed through an examination of how accurate admissions test scores are at predicting college attainment for different groups. It is clear that test scores do not always equate to attained college grades, but as the SAT accounts for a relatively small proportion of the variance in college grades, this is not surprising. There is consistent evidence that the SAT under-predicts female attainment, even after differential selection of courses for grading difficulty is

allowed for. The findings for ethnic groups are less consistent, with some researchers arguing test scores are a fair representation of attainment, whilst others provide contrary evidence.

8. The SAT was established by the College Board with the goal of standardising the admissions process to colleges in the United States. Although relatively successful in this goal, criticisms of the SAT have led to variations in how colleges use it, with some no longer requiring SAT scores at all. The extent to which different colleges consider admissions test scores and high school record to be important varies considerably, although test scores appear to have relatively little impact on students' choice of colleges.

9. Due to the importance attached to the SAT, it falls on high school tutors to prepare students to take this test. In doing this, students are distracted from their normal high school studies. Although the effects of this are hard to establish, critics of the SAT have targeted the time spent on SAT preparation at the cost of high school work as being a detrimental consequence of the test. A considerable industry has also built up around preparation and coaching for admissions tests. Controlled studies show the effects of coaching to be small, although this may have an impact when students apply to more selective colleges. Despite the effects of coaching being modest, the costs involved suggest that this may be a potential source of bias, as students from more affluent backgrounds may have greater access to preparation materials and courses.

10. Whereas some of the consequences of the SAT are obvious, it is less clear what would happen if it was abolished. It has been argued that this could lead to inflation of high school grades, as in the absence of a national attainment testing programme, the SAT acts as a measure against which high school performance can be judged. The effect of removing the SAT could therefore be to lower standards. However, in cases where reporting of SAT scores has been made optional, there is little evidence that a fall in standards has occurred.

11. In Britain there has been far less open debate on the fairness of A-levels, and relatively little research has been conducted into their effectiveness as a selection tool for higher education. One exception to this was a series of studies which investigated the possibility of using aptitude tests as part of the selection procedure into higher education, conducted during the 1960s and 1970s. This research found that A-levels were modest predictors of degree performance, consistent with other studies which have examined the link between A-levels and performance in higher education. The findings for the aptitude test were less encouraging, as this added virtually nothing to the prediction of degree grades in addition to A-levels.

12. As part of the debate on university access, it was revealed that a number of British universities were piloting assessments to provide additional information on students. Interviews were conducted with staff at three of these institutions, which explored the reasons behind considering additional tests, the nature of these tests, the research that was being conducted on them, and how they would be used if introduced.

13. Any use of a university admissions test in Britain must be considered in the context of the current British education system. The British system differs from that in the United States in that students are generally more selected by the time they apply to university, and that this selection takes place through a national assessment system. To an extent, the SAT

acts as a national benchmark in the absence of standardised assessments in the United States.

14. An admissions test such as the SAT could help admissions tutors at British universities discriminate between students. This can currently be difficult due to the large number of students attaining high A-level grades, but planned revisions to the A-level system may at least partially resolve this problem. If an admissions test was introduced, this would be likely to reduce the focus on A-levels and place additional demands on students and tutors. However, without further research it is not possible to say whether an admissions test could provide a fair assessment of potential for university study, and provide useful information in addition to that given by the current exam system.

1. Overview

This document reports the findings from a literature review on aptitude testing for university entrance. The work was commissioned by The Sutton Trust and conducted by the National Foundation for Educational Research (NFER).

This literature review was conducted at a time of increasing scrutiny of British universities and their admissions policies. Prompted largely by research conducted by The Sutton Trust, much of this attention concerned the apparent bias in university selection procedures. The work of The Sutton Trust focused specifically on the 13 universities ranked highest according to league tables published by *The Times, Telegraph, Sunday Times* and *Financial Times* (The Sutton Trust, 2000). The proportion of students attending these universities was analysed by the type of school they had attended and their social class. It was found that students from independent schools were over-represented in the 13 highest-ranked universities, and students from the lower social classes were under-represented. When the five highest-ranked universities were considered, the bias in favour of independent schools was even more pronounced. This bias remained once the academic performance of students had been taken into account.

These statistics came to public attention around the same time as a high-profile case of a student from a state school who attained five A grades at A-level but was rejected by Oxford University (e.g. Stein, 2000), and together fuelled the debate on university access. From near the beginning of this debate, it was argued that the entrance procedures used in the United States of America offered a model which could be adopted in Britain to make university admissions fairer. Central to the American system is the use of the Scholastic Aptitude Test (SAT; renamed since 1994 the Scholastic Assessment Test). It was claimed that the SAT measures students' aptitude for college education, but more importantly, that it does this independently of social factors such as students' sex, ethnic background or social class (e.g. Clare, 1999). In terms of the debate on university access in Britain, this suggested the possibility of an assessment of potential for university education which could provide a fairer reflection of students' potential, regardless of their background or educational experiences.

Interest in aptitude tests and other tests of potential has been expressed by a number of British universities, and the SAT has been studied by the Qualifications and Curriculum Authority (QCA). As part of the investigation into the potential of the SAT, The Sutton Trust asked the NFER to conduct a pilot study examining the association between SAT scores and A-levels in high- and low-attaining schools. The results of this work have been presented separately (McDonald *et al.*, 2001). Due to the increased interest in the SAT, The Sutton Trust also commissioned the NFER to conduct the present literature review on the SAT and other aptitude tests used for university entrance. The evidence for this review was obtained from searches of electronic databases (see Appendix A for details of searches conducted), secondary sources obtained from these articles, and documents taken from the websites of Educational Testing Services, the College Board, American College Testing Program, and the Ministry of Education in Singapore.

This review initially describes the two tests most frequently used for university (more often referred to as 'college') admission in the United States - the Scholastic Assessment Test and the American College Test. The role of these tests in the admissions process is discussed, from the perspective of both admissions tutors and students. Overviews are then given of the admissions tests used for university entrance in a number of other countries.

The primary role of aptitude tests is to predict students' likely performance on a college or university course, and the ability of these tests and high school record to do this is examined in the following section. Considerable controversy has surrounded group differences in test scores, and evidence for score differences according to sex and ethnic status is studied. In an attempt to determine whether these differences are an accurate reflection of students' academic potential or result from limitations of the tests, research on the extent to which predicted grades are an accurate reflection of actual college attainment is reviewed.

Any testing system in education, such as that for university admissions, will inevitably have an impact on other areas of the educational process, as well as having broader social consequences. Considerable debate has centred on the effects of admissions testing in the United States, and what the positive and negative consequences of this are. Although much of this debate is necessarily speculative, particularly regarding the possible effects of no longer using tests such as the SAT, this is presented as it provides valuable insights into the possible consequences if aptitude testing was introduced in Britain.

In Britain, A-levels remain the most common way through which students gain access to university. Despite this, the ability of these qualifications to predict subsequent university performance has received only limited attention. An overview of the research in this area is given, along with a detailed account of research conducted on aptitude testing for university entrance by the NFER during the late 1960s and early 1970s. As part of the debate on university access and the SAT, it was revealed that a number of British universities were looking at assessments to provide additional information on applicants. Interviews were conducted with three universities looking at different forms of assessment, and the outcomes from these are also reported.

The findings from the overall review are brought together in the final section, where they are discussed in relation to the possibility of introducing aptitude testing for university entrance in Britain.

2. The use of aptitude testing in university entrance

2.1 Introduction

This section of the review describes how aptitude tests are used in admission to higher education in a number of countries. This review is necessarily selective, and focuses particularly on the American system as this has generated the greatest amount of research and discussion. The tests used in each of the countries are also described.

2.2 United States of America

The most widely known and intensively researched aptitude test for selection into higher education is the SAT, used in the United States since 1926. 'SAT' originally stood for Scholastic Aptitude Test, but as part of a revision in 1994 it was renamed the Scholastic Assessment Test. However, the SAT is not the only selection instrument for higher education in the United States, as the American College Testing Program has offered the ACT since 1959. This review firstly describes the SAT and the ACT, before looking at how they are used by American colleges in the selection of students and how students make use of the test scores.

2.2.1 The Scholastic Aptitude (or Assessment) Test

In 1900, in an attempt to reconcile the considerable variation in entrance procedures between colleges in the United States, the College Board was formed. The tests initially offered by the College Board from 1901 were open-ended and assessed students' understanding of specific subject areas including English, math, Greek, Latin, history and chemistry, rather than general aptitude for college-level work. The test that is recognised as the SAT was first offered in 1926, its development having been influenced by the extensive use of objective testing for army recruits during the First World War. Initially SAT results were only sent to colleges, but from 1958 test takers also received their scores. In 1947, the College Board helped found Educational Testing Services (ETS), and since then ETS have been responsible for developing the SAT.

The SAT gives two scores: verbal and math. Each of these is reported as a scaled score which has a mean of 500 and a standard deviation (a measure of the spread of scores) of 110. The maximum score for each part of the SAT is 800 and the minimum 200. Using these scaled scores it is possible to obtain an indication of how the verbal or math scores for an individual student compare with the scores of others, as this transforms all scores into an approximately normal distribution. As the scaled scores approximate a normal distribution, we know that 68 per cent of students will have a math and verbal score that lies within one standard deviation either side of the mean (that is, between 390 and 610) and 96 per cent will have a score within two standard deviations of the mean (that is, between 280 and 720). Only two per cent of test takers would be expected to score less than 280, with a further two per cent scoring above 720.

During the early years of the SAT, no attempts were made to relate the test scores obtained from one version of the test to those obtained from other versions. This changed in 1941, when the scores from the April administration of the SAT formed the basis of a scale to which subsequent tests were linked. This linking is done through a mechanism in which each new version of the SAT includes approximately 20 per cent of items from a previous version (Wainer, 1999). It is through these common items that scores from each new version of the SAT are linked to the mean of 500 established in the 1941 test. However, in 1995 the scaled scores were re-centred for the first time since 1941, as mean scores on the math and verbal tests had shifted. This had resulted in the mean verbal score being 428 and mean math score

482 in 1995. Due to the potential confusion this could cause, the 1995 scores were re-centred with each being set again to a mean of 500. This linking of scores allows standards to be monitored over time, but more importantly means that the scores of students who have taken the SAT at different sittings are directly comparable to each other.

Although modifications to the question formats and the focus of the test's content occurred over the years, the SAT remained largely unchanged until March 1994. The major changes to the SAT which occurred in 1994 are documented by Steven Graff (1993), project director of the working group responsible for the redevelopment of the SAT. Prior to 1994 the verbal section of the SAT consisted of:

- analogies measuring reasoning skills and knowledge of vocabulary (20 questions);
- sentence completions primarily measuring logical relationships among parts of a sentence (15 questions);
- antonyms measuring knowledge of vocabulary (25 questions);
- five or six reading passages of between 200 and 450 words each, with questions assessing inference, application or evaluation of logic or style, and questions relating to the main idea of the passage. Passages were drawn from a range of subject areas, and included at least one passage that was considered 'minority-relevant' (25 questions).

These questions were organised in two 30-minute sections. A 30-minute Test of Standard Written English (TSWE) was also included as part of the verbal test.

The major changes to the revised SAT involved dropping the antonym section and replacing the TSWE with a separate test which also included English composition - the SAT II: Writing Test. The emphasis on the assessment of reading was also increased in the redevelopment. This was done as reading, particularly the ability to read critically, was seen as being of central importance to any course of study at college. The revised, current verbal SAT consists of:

- analogies measuring reasoning skills and knowledge of vocabulary (19 questions);
- sentence completions measuring logical relationships and vocabulary in the context of the sentence (19 questions);
- four reading passages of between 400 and 800 words each, with questions assessing extended reasoning skills, literal comprehension and vocabulary in context. Two of the passages are paired and complement each other in some way (e.g. represent differing sides of an argument), and so support questions relating to differing styles of writing or points of view, and one passage is described as being 'minority-relevant' (40 questions).

These questions are organised in two 30-minute sections and one 15-minute section.

Prior to 1994 the math section consisted of standard multiple-choice items (40 questions) and quantitative comparisons (20 questions). These questions covered the areas of arithmetic, and algebraic and geometric reasoning, and were organised in two 30-minute sections. These two question types have been retained on the revised math section. The major differences with the revised math test are that students are now encouraged to use calculators and some questions require students to produce answers rather than choosing from a list of possible answers.

Both of these changes are seen to be in accordance with the greater emphasis on problem-solving in math. There are 35 multiple-choice questions in the revised math section, 15 quantitative comparisons and 10 open-ended questions, or 'student-produced' responses. As with the revised verbal test, the questions are organised in two 30-minute sections and one 15-minute section.

Minke (1996) reported that the estimated reliabilities of the revised SAT (an indication of the extent to which all of the questions 'hang' together and measure a single construct) were 0.92 for both the verbal and math sections. In the revision process the Scholastic Aptitude Test was renamed the 'Scholastic Assessment Test', abbreviated to 'SAT I: Reasoning Test'. This renaming took place due to mounting evidence that Blacks and other ethnic and social groups consistently scored lower on the SAT. Including the word 'aptitude' in the test title implied that the SAT measured innate differences. Commenting on the lower scores obtained by Black students, Jencks observed that this 'suggested that blacks might suffer from some sort of innate disability' (1998, p. 66), whereas using the word 'assessment' removed the innate connotation.

A second series of tests (SAT II: Subject Tests) has also been introduced. These have been developed from the College Board's Achievement Tests, and cover the major areas of academic study: English, math, social studies, natural sciences, and languages. Each of these tests lasts for one hour, and they are approximately equal to the difficulty of a high school final exam. The purpose of the SAT II: Subject Tests is to provide admissions tutors with an indication of the depth of subject-matter knowledge a student is likely to bring to their chosen course. Students choose whether to take subject tests, and although some colleges do not require students to take these, many recommend that they do.

The College Board (collegeboard.com, 2000) also offers the Preliminary SAT (PSAT), which is intended to be taken by students in their second year of high school. The PSAT is very similar in structure to the SAT, having verbal and math sections which contain the same question types as the full SAT. The main difference is that it also contains a test designed to measure writing skills, similar to the SAT II: Writing Test. In total there are 52 verbal questions, 40 math questions and 39 writing skills questions, and the whole test lasts two hours ten minutes.

The College Board offer a number of reasons for students to take the PSAT, including the opportunity to receive feedback on strengths and weaknesses in skills necessary for college study, as a comparison with other college-bound students, and as preparation for the SAT. The feedback students receive from the PSAT also provides guidance on college courses and possible careers, and students can choose for their scores to be entered into national scholarship competitions.

2.2.2 The American College Test

Although the SAT is by far the best-known college admissions test, it is not the only one used in the United States. The major competitor to the SAT is a test produced by the American College Testing Program, the ACT, which was first offered in 1959.

The SAT was intended to assess a student's potential to engage in college-level work through the assessment of verbal and math skills. In contrast to this, the ACT is oriented towards the major areas of high school and college instructional programmes, and so is more directly related to the student's educational progress. Whereas the SAT was developed more as an aptitude test, the ACT 'measures the knowledge, understanding, and skills that you have

acquired up to now' (American College Testing Program, 2000b). Despite the SAT and the ACT purporting to measure quite different constructs, the two have been seen to correlate around 0.9 (Schneider and Dorans, 1999), suggesting that the underlying constructs they measure are virtually identical. An examination of the subtests further indicates that although these cover four commonly studied areas, the question formats and emphasis on reasoning mean that the ACT shares much in common with the SAT.

The current ACT consists of four subtests:

- The English test contains 75 questions and lasts 45 minutes. Questions are designed to measure understanding of standard written English conventions and rhetorical skills. The test consists of five stimulus passages, each of which is followed by multiple-choice questions.

- The math test contains 60 questions and lasts for 60 minutes. It is designed to assess the skills students will typically have gained by Grade 12, but complex formulas and calculations are not required to answer questions. The test covers pre-algebra/elementary algebra, intermediate algebra/co-ordinate geometry and plane geometry/trigonometry.

- The reading test contains 40 multiple-choice questions and lasts for 35 minutes. It tests reading comprehension by requiring test takers to derive the meaning from four prose passages on different topics: social studies, natural sciences, prose fiction and the humanities.

- The science reasoning test contains 40 multiple-choice questions and lasts 35 minutes. Test takers are presented with seven sets of scientific information in the form of data representation (e.g. graphs), research summaries or conflicting points of view. Each of these is followed by questions designed to assess the interpretation, reasoning, analysis and problem-solving skills that are necessary for natural science courses.

The ACT gives separate scaled scores for each of the subtests which range from one to 36, and a composite score formed from averaging the scores to the four subtests.

2.2.3 Use of test scores by institutions

Despite it being widely believed that the use of the SAT or ACT is virtually ubiquitous by American colleges, a survey by the College Board has shown this was not so (Schaffner, 1985). Although the majority of colleges required applicants to take admissions tests, this was not universal. More recently Rooney with Schaeffer (1998) found that at least 275 selective four-year colleges were not using either the SAT or the ACT to make selection decisions about some or all of their students. Overall, in 1999 it was reported that about 85 per cent of colleges generally required students to take admissions tests (Schneider and Dorans, 1999).

Possibly because of the long history of the SAT and the ACT, and the controversy that has surrounded particularly the SAT, where tests are used there appears to be no single way in which their results are incorporated into the admissions process. Schaffner (1985) reported that 55 per cent of colleges routinely considered admissions tests when reaching decisions, whereas 21 per cent did not require them or rarely used them. Thirteen per cent said they were required or recommended but rarely actually used in admissions. Sixty-five per cent of institutions reported high school achievement as being ‘very important’ or ‘the single most important factor’ in reaching decisions, whereas only 42 per cent thought this of test results.

Further variations are highlighted by Smyth (1995), who presented the results from a survey of 360 colleges on their use of the SAT and ACT. Of the colleges who responded, 93 per cent reported that they accepted scores from either the SAT or the ACT, although highly selective colleges were more likely to require the SAT. Similar figures were reported by Schneider and Dorans (1999), who found that the majority of colleges accepted SAT or ACT scores. Only seven per cent of respondents to the survey said that they absolutely required achievement tests (e.g. SAT II: Subject Tests), although as this sample was somewhat biased towards more selective institutions, Smyth (1995) noted that this figure was likely to underestimate their use.
The ACT gives scores across four areas - English, math, reading and science reasoning - and a combined score. Eighty-eight per cent of colleges tended to focus on the composite ACT score, whereas 75 per cent considered the SAT verbal and math scores separately. However, in states where students predominantly took the ACT, there was a greater tendency to use the combined SAT score (Smyth, 1995).

In actually using SAT or ACT results, the most common approach involved combining these with information such as high school class rank (HSCR) and grade point average (GPA). High school information has consistently been identified as the best predictor of academic progress in college (see Section 3.2), and the additional value of the SAT has been questioned. Some have argued that the SAT makes little difference to selection decisions (e.g. Crouse and Trusheim, 1988), whereas other data shows that it can be useful in addition to high school record (e.g. Bridgeman *et al.*, 2000).

Using information from students who have already attended the college, it is possible to determine the association between the different predictor variables at the time of admission (e.g. SAT, ACT, HSCR, high school GPA) and college GPA. More sophisticated models may weight each predictor separately, according to their association with college grades. This process results in a prediction table, which shows likely college performance given an applicant's current attainment. Admissions tutors can use these tables to rank applicants according to probable college grades or to set a minimum cut-score which all applicants must exceed before they are considered further.

In some cases separate prediction tables may be developed and applied to applicants from different groups (e.g. males/females or according to their ethnic status) and academic disciplines. Blackburn (1990) observed that the use of separate prediction tables was more likely in highly selective colleges. This was justified as the impact of scores from tests such as the SAT can be greater under highly selective conditions, when the goal is to select high-attaining students (Ben-Shakhar *et al.*, 1996). However, a number of states have now prohibited affirmative action policies, and so use of these would be questionable if they led to the preferential treatment of one group over another (Perfetto *et al.*, 1999). Whether the use of such differential selection methods would be considered legal in some countries, including Britain, is doubtful.

A number of institutions have reported even wider variations in how they treat applicants' test scores. For example, colleges may allow students the option of withholding their SAT scores at the time of their application. An analysis of one such college by Schaffner (1985) suggested that those who withheld their scores tended to have slightly lower HSCR and SAT math and verbal scores. However, when SAT scores were collected on acceptance to the college, no differences were seen in the predictive validity for those who withheld SAT scores and those who submitted them. Schaffner reported that one of the motives behind making reporting of scores optional was to encourage admissions tutors to carefully screen

withholders, to see if they had potential that may not have been reflected in the SAT. Overall, this policy of making the reporting of SAT scores optional was seen as a success, due to it creating a more diverse student population whilst maintaining academic standards (Crouse and Trusheim, 1988; Rooney with Schaeffer, 1998).

Recent work by the College Board has attempted to define a taxonomy of the decision-making processes college admissions tutors employ (Perfetto *et al.*, 1999). From discussions with 50 admissions tutors, two broad models emerged: the eligibility model and the selective model. The eligibility model sets out objective and public criteria, and all students who meet these criteria are admitted. This was contrasted with selective models where comparisons may be made between students rather than just against set criteria. The report acknowledged that very few institutions operated simply according to the eligibility model, although many had eligibility criteria which must be met before a student's application could be considered further - that is, to progress to the selective stage.

In selective models, students may be selected according to three groups of factors. The first of these was labelled 'personal qualities' and included academic attainment and attributes, such as evidence of motivation and perseverance. The second group concerned the effect of education on the individual, and referred to the potential of the individual to benefit from education. This group also took into account the extent to which students had been able to overcome any educational adversities (e.g. attending a very poor high school). The final group of factors looked at the potential of the student to contribute to various areas. These included their potential contribution to wider society and to the student body (e.g. to sports teams), and also, particularly as college funding in America has become an area of particular concern, their ability to contribute financially towards their college education.

Each of these factors was presented as a distinct method of selection, but it was acknowledged that any single institution may use many of these methods in selecting a single intake. Often multi-stage selection methods were used, with the different factors being used at successive stages. Alternatively, many factors could be considered at the same time, with each of these being assigned different weights.

2.2.4 Use of test scores by students

A key difference between the American system and, for example, the British system, is that students in America tend to receive their SAT or ACT scores before they make their final applications to colleges. They are therefore able to use their test results to guide these decisions. This is likely to be particularly important when students are considering applying to more selective institutions, as they will be able to match their test scores to the general entry requirements of specific colleges. If a student's scores are too low for one of their intended institutions, they may revise their list of intended colleges and so obtain a greater chance of gaining a place. This can be contrasted with the British system where students have to select institutions on the basis of predicted grades - predictions that are often not very accurate (Delap, 1994; 1995).

The American system sounds good in theory, but does it accurately reflect what students do? Some answers to this question can be seen from a survey of students' reactions to the SAT conducted by Baird (1987). When looking at students' perceptions of what they thought were the most important factors in their admissions, high school GPA and SAT scores came out top. Fifty-eight per cent thought GPA had a 'great deal' of importance, with 34 per cent endorsing this view of SAT scores. Aptitude test scores appeared not to feed heavily into

students' decisions, as 72 per cent of SAT students said that their results had no effect on their choice of colleges, and ten per cent had already applied by the time they received their test results. However, a further ten per cent decided to apply to less selective colleges, and seven per cent applied to more selective colleges, because of their SAT scores.

Baird (1987) analysed responses by SAT scores and demographic variables. Students who obtained lower SAT scores were somewhat more likely to say that they affected their choice of colleges, with these students tending to apply to less selective colleges. Family income was not related to survey findings, but ethnicity was, as more Blacks and Hispanics than Whites said that their decisions were influenced by test results.

Overall, Baird's survey suggested that SAT results had less influence on students' decisions than many may assume. When they did influence choices, there was a tendency for them steer students towards less selective colleges, a pattern which was more pronounced amongst low scorers and those from ethnic minorities.

Grades from tests such as the SAT, PSAT or other test batteries are also used by high school and college counsellors when providing academic guidance to students. In order for test scores to be used in guidance, it is necessary that they have some predictive validity. When the PSAT and an assessment of interests were used together, Stricker *et al.* (1996) reported a median validity of 0.42 for the prediction of first-year grades in major college fields. This figure is high enough to suggest these measures would be useful for guidance, although not sufficiently high to warrant interpretation without caution. This work concluded that aptitude test scores can help students better understand what their grades in different subjects may be, particularly when the differences in grade distributions between subjects are taken into account (i.e. allowing for the fact that some college subjects will produce higher GPAs than others). This may be particularly important when achieving a high grade is a primary concern, as Stricker *et al.'s* (1996) research showed that a student may obtain very different grades in different majors.

2.3 Aptitude testing in Israel

The Israeli system of access to higher education is described by Beller (1994). Israel has national achievement tests similar to the Baccalaureate, taken at the end of high school. Successful students receive a matriculation certificate (Bagrut) which is based on a combination of these national exams and school assessment, and is needed for university entrance. Beller described that as the demand for higher education grew, many universities started to administer their own aptitude tests that were intended to be less dependent on applicants having studied a specific curriculum. In 1981, the National Institute of Testing and Evaluation was established, which aimed to construct a single test for entrance to university. The test they developed is called the Psychometric Entrance Test (PET).

In Israel, universities determine their own admissions policies, although all major institutions routinely require students to take the PET. Students apply to a department within a university, and selection is typically made on the basis of a composite score from the Bagrut and the PET. Generally, candidates are placed in rank order according to their composite score, and cut-off points are then established according to the ability of the applicants and the availability of places. For some courses, minimum cut-off points may be set regardless of the quota to be selected, in order to ensure that students meet basic requirements. Additional selection methods such as interviews and other proficiency certificates are also used for a limited number of courses (e.g. medicine, music).

2.3.1 The Psychometric Entrance Test (PET)

The PET 'measures various cognitive and scholastic abilities to estimate future success in academic studies' (Beller, 1994, p. 13). Since 1990, the test has consisted of three sections: verbal reasoning, quantitative reasoning and English.

- Verbal reasoning: 60 items assess verbal skills and abilities seen as necessary for academic studies. Question types include antonyms, analogies, sentence completions, logic and reading comprehension.

- Quantitative reasoning: 50 items look at ability to use numbers and mathematical concepts. Question types include algebraic problems and equations, and geometric problems. Only basic knowledge of maths is needed to solve these questions and any explanations or formulae that may be needed are given in the test booklet.

- English: 54 items assess command of English considered necessary for reading academic texts. The questions include sentence completions, restatements and reading comprehension. Scores from this part of the test can be used for the placement of students in remedial English classes.

All items are multiple-choice and the weighting is 40, 40 and 20 per cent for the three test sections respectively. Test results are reported as a scaled score with a mean of 500, a standard deviation of 100, and a range from 200 to 800. Beller (1994) reported median internal consistencies for the verbal reasoning, quantitative reasoning and English tests of 0.89, 0.90 and 0.93 respectively, with a corresponding figure of 0.95 for the total score. The PET correlates highly with the SAT (0.82), particularly the maths section (0.85), but correlations between the PET and the Bagrut are much lower, typically around 0.45.

Currently the PET is translated into Arabic, Russian, English, French and Spanish, with the exception of the English section, as it is considered that the problems in equating are far less than the difficulties that would result from asking students to take the test in a second language. Those who take one of the translated versions are also required to take a Hebrew proficiency test.

Jones (1994) observed that the PET shares many common features with the SAT, particularly in terms of factor structure. Factor analysis, a statistical method of reducing a large number of items down to their underlying constructs, has revealed that the PET assesses two factors, which correlate with SAT verbal and math. These factors have also been linked to fluid (flexible problem-solving skills) and crystallised (accumulated knowledge) abilities. PET scores have also been seen to differ according to intended major, as have scores on the SAT, with natural sciences and engineering students obtaining higher scores than those taking education, nursing, or social work courses. PET scores generally predict first-year degree performance better than high school grades, whereas the opposite pattern has been seen with the SAT. Jones (1994) has argued that this may be due to the closer proximity of the PET to college entrance.

2.4 Aptitude testing in Sweden

The SweSAT was first introduced in 1977. Prior to 1977, the upper secondary leaving certificate was the main way in which students were selected for higher education. The SweSAT was introduced partially because older students were increasingly being admitted to

certain university programmes, often on the basis of work experience. It was offered as a way of standardising the assessment of older applicants, but was also developed to encourage more older students to apply to university.

At the time the SweSAT was introduced, it could only be taken by people who were aged 25 or over who had at least four years' work experience. In 1991, access to the SweSAT was changed, allowing all students who wanted to apply to university to take the test either in an autumn or spring sitting. One reason for extending use of the SweSAT was the belief that it would show smaller socio-economic differences than the school leaving certificates. Reuterberg (1998) reported that the greater use of the SweSAT has been accompanied by far greater interest in it, particularly in terms of its adequacy as an assessment instrument.

University applicants are judged on the most favourable of their SweSAT scores or upper secondary school GPA. Places are allocated on a quota system, with about two-thirds of students being selected on the basis of their school GPA and a third on their SweSAT scores (Henriksson and Wolming, 1998). More recently the admissions process in Sweden has been decentralised and universities are now allowed to design their own admissions systems, although most still use the established system.

2.4.1 The SweSAT

The SweSAT is composed of six multiple-choice subtests, each of which is timed. The range of subtests has been designed in an attempt to cover the demands of different courses. The six subtests are:

- vocabulary - 30 items in 15 minutes;
- reading comprehension - 24 items in 60 minutes;
- English reading comprehension - 24 items in 50 minutes;
- data sufficiency - 20 items in 45 minutes;
- interpretation of diagrams, tables and maps - 20 items in 55 minutes;
- general information - 30 items in 25 minutes.

When the SweSAT was first introduced in 1977, it included a section designed to measure complex study skills, which required test takers to find necessary information using sources such as indexes. This was replaced by the English reading comprehension subtest in 1991.

In total the test lasts for four hours ten minutes, and the student's total score is the number of questions they answer correctly. The SweSAT is not generally speeded, but some older candidates and some females have reported concerns over the time limits (Wedman, 1994). Stage (1992, cited in Wolming, 1999) reported a correlation of 0.51 between SweSAT and GPA, showing that the two were measuring relatively distinct constructs.

As with the Israeli PET, a number of similarities can be drawn between the SweSAT and the SAT, including the emergence of two factors from factor analyses which correspond to verbal and mathematical reasoning (Jones, 1994). Males also tend to score higher on the SweSAT than females, as they do on the SAT.

2.5 Aptitude testing in Singapore

Singapore has a national assessment system based on O-levels and A-levels. Admission to university is currently largely dependent on A-level attainment, but changes to this system are planned. Details of these changes were taken from the Singapore Ministry of Education website (Ministry of Education, 2000).

In 1998, the Government set up a committee to explore possible developments in the university admissions process, which reported its findings in 1999 (Ministry of Education Committee on University Admission System, 1999). The committee recommended retaining the use of A-levels as the major component in university entrance decisions, as these were considered to provide an 'effective measure of content knowledge'. It was also considered that these exams involved considerable problem-solving and analytical skills, and that preparation for the exam involved 'perseverance and discipline in students'.

The major change to the current system recommended by the committee involved the introduction of an aptitude test. It was argued that such a test would be able to assess students' analytical thinking skills, in the absence of subject-specific knowledge. This ability was also related to the broader goal of selecting into higher education those individuals who would be most able to adapt to the challenges faced by Singapore society in the future.

The use of additional indicators of suitability for higher education was also suggested, including an assessment of project work and co-curricular activities. Project work was proposed as it would allow students to develop and demonstrate qualities like creativity and resourcefulness. Co-curricular activities (e.g. participation in sporting and artistic activities) were seen to provide information that supplements cognitive indicators obtained from tests and exams, by assessing qualities such as leadership and teamwork. The consideration of a range of factors in determining university admission was seen as being in accordance with the emphasis on a 'holistic' approach to the development of young people in Singapore.

The Ministry of Education in Singapore have decided that the SAT, as used in the United States, will be the reasoning test used as part of the admissions process. The Ministry of Education has not precluded developing their own reasoning test, but due to the time this could take have decided to use the SAT initially. The SAT will first be used in 2003. Precise details of how the test will be administered were still being worked out at the time of writing this report.

For the majority of university applicants – those with A-levels – it is proposed that A-level grades contribute 65 per cent towards the application, the SAT II5 per cent and project work ten per cent. The co-curricular activities can also be included, adding a bonus of up to five per cent to a student's overall admission score. Slightly different models have also been proposed for students who have come through different educational routes, mature students and those with international qualifications.

An interesting feature of the proposed system is that, even before it has been implemented, it acknowledges that flexibility will be needed if it is to accommodate all students. This flexibility is partly highlighted in the differing routes proposed for different students (e.g. mature students). It has also been stated that some faculties may want to use additional assessments, and some students who show particular aptitude in certain subjects be offered direct admission to university.

The planned changes in Singapore are particularly relevant, due to both A-levels and SAT scores being used together for university admissions – as might happen in Britain if aptitude testing was introduced. Unfortunately, it will be a number of years before evidence on how these two assessments function together is available.

3 Predicting college performance

3.1 Introduction

This section examines the ability of tests such as the SAT to predict college performance, and whether they do this fairly for all groups of test takers. Three strands of evidence will be considered in order to address this issue. First, the association between aptitude tests and college performance will be examined, as will the extent to which aptitude tests provide information in addition to that already available from high school records. Second, probably the most controversial issue surrounding aptitude tests is the score differences observed between different social and ethnic groups. Evidence will therefore be reviewed in an attempt to determine whether these score differences are a fair reflection of students' academic aptitude, or whether they result from deficiencies with the tests themselves. Finally, these two strands of evidence will be brought together by looking at the accuracy with which SAT scores and high school record predict college attainment for different groups.

3.2 The predictive validity of the SAT and other indices of performance

3.2.1 Issues to consider in assessing predictive validity

Although the SAT has been relabelled as an 'assessment' test rather than a test of 'aptitude', it is still generally conceived as a test of academic aptitude. Implicit within the term 'aptitude' is the concept of predictive validity. The purpose of an aptitude test is to measure an individual's potential for obtaining a certain goal. In the case of the present review, the goal is successful completion of a course at college or university. If a high proportion of applicants who score well on a certain test go on to successfully complete their degrees, and those who score lower are somewhat less likely to be successful, we would say that the test has predictive validity.

In considering an aptitude test, it is never a simple matter to declare it valid or not. It is also necessary to look at the strength of its predictive validity, and to consider whether its predictive ability can be generalised across a range of situations (in the present case across different students, courses and colleges). Estimates of predictive validity are generally based on correlation coefficients or variations of these. Correlation coefficients indicate the extent of the association between two variables. A correlation of zero indicates that two variables are unrelated to each other and a correlation of one indicates a perfect, linear association between them. The stronger the correlation between an aptitude or attainment test and an outcome measure, the better its predictive validity is said to be. The outcome measures which have most frequently been used in this work are first-year, or freshman, grade point average (FGPA), and cumulative grade point average (CGPA), which in most cases reflects final degree attainment.

Evidence for the predictive validity of aptitude tests used for college admission is presented below. When interpreting this, a number of points need to be borne in mind. Most importantly, it is not sufficient to consider the simple association between aptitude test scores and outcome measures. Any statistics need to be presented in the context of other information that is readily available when admissions decisions are made. In the United States, where the majority of the work has been conducted, this information usually takes the form of HSCR and high school grade point average (HSGPA). Aptitude tests will only be useful if they can tell us something about students in addition to this information which is routinely available.

3.2.2 Evidence on predictive validity

The most recent data from the College Board looked at the ability of the SAT and HSGPA to predict FGPA in over 48,000 students from 23 colleges. Across all colleges studied, the association between the SAT and FGPA was 0.35 (Bridgeman *et al.*, 2000). On average, the SAT was therefore able to account for just over 12 per cent of the variation in first-years' performance at college. The figure for HSGPA was comparable to that for the SAT, being 0.36. These summary statistics mask a number of variations by sex and ethnic group, as can be seen in Table 3.1.

Table 3.1: Correlations of first-year GPA with SAT and HSGPA

	African American		Asian American		Hispanic/Latino		White	
	Males	Females	Males	Females	Males	Females	Males	Females
SAT verbal	0.23	0.29	0.24	0.26	0.19	0.29	0.28	0.30
SAT math	0.30	0.34	0.32	0.32	0.19	0.31	0.30	0.31
SAT total	0.34	0.37	0.36	0.37	0.34	0.37	0.33	0.35
HSGPA	0.34	0.29	0.28	0.26	0.30	0.29	0.38	0.34

SAT verbal scores were most closely associated with FGPA in African American females and Hispanic/Latino females, whereas they showed least association for Hispanic/Latino males. Math scores also showed the lowest association with FGPA for this group, and the highest for African American females. When SAT verbal and math scores were combined, the variation in associations with FGPA becomes less pronounced for the different groups, suggesting an averaging effect. HSGPA also showed a modest variation between groups, with this being most closely associated with FGPA for White males, and the lowest association for Asian American females.

Bridgeman *et al.* (2000) also presented separate data for the 23 colleges whose results were summarised in their report. This showed that the association between the SAT and FGPA varied between colleges from a high of 0.72 to a low of 0.37. Although this highlights the variation that can exist between colleges, these values are not directly comparable to the other correlations presented by these authors. This is due to the figures given for each college being adjusted to account for range restriction (i.e. a statistical adjustment for one or more of the factors having a limited range of values, due to the group under consideration having been selected on one or more of these), whereas the summary statistics and those presented in Table 3.1 are not. Variation in college-level correlations for HSGPA ranged from a high of 0.68 to a low of 0.44, with these again being corrected for range restriction.

Studies conducted independently of the College Board have revealed a much more variable picture in terms of the predictive validity of the SAT. The associations between SAT scores and GPA for different ethnic groups have been an area of particular interest, and examples of this research are given below. It should be noted that as these have not been corrected for range restriction, the correlations they report are directly comparable to the figures given in Table 3.1.

Fuertes *et al.* (1994) found SAT verbal scores to correlate with CGPA between 0.15 and 0.22 for Asian American students, and math scores to correlate 0.31 to 0.38. When the same associations were examined for Hispanic students, these ranged from 0.20 to 0.40 for SAT verbal scores and 0.22 to 0.34 for math. Lawlor *et al.* (1997) found that SAT math correlated 0.14 with CGPA in European Americans and 0.12 in African Americans, with the verbal score correlating 0.33 and 0.61 in these groups respectively.

Data from a sample of predominantly White students at the University of Pennsylvania has been presented by Baron and Norman (1992). Total SAT score correlated 0.26 with FGPA and 0.20 with CGPA after three or four years of study. The predictive power of HSCR and highest three achievement test scores, which included a test of English, were also studied. Achievement tests correlated 0.33 and 0.26 with FGPA and CGPA after three or four years respectively, with these figures for HSCR being 0.34 and 0.31.

As African American students are known to score lower on tests such as the SAT and ACT (e.g. Bridgeman *et al.*, 2000), the predictive validity of these tests has been particularly closely scrutinised for this group. Fleming and Garcia (1998) have provided a recent review of work in this area. The average predictive validity of the SAT for Whites was seen to be 0.34, accounting for just under 12 per cent of variation in college grades and comparable to the figures from the College Board data (Bridgeman *et al.*, 2000). The average correlation for Black students was 0.31, showing that SAT scores accounted for almost ten per cent of the variance in their college grades. However, Fleming and Garcia observed that predictions for Blacks' grades were far more variable than for Whites, with figures ranging from -0.01 to 0.48. This indicates that in some cases, higher SAT scores may even correspond to marginally lower college grades, and at best they account for 23 per cent of the variance in grades. They also noted that the reasons for this variation have not been adequately explained.

From their own research, Fleming and Garcia have shown that the predictive validity of the SAT was slightly higher for Blacks in predominantly Black colleges, than Blacks in predominantly White colleges. Although these differences were small, a number of more significant sex differences were seen. For example, SAT scores were generally more predictive for Black freshmen males in Black colleges, whereas for Blacks in predominantly White colleges a very variable pattern of associations were seen, including negative, zero and positive correlations. Contrary to this, the predictive validity for Black freshmen females was higher in predominantly White colleges. Patterns of prediction were also seen to vary as the students progressed through college. In summarising this work, Fleming and Garcia argued that sex and year of study are the factors that account for the most variation in GPA, not type of college. This, coupled with the significant number of students who did not complete their courses, suggests that the extent to which different groups are able to adjust to their college environment may underlie much of the variation in prediction.

A further factor which has been seen to affect the predictive validity of the SAT is students' age. For example, Moffatt (1993) found that in people who took the test before they were 30, the correlation of SAT verbal and math with CGPA was 0.50 and 0.47 respectively. In those who took the test after 30 years of age, correlations were 0.31 and 0.15. Zeidner (1987) previously reported comparable results for the PET which is used for university admissions in Israel, concluding that test scores showed least validity as predictors of FGPA for students aged over 30.

The studies reviewed above, which have focused predominantly on SAT data, have shown that its predictive validity can vary according to a range of factors. One further line of research in this area has concerned the extent to which the predictive validity of the SAT and high school grades have changed over time. For example, Fincher (1990) examined data from the University of Georgia from between 1960 and 1985, to identify changes in the predictive validity of the SAT and HSGPA. In 1960, HSGPA correlated 0.53 with FGPA, compared with a figure of 0.51 in 1985. Far greater falls in the predictive validity of the SAT were seen across this time, with figures for SAT verbal being 0.43 in 1960 and 0.36 in 1985, with the corresponding scores for math being 0.41 and 0.32. The decreases for the SAT were seen to be most significant over the six to seven years prior to 1985, and it was proposed that this is linked to changes in university policies, including the introduction of remedial teaching programmes and the inclusion of non-traditional students, such as older applicants.

Similar findings were reported by Morgan (1990), who studied the predictive validity of the SAT, achievement tests and HSGPA in almost 300,000 students from 198 colleges. The data was taken from classes enrolling in 1978, 1981 and 1985. The correlations for SAT verbal declined from 0.36 in 1978 to 0.33 and 0.32 in 1981 and 1985 respectively, with these figures for SAT math being 0.37, 0.33 and 0.32. The predictive validity of HSGPA fell from 0.46 to 0.44 and 0.43, with similar figures being observed for achievement tests. When split by sex and ethnic background, all correlation estimates were higher for females than males, and females showed less of a decline in the predictive validity of the SAT and HSGPA than males. Ethnic group analyses showed no noticeable drop for prediction of Asian-American or Black students' grades, and prediction actually rose for Hispanic students over this time.

Overall, Morgan concluded that the decline in predictive validity of the SAT and other frequently used predictors could be seen for most freshmen, although less change in predictive validity occurred for students in the top third of their college classes. Morgan, in accordance with Fincher (1990), argued that this may be due to colleges taking steps to reduce student failure. A further study published by Schurr *et al.* (1990), but using data up to 1987, also found that the predictive power of the SAT had declined over time. In 6,278 freshmen, SAT verbal scores predicted FGPA 0.40 in 1983 and 0.38 in 1987, with the figures for math being 0.46 and 0.37.

The data reviewed above has all come from the United States, although similar research has been conducted in other countries which use aptitude tests for university entrance, notably Sweden and Israel. In a review of work on the Israeli PET, Beller (1994) reported

this to be a good predictor of FGPA, with the average correlations from 705 validity studies being 0.53 for liberal arts, 0.5 for sciences, 0.45 for social sciences, and 0.43 for engineering. The maths and verbal reasoning subtests of the PET were seen to make the greatest contribution to the prediction of FGPA across study disciplines (other components of the PET are general knowledge, figural reasoning and English). Similar results were found for the prediction of GPA at the end of undergraduate studies.

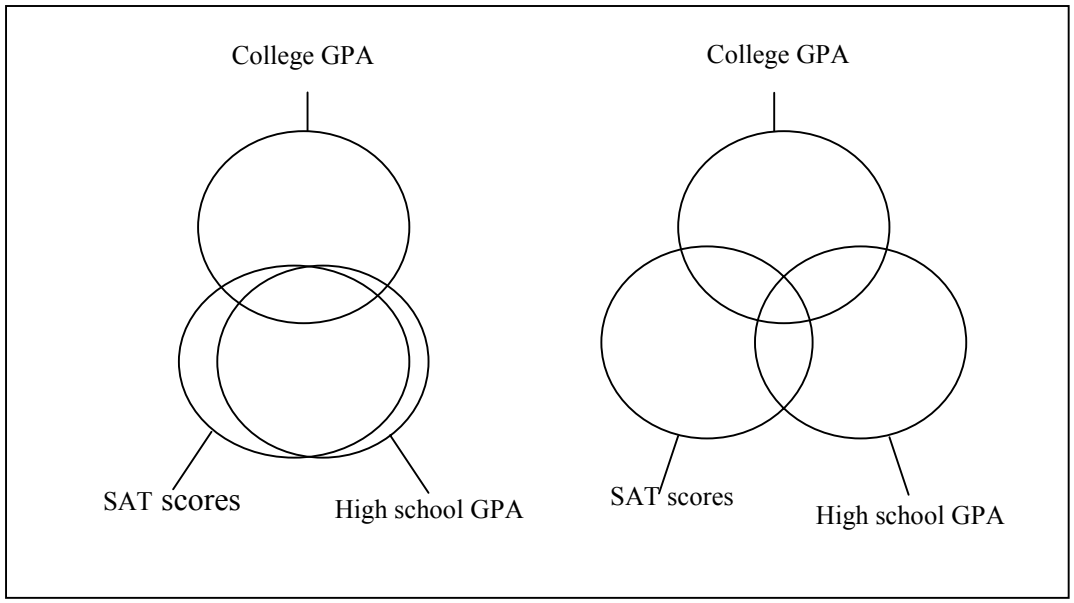
Across all fields of study, it was found that PET scores were correlated more highly with college grades than matriculation scores (overall validity 0.38 and 0.32 respectively). This was highlighted as being contrary to many studies from the United States, where HSGPA or HSCR had been identified as the best predictors of college success. Beller argued this may be due to Israeli students entering college two to five years after graduating from high school, because they are required to complete military service.

3.2.3 Do aptitude tests tell us any more than we already know?

Scores from tests such as the SAT have been generally observed to predict achievement in college to about the same extent, or slightly less than, other information on students such as high school grades and class rank. Although the evidence reviewed above shows that prediction, particularly by SAT scores, can be very variable, it is still sufficient to suggest that the SAT may be of value to admissions tutors when allocating college places. However, in order to determine whether the SAT is of real value, we need to understand what it tells us about each student's potential in addition to what is already known about them.

If Bridgeman *et al.*'s (2000) figures are accepted, showing that on average both the SAT and HSGPA accounted for about 12 per cent in the variation of FGPA, we need to know if these accounted for approximately the same 12 per cent in GPA or whether each accounted for a unique part of this. This question is illustrated graphically in Figure 3.1. In the left-hand figure, there is a considerable overlap between the variance in college GPA accounted for by the SAT and HSGPA, whereas in the right-hand one the overlap is much smaller and the unique variance attributable to each correspondingly higher. If the left-hand figure is an accurate representation of the link between the SAT and HSGPA, then the SAT is effectively redundant. Alternatively, if the right-hand figure is a closer representation of their association, then the information provided by the SAT will be of far more value.

Figure 3.1: Possible relationships between the SAT, HSGPA and college GPA



The evidence reviewed above has considered single predictors of college GPA (i.e. SAT scores, HSGPA, HSCR) to illustrate the variation in associations. In order to identify what SAT scores can tell us in addition to routinely available information such as HSGPA and HSCR, it is necessary to consider more than one predictor at the same time, using methods such as multiple regression. An overview of studies which have done this is given below.

Generally, the SAT has been found to account for a modest amount of variance in performance, after the information regularly available on students has been taken into account. For example, Bridgeman *et al.* (2000) found HSGPA to predict 13 per cent of the variance in FGPA. Adding the SAT to this increased prediction to 19.4 per cent across all students. The extent to which the SAT provides additional information to HSGPA was seen to vary according to ethnic group and sex, and the variance initially accounted for by HSGPA. Table 3.2 shows that the SAT added most to the prediction for Asian American females (11.7 per cent) and least for White males (5 per cent).

Table 3.2: Percentage of variance in first-year GPA accounted for by HSGPA and SAT

	African American		Asian American		Hispanic/Latino		White	
	Males	Females	Males	Females	Males	Females	Males	Females
HSGPA	11.6	8.4	7.8	6.8	9.0	8.4	14.4	11.6
HSGPA + SAT	20.3	19.4	19.4	18.5	14.4	19.4	19.4	18.5

Bridgeman *et al.*'s work indicated that the SAT can be a useful addition in predicting college attainment, but others have previously questioned this. For example, a study of almost 4,000 students at the University of Pennsylvania by Baron and Norman (1992) found that the SAT accounted for an additional two per cent of variation in CGPA over HSCR, but added nothing to prediction after the average performance in each student's three best achievement tests had been taken into account. Together HSCR and average achievement test scores were the best predictors, accounting for 13.6 per cent of the variance in CGPA. The authors concluded that when achievement tests are available, the SAT has very little predictive validity even at very selective institutions. As many colleges require or at least recommend that students take one or more of the SAT II: Subject Tests, the validity of the SAT I: Reasoning Test over these clearly needs further study.

Fincher (1990) provided evidence that the predictive power of SAT scores and HSGPA had fallen over time. Correspondingly, the ability of SAT scores to account for FGPA in addition to that of high school grades also appears to have fallen. Fincher reported that adding the SAT to HSGPA raised the prediction from 26 per cent to 34.8 per cent in the 1970s compared to a rise from 27 per cent to 32.5 per cent in the 1980s. Whereas the predictive validity of HSGPA increased slightly when the figures for these decades were examined, the corresponding values for the SAT have decreased.

Tests such as the SAT do appear to increase prediction of students' college grades, but this increase is modest and can depend on what other predictors are being considered. Research also indicates that the general ability of all information to predict college success fell during the 1970s and 1980s. This has been put down to colleges providing greater assistance to

students who need support in their academic work (e.g. Fincher, 1990). Providing this support to students who may otherwise have failed to complete a course or obtained a low GPA effectively seeks to break the association between poorer SAT or high school performance and subsequent poor college performance.

In determining the worth of the SAT, Crouse and Trusheim (1991) have criticised the validity service provided by the College Board, which gives an individual analysis for each college of the link between students' high school records and SAT scores, and subsequent college performance. They argued that the reports this service produced overestimated the role of the SAT by not showing the predictive statistics for GPA alone. Overall, predictive statistics showed that adding the SAT to the high school record improved a college's estimates of students' academic performance, with predicted college grades being a fair approximation of actual attainment. However, they did not show how much less accurate the grades would have been if only high school record was used. Crouse and Trusheim's own data suggested the SAT made little difference to the prediction of college attainment, once high school record had been taken into account.

Despite this apparent redundancy, Ben-Shakhar *et al.* (1996) argued that traditional validity estimates may not be the best way of evaluating tests such as the SAT. Instead, they argue it is necessary to consider the impact of the scores on the actual admission process, and the costs and payoffs of using them. In a study of successful and unsuccessful applications to a liberal arts faculty in Israel, the effects of the PET were seen to be highest when a 'strong' model of success was adopted - that is when high college achievement was the outcome. As weaker models of achievement were adopted, that is less emphasis was placed on students attaining top grades but just successfully completing a college course, the value of the PET became less.

On the basis of these results Ben-Shakhar *et al.* argued that if the primary goal is to produce excellent students, the use of the PET is justified. If the purpose of the admissions process is to provide general access to college, then high school grades alone are adequate. This corresponds to evidence from the United States, where the incremental validity of the SAT tends to be higher in more selective, high-attaining colleges. It was argued that these findings result from the nature of tests such as the SAT and PET, as these are designed to measure academic potential rather than likelihood of actually completing a college course. They are less successful in predicting retention as failure to complete a course is often not determined by academic reasons (e.g. Choppin and Orr, 1976).

3.3 Is aptitude testing fair?

The evidence reviewed above shows that the SAT can indicate potential to succeed in college, although once high school record has been allowed for, its predictive validity is often quite modest. Considerable variation in the predictive validity of tests such as the SAT have also been seen, particularly for different ethnic groups. This leads us to probably the most controversial issue surrounding aptitude tests: whether they assess aptitude fairly for all social and ethnic groups. That is, whether differences observed in test scores between groups are an

accurate reflection of their chances of succeeding in subsequent education, or whether they result from limitations of the test itself and so do not reflect genuine educational potential.

Test fairness or bias can been examined at two levels - scores on the overall test and responses to individual items or questions. Evidence for test- and item-level bias is presented below, and possible reasons for the observed differences explored.

3.3.1 The concept of test bias

It is easy to argue that tests are biased, but far more difficult to prove that they are not. Test takers can be grouped in a vast number of ways (e.g. sex, ethnic status, socio-economic status, geographical area, education), and the chances are that differences in average test scores will be observed between some of these groups. Such differences do not prove that a test is biased, although they are often taken as evidence of this. Equally, finding no differences in test scores between groups could also indicate bias, although this situation is far less likely to be interpreted so.

The difficulty in identifying bias stems from the fact that we have no objective measure against which tests can be compared. For most physical properties we are able to take objective measurements. For example, if we measured the height of a group of males and a group of females and found that, on average, males were taller, would we say that our method of measurement was biased? Probably not, as there are commonly accepted scales against which height can be measured and recognised units for its measurement. This would lead us to conclude that the observed differences in height reflected real differences between these groups, and were not a result of biased measurement.

In assessing mental abilities, difficulties arise because, unlike in most physical sciences, the constructs of interest are not directly observable, but have to be inferred from the measurements that are made. From performance on a test of verbal reasoning, we infer a person's verbal reasoning ability. But how do we know that this test is an adequate measure of verbal reasoning? Usually this question is answered through expert judgements on the test coverage and its associations with existing tests which purport to measure the same construct (convergent validity). Judgements are also made on factors such as the content of the test and the language it uses to determine whether this may unintentionally have made the test easier or harder for some groups of test takers. However, unlike the example of height given above, there is no absolute standard against which we can evaluate our test or the results obtained from it. Therefore, it is not possible to conclusively say whether group differences in scores on the verbal reasoning test arise due to limitations of the test or real differences in the verbal reasoning ability of these groups.

Despite this, it is possible to look for evidence from a number of sources so as to make an informed judgement about whether a test is biased or not. This can be done firstly through determining whether scores on aptitude tests, and the educational outcomes they are used to predict, vary together. The associations between aptitude test scores and scores on tests designed to measure similar constructs can also be examined, again to determine whether the two co-vary. A further method is to take into account group factors known to influence test scores, and to see how scores compare between groups after these factors have been allowed for.

The following section will summarise the work on how SAT and ACT scores vary between different groups, before looking at whether this may reflect real differences between test takers or suggest that tests are biased.

3.3.2 Evidence of score differences

The College Board publishes details of the SAT scores each year. Table 3.3 presents the data for students who graduated from high school in 2000 (College Entrance Examination Board, 2000).

Table 3.3 shows that male and female high school graduates in 2000 had comparable scores on the verbal section of the SAT, but males performed better on the math section by almost a third of a standard deviation. Comparisons between ethnic groups show that Whites outperformed all other groups on the verbal section of the SAT. This can be explained through the SAT being a test of developed reasoning abilities in English, the first language of most White test takers. In addition to the high scores of Whites, the lower scores of African Americans were also noticeable, with there being almost a full standard deviation between the average scores of these groups. Low verbal scores were also seen for Mexican Americans, Puerto Ricans and Hispanics/Latinos.

Table 3.3: 2000 average SAT scores by sex and ethnic status

	SAT Verbal	SAT Math
All test takers[1]	505	514
	(111)	(113)
Males	507	533
Females	504	498
American Indian, Alaskan Native	482	481
Asian, Asian American, Pacific Islander	499	565
African American/Black	434	426
Mexican American	453	460
Puerto Rican	456	451
Hispanic/Latino	461	467
White	528	530

[1] Figures in parentheses show population standard deviation

The highest average score on the math section of the SAT was achieved by students in the ethnic group labelled Asians, Asian Americans and Pacific Islanders. These students had an average score almost a third of a standard deviation higher than Whites, and a full standard deviation higher than African Americans and Puerto Ricans. Figures presented by the College Board also show that both verbal and math scores were positively associated with family income and parental education.

In the report of the 2000 SAT, comparisons were also made with the results obtained from a decade before in 1990. Average verbal and math scores in 1990 were 505 and 521 for males respectively, and 496 and 483 for females. A comparison between these figures and those in Table 3.3 shows that the differences in verbal scores between males and females declined over this time from nine to three scale points. Relatively, the difference in average math scores declined far less, from 38 points in 1990 to 35 points in 2000.

Changes in scores over the ten-year period can also be examined for ethnic groups. For the verbal section of the SAT, average increase across all groups was five points. The largest increases were seen for Puerto Ricans (21 points), American Indians and Alaskan Natives (16 points) and Asians, Asian Americans and Pacific Islanders (16 points). Below average increases were seen for Hispanics/Latinos (2 points) and the scores of Mexican Americans fell by four points over this decade. The average score for African Americans rose six points over this time, with Whites' rising nine points.

A similar pattern was seen for average math scores, with Asians, Asian Americans and Pacific Islanders (19 points), Whites (15 points) and Puerto Ricans (14 points) showing score increases above the average of 13 points. Mexican Americans (no change), Hispanics/Latinos (3 points) and African Americans (7 points) had score changes less than the average increase. It should be noted that over this decade, the SAT underwent a considerable revision (see Section 2.2.1). What effects, if any, this revision could have had on the score changes discussed above is unknown, even though scores for the two time-points are given on the same scale.

Although the ACT is somewhat different from the SAT in its format and what it claims to measure, it is used in very similar ways for college entrance, and so score differences on this are also examined. Table 3.4 shows the mean scores on the subtests and composite scores for all students who took the ACT in 2000, and these scores broken down by sex and ethnic status (American College Testing Program, 2000a).

Looking at sex differences first, it can be seen that males had a slightly higher composite score on the ACT than females. Males scored more highly than females in the mathematics and science reasoning subtests, whereas females had higher scores in English and reading. In terms of ethnic groups, Caucasians and Asian Americans clearly outperformed all other ethnic groups both on the composite score and on each of the subtests. Caucasians tended to perform slightly better than Asian Americans, although Asian Americans were somewhat at an advantage on the mathematics subtest.

Table 3.4: 2000 average ACT scores by sex and ethnic status

	English	Mathematics	Reading	Science Reasoning	Composite Score
All test takers[1]	20.5	20.7	21.4	21.0	21.0
	(5.5)	(5.0)	(6.1)	(4.5)	(4.7)
Males	20.0	21.4	21.2	21.6	21.2
Females	20.9	20.2	21.5	20.6	20.9
African American	17.4	17.6	17.8	17.9	17.8
American Indian	19.7	20.0	20.9	20.6	20.4
Caucasian	22.3	22.4	23.1	22.5	22.7
Mexican American	18.6	19.6	19.7	19.6	19.5
Asian American	21.3	23.9	22.0	22.0	22.4
Puerto Rican/ Hispanic	19.8	20.6	20.9	20.4	20.5

[1] Figures in parentheses show population standard deviation

The ethnic group that probably stands out more than any other in Table 3.4 is African Americans. This group had the lowest composite score, and also had lower scores on each of the subtests than all other ethnic groups. In terms of the population standard deviations, the average score for African Americans was over a full standard deviation below that of Caucasians and almost as much below that of Asian Americans.

Despite the differences in format between the SAT and the ACT, a comparison of average scores by background variables reveals remarkably similar patterns. Firstly, males were seen to have higher math scores than females on both tests, with this pattern being reversed for the verbal, English and reading tests. Overall, males had slightly higher average scores on the ACT and SAT.

In terms of ethnic background, Whites tended to outperform all other groups. The exception to this was in math where Asians, Asian Americans and Pacific Islanders had the highest scores on both the SAT and ACT. African Americans performed very poorly on both tests, having the lowest average scores of any ethnic group on both SAT sections and all of the ACT subtests. Overall, scores from both tests placed ethnic groups in approximately the same rank order.

3.3.3 SAT scores, academic attainment and scores on tests measuring similar constructs

Due to there being no objective standard against which scores from tests such as the SAT or ACT can be compared, it is difficult to determine conclusively whether the observed score differences reflect bias in the tests or real differences in the ability or attainment of these groups. In the absence of a clear method of answering this question, it is necessary to consider information from a number of sources in an attempt to reach an informed judgement. The majority of this information comes from work with the SAT, although as the SAT correlates highly with tests such as the ACT and Israeli PET (e.g. Beller, 1994; Schneider and Dorans, 1999), similar arguments are likely to apply to these.

One way in which the issue of bias can be addressed is by looking at how scores on tests such as the SAT relate to other information on academic attainment. This can be done by comparing school or college performance between groups of students, after matching their SAT scores. Alternatively, attainment can be matched and SAT scores examined. If both groups show similar attainment when matched on SAT scores, this might suggest that the test is not biased in favour of one group over the other in terms of predicting academic performance. However, this does not preclude the possibility that both assessments are biased in the same way.

Bridgeman and Wendler (1991) reviewed gender differences in SAT and ACT math scores, and found that males tended to outperform females by between 0.3 and 0.4 of a standard deviation. However, this difference was not reflected in general college attainment, which tended to be equal for males and females, or where differences did occur, females showed higher attainment. Bridgeman and Wendler (1991) argued that these findings could be due to differences in the grading of courses selected by males and females (i.e. males generally selected harder courses), and set out to test this by studying algebra, calculus and pre-calculus courses in nine universities. Results suggested that differences in attainment scores were not the result of course selection patterns, as even within specific courses females had comparable grades to males, despite males having higher SAT math scores. This finding was consistent across courses and across institutions. It was also found that students on each course tended to have similar high school experiences, showing that prior differential course taking could not account for the results.

Pearson (1993) compared the SAT scores and CGPA of 220 Hispanic and 892 non-Hispanic White students who entered the University of Miami in 1988. A survey of students confirmed that the Hispanic students came from predominantly middle-class backgrounds and had been educated in the United States, making them comparable to non-Hispanic Whites. The majority of the Hispanics were bilingual. Hispanics were seen to have verbal SAT scores on average 42 points lower than non-Hispanic Whites, and math scores 49 points lower. These differences were not reflected in the CGPAs of these groups, as these were marginally higher for Hispanics. Pearson suggested that the SAT score differences may be due to the Hispanics being bilingual and to differences in test taking strategies, as Hispanics appeared to work more slowly and more precisely.

The association between SAT scores and CGPA has also been observed to vary according to the age of the test takers. For example, Moffat (1993) found that students who took the SAT before they were 30 had higher scores on both sections, particularly the math. However, this difference did not correspond to their CGPA, as these were virtually identical for both groups. Students in this sample were also grouped according to their age when they entered college.

Similar differences in SAT scores were again seen, but here older students attained higher CGPA.

Corley *et al.* (1991) examined SAT and college attainment differences between students from urban and rural communities in a predominantly Black college. Rural students had lower SAT verbal scores than urban students, but higher GPAs in high school and college. When the samples were matched on distributions of SAT scores, the same pattern of higher grades for rural students emerged. Scores were also examined by parental income. Students from more affluent backgrounds had higher SAT verbal and math scores, but little difference was seen in the high school or college grades of these groups. These authors concluded that as differences persisted through high school and college, they were not due to differing standards between rural and urban high schools.

The evidence presented above indicates that although SAT scores may vary considerably between groups defined according to a range of factors, the same groups will not necessarily differ in academic attainment. This evidence should not be interpreted as showing that SAT scores are totally unrelated to academic attainment - the SAT is able to predict college grades at a moderate level (e.g. Bridgeman *et al.*, 2000).

An alternative indicator of possible test bias is to look at how SAT scores relate to scores from tests which claim to measure similar constructs, and the effect of background factors on this association. Evidence of this type is presented by Hyde and Linn (1988), who conducted a meta-analysis which combined the results from 165 studies of verbal ability on almost 1.5 million participants. From these studies, which had used a wide range of measures, there emerged only a very small sex difference in verbal ability, with females tending to outscore males by 0.11 of a standard deviation on average. In contrast to this trend, males tended to outperform females on the verbal section of the SAT, a finding which can still be seen in analyses of recent SAT data (e.g. see Table 3.3). Although some evidence suggests that this could result from male SAT takers being more highly selected than females, it was questioned whether this was sufficient to account for the differences observed. An alternative possibility suggested by Hyde and Linn was that the more technical nature of the material in the verbal section favoured males.

This meta-analysis also showed a decline in gender differences over recent years. Hyde and Linn argued this may have resulted from changing roles in society or from a change in publication practices, as once initial research had suggested the existence of gender differences, subsequent evidence was collected to refute this. Despite the SAT still showing male superiority on verbal reasoning ability, data from the College Board supports this continuing decline in gender differences.

A similar meta-analysis on mathematical ability was conducted by Hyde *et al.* in 1990. This analysis combined the results from 100 studies, which had collected test results from over three million individuals. Across all studies the gender difference in maths was seen to be 0.20 of a standard deviation in favour of males, with this falling to 0.15 when SAT data was excluded. When studies of general population samples were analysed, that is, selected samples were excluded, the difference was 0.05 of a standard deviation in favour of females. This finding is significant and shows that results from pre-selected samples, or samples which are self-selected in some way, may not be a true reflection of abilities in the wider population.

From this analysis it was noted that the SAT produced somewhat discrepant results, as the verbal test had in the previous meta-analysis. Overall, the difference in mathematical ability

was 0.15 standard deviations when SAT scores were excluded, but data from the SAT alone showed an effect size of 0.40 in favour of males. Hyde *et al.* suggested that the reasons behind differences in the math SAT may be similar to those on the verbal section, including the selected nature of the sample, male SAT takers being more advantaged than females, and the content of the test being of a technical nature.

The meta-analyses conducted by Hyde and colleagues on verbal and mathematical ability both suggested that the SAT assesses somewhat different constructs from other tests in these areas. These differences appear to provide evidence that the SAT is biased in favour of males. One further example of this is provided by Sheehan and Gray (1992), who looked at the association between the SAT, GPA and the Descriptive Test of Mathematics Skills (DTMS), which is also produced by ETS for the College Board. Data was obtained from almost 3,000 freshmen who took a mathematics course in a selective, private college. No sex differences were seen in the DTMS, but females had significantly higher GPA and males significantly higher SAT scores. Although these findings again do not directly show which, or even if any, of the tests were biased, it again highlights discrepancies between the SAT and other tests measuring comparable constructs.

3.3.4 Item-level analysis of SAT data

The evidence reviewed above has been concerned with test-level bias, that is, differences in the total test scores achieved by different groups. An alternative way of addressing bias is to look at performance at the level of individual test items. Item-level bias is most frequently studied through statistical methods which produce indices of differential item functioning (*dif*). *Dif* analyses compare the performance of two groups on a test item, once their overall performance on the test has been equalised. Comparisons are typically made between males and females, and between groups based on their ethnic status.

All new SAT items are pre-tested by being inserted into actual administrations of the test, so making it possible to examine *dif* under normal testing conditions. Since 1989, SATs have been constructed from item pools which have been screened for *dif* (Burton and Burton, 1993). Although all SAT items go through an extensive review process before being trialled, the purpose of the *dif* analysis is 'to identify those items that escape the conventional review process' (Freedle and Kostin, 1990, p. 329). Items which show unacceptably high *dif* are generally removed from the pool of items available for SAT construction, unless they are needed to meet the test specification for a particular version of the SAT (Burton and Burton, 1993). Once each SAT has been administered, the live data is again analysed for *dif*, to determine the effectiveness of the screening of the trialling data.

It is necessary to recognise that simply removing items which show a significant bias in favour of one group or another will not necessarily result in the two groups having identical scores on the overall test. *Dif* analyses describe the magnitude of the bias. During the development of the SAT, those items showing extreme bias are removed where possible, but this does not preclude one group obtaining a higher test score than another due to them showing moderately better performance on a number of items. *Dif* is often seen in some SAT items, and examples of this will be discussed below. However, the pre-trialling development process for the SAT appears to be relatively successful in removing biased items, as only a modest number exhibit *dif* when trialled (e.g. Burton and Burton, 1993).

The results from *dif* analyses can sometimes be very difficult to interpret - the statistics may say that an item shows significant *dif*, but it is often not apparent why this is so. *Dif* analyses

also require quite large sample sizes if they are to produce reliable results. Due to the large number of students who take the SAT, this has encouraged many researchers - not just those directly connected with the SAT - to use SAT data in an attempt to identify factors that may result in *dif*. The results from recent work in this area are summarised below.

Looking at verbal SAT items, Freedle and Kostin (1990) compared item functioning between Blacks and Whites. *Dif* was seen to interact with item difficulty, as harder items were generally answered better by Black students and easier items were answered better by White students. It was suggested that this finding may result from the different strategies used by Black and White test takers. Subsequent work by Freedle and Kostin (1997) further explored reasons for *dif* in verbal analogy items. They again replicated the finding that Black students performed better on harder items, and argued that easier items contained concepts that were differentially less familiar to Black students, reflecting their different experiences and activities.

An attempt to synthesise the literature on verbal *dif* was presented by Schmitt and Dorans (1990). They found evidence that *dif* was related to certain item characteristics that could be generalised across different ethnic groups. Items which had 'content of interest' were found to be easier for Black and Hispanic students, with these most often being seen on sentence completion and reading comprehension items. The second major finding was that verbal items which involved homographs (words with more than one meaning) were more difficult for all ethnic minority groups. A further factor that may have contributed to *dif* was speededness, as Blacks and Hispanics were less likely to reach items at the end of the verbal section.

Dif analyses of the math section of the SAT have also revealed interesting group differences. For example, Harris and Carlton (1993) found that male and female students who attained similar math scores, achieved these through different means. Males performed better on geometry and geometry/arithmetic items, whereas females performed better on arithmetic/algebra and miscellaneous items. They concluded that these findings support the view that males perform better on items that have a practical application, possibly because they see math as being more relevant to their daily lives than females.

An item-level analysis of the revised SAT which was introduced in 1994 was reported by Burton (1996). Overall, no significant *dif* was seen for males or females on the two types of math question used in the previous version of the SAT (multiple-choice 'regular math' questions and quantitative comparisons). On the new item type - student-produced responses - a slight bias was seen in favour of females, but this was quite modest. Overall, the revisions were seen to have had no noticeable difference on the male/female *dif* for the math section, although revisions appeared to have slightly reduced the *dif* on the verbal section of the SAT.

Lawrence *et al.* (1995) also examined the *dif* for males and females, and Blacks and Whites, using data from the revised SAT. The male/female comparison produced negligible *dif* for both the multiple-choice and student-produced response items. The comparisons between Blacks and Whites revealed a number of items to show *dif* in favour of Whites, particularly the student-produced response items, although none of the items exhibited extreme *dif*. On the basis of these findings, Lawrence *et al.* suggested that 'the item types measure mathematical ability somewhat differently for African American test takers and White test takers' (p. 15). However, they noted that it was not possible from their analysis to conclude whether this multidimensionality was irrelevant to the construct of interest, and so an additional source of test error.

The test-level score differences reviewed above show that efforts to screen the SAT for items which display *dif* during development have not resulted in the scores of different groups being equalised. Score differences are therefore likely to result from small differences in performance across a number of questions, rather than large differences on a few. This form of bias is far harder to detect using *dif*, and would also make test construction very difficult if more rigorous criteria for *dif* were adopted.

Overall, the process of removing biased items from the SAT appears to have had little effect on the whole test. For example, Burton and Burton (1993) compared data from before and after routine screening of SAT items for *dif*, finding performance of different groups to be 'virtually unchanged'. Although surprising, they noted that this was likely to have occurred because items which favoured the minority groups (e.g. Blacks) were also removed from the item pool, as well as those which were biased against them. They also found that screening for *dif* had little effect on item discrimination (the ability of items to distinguish between able and less able students).

Although the general view may be that removing biased items improves measurement quality, recent work by Roznowski and Reith (1999) has failed to support this idea. They found removing biased items to have little effect on predictive validity or overall measurement quality, and argued that too great a focus on removing biased items can be 'unnecessarily limiting to test development'.

3.3.5 Does the evidence suggest the SAT is biased?

It has been seen that the performance on the SAT differs by sex and ethnic group, at both the test and item level. Discrepancies have also been observed between the SAT and academic attainment, and SAT scores are also differentially related to background variables when compared with tests which measure broadly similar constructs. But does this suggest that the SAT is a biased measure of developed reasoning abilities, or do scores reflect real differences between groups of test takers?

3.3.5.1 Sex differences

Considering sex differences first, much has been written about sex differences in performance, particularly in the area of mathematical ability. This may partly reflect the considerable differences in mathematical attainment seen between males and females, particularly on tests such as the SAT. Rudisill and Morrison (1989) have argued that there are strong suggestions that physiological differences affect mathematical ability, with these resulting in the superior performance shown by males. The most conclusive evidence for this is in the area of visualisation and spatial ability. However, physiological differences cannot be conclusively seen as causal, as they may result from experience. It is also unlikely that these differences are sufficient to account for the SAT data, as visualisation and spatial ability are not major parts of the SAT math section.

Authors such as Rosser (1989) and Gallagher and De Lisi (1994) have argued that the format of tests such as the SAT may put females at a disadvantage. Rosser conducted a detailed analysis of question responses in order to determine the differential performance of males and females. It was found that more females than males left questions blank, and an even larger number omitted the last five verbal questions and the last ten math questions on the SAT that was analysed. It was suggested that this may be due to females being less likely to take risks and guess answers - the warning about the guessing penalty in the SAT instructions may be

taken more seriously by females. It was also suggested that females may have a greater problem with the time pressures on the math test than males.

Test questions can be categorised along a number of dimensions, and Gallagher and De Lisi (1994) have argued that analysing SAT math questions on this basis may provide insights into sex differences. These dimensions include computational versus word problems, algebra versus geometry, and well-defined versus more loosely defined problems. This last dichotomy was argued to come closest to the distinctions between classroom grades (based on tightly defined problems for which students have explicitly been taught solutions) and standardised test items (more loosely defined problems which students may not have previously encountered). To investigate whether this distinction could account for sex differences in actual performance on SAT math questions, think-aloud protocols were used to identify problem-solving strategies in a sample of high-ability high school students.

Females were seen to do better on conventional than unconventional problems, whereas males showed the opposite pattern. There was a substantial overlap in the problem-solving strategies used by males and females, but females tended to rely more on conventional strategies for problem-solving. The use of conventional strategies also correlated with negative attitudes such as dislike of maths and seeing it as less relevant, whereas more positive attitudes were associated with the use of unconventional strategies. These results supported the view that gender differences in test scores are at least in part due to solution strategies, as females were seen to rely more on strategies communicated to them by teachers. In a similar vein, Linn and Hyde (1989) have argued that lower female performance may be due to them being less prepared for the SAT than males.

Taken together, these arguments suggest that one way in which to improve female performance may be to encourage more flexibility in the way they approach test problems. Linn and Hyde (1989) have also highlighted the different approaches that males and females may take to the SAT, as has the work by Rosser (1989) on test strategies. If more of the questions in the ACT or SAT were in a format that favoured males, and the format was unrelated to college performance, this would suggest one way in which the tests were biased in favour of males.

Taking a broader view of sex differences, Davies and Guppy (1997) have argued that there is a large degree of self-perpetuation in the American academic system. They looked at the effects of a range of factors on educational progress, including field of academic study, mean monthly income for the different fields of study, SAT scores and selectivity of institutions. Males were more likely than females to enter high pay-off fields and selective colleges. Within the selective colleges, students from higher socio-economic backgrounds were also more likely to study areas with potentially higher earnings. Females were less likely to enter lucrative fields of study or to be studying these fields in selective schools, both of which translated into higher earnings.

From these findings, Davies and Guppy suggested that high socio-economic status may allow access to more selective colleges and, in turn, increased earning potential. The interactions between these variables are complex and yet to be fully understood. However, Waller (1971) argued that the relationship between social class and IQ was probably reciprocal – high social class may be associated with higher IQ, but higher IQs may allow people to improve their social class. This seems to correspond with Davies and Guppy's findings of self-perpetuation in college education, as does the view of Neisser *et al.* (1996) that certain jobs may affect IQ. For example, jobs that are more varied and demanding may encourage more 'intellectual

flexibility'. If, as seems reasonable to argue, these jobs also carry the greatest financial rewards, the children of these job incumbents are more likely to be successful in their education, so leading to the self-perpetuation described by Davies and Guppy. This argument appears particularly applicable to ethnic minority families, who may be disproportionately highly represented in the lower socio-economic classes. Factors that may affect the scores of these groups are now discussed.

3.3.5.2 Ethnic differences

In 1994, Herrnstein and Murray published a book called *The Bell Curve*, which argued that genetic factors accounted for much of an individual's success or failure in the United States, and accordingly for group differences on tests such as the SAT. If this is the case, it would suggest that the SAT is not biased against ethnic minority students, particularly African Americans. The controversy that followed the publication of this book led the American Psychological Association to assemble a task force to produce an authoritative paper on what was scientifically known about intelligence. The conclusions of this task force are presented by Neisser *et al.* (1996). Although this paper is concerned with intelligence and does not directly address tests such as the SAT, it is highly relevant to this discussion due to considerable evidence that '*g*', or a general intelligence factor, underlies much human performance including the SAT and ACT (e.g. Brodnick and Ree, 1995).

Addressing what was Herrnstein and Murray's most controversial conclusion, that of genetic differences, Neisser *et al.* concluded that there was little evidence to support the genetic hypothesis for the difference in IQ between Blacks and Whites. However, at the individual level, genetics could account for a substantial proportion of IQ. For example, correlations between IQ scores for identical twins raised apart have been reported in the range of 0.7 to 0.8. Correlations between unrelated, adopted children raised in the same family have been found to be virtually zero by some researchers and no higher than 0.2. Although this suggests a strong heritability factor in IQ, this does not imply that IQ is unchangeable.

The relationships between socio-economic status, IQ and schooling are complex and yet to be fully understood, but do give some clues as to the causes behind ethnic score differences. For example, parental socio-economic status has been seen to predict about one-third of children's social status and about one-fifth of the variation in their income (Jencks, 1979). About half of this effect of parental socio-economic status can be explained through it predicting children's IQ, which in turn is related to social status and income. Higher IQ has also been linked to schooling in a reciprocal relationship - childhood IQ predicts how long a person will stay in school, but longer schooling is also associated with changes in mental abilities, particularly those measured by intelligence tests. Neisser *et al.* also noted that the quality of the school is an important factor, as general skills such as problem-solving, the ability to concentrate and motivation can all be passed on through schools and have important effects on subsequent learning. Socio-economic status is therefore likely to be a contributory factor to score differences, particularly as certain ethnic minorities may be over-represented in the lower social classes in America and many other western countries.

In addressing the issue of test bias, particularly in relation to the lower scores of African Americans, Neisser *et al.* (1996) observed that from the viewpoint of simple equality, tests are biased against this group, but so are many other outcomes in American life (e.g. lower representation in highly paid professions). An alternative point of view is to acknowledge that the main function of tests is as predictors, and in this sense they predict educational attainment for Blacks as well as for Whites. As was seen earlier in this section, not all

research agrees with Neisser *et al*'s conclusions on this issue. Further evidence on predictive validity is also presented in Section 3.4.

The nature of the actual tests has also been cited as a possible source of bias. Language is one area which has been highlighted as a potential problem, as tests are usually written in a very standard form of English, and so may use phrases that Black students and those from other ethnic minorities are less familiar with. Language has been argued to be a particular barrier to Hispanics, as they often speak English poorly. Although language may play a small role in affecting test scores, attempts to allow for this have had little success in reducing the score gap. A related argument is that tests clearly reflect White values, and so Blacks may not be motivated to perform well on them.

Neisser *et al.* observed that the lower scores of ethnic minority groups on IQ tests are not unique to America. In other countries which have disadvantaged social groups, similar differences have been observed (e.g. New Zealand, India). These groups have been described as being 'caste-like' in that children born into these groups 'grow up firmly convinced that one's life will eventually be restricted to a small and poorly-rewarded set of social roles' (Neisser *et al.*, 1996, p. 94). Across the world, children who are born into these minorities tend to do less well in education and drop out earlier. The American school system has also been argued to conflict with many structures deep in the African American culture, a situation which again could hinder the educational progress of these students.

Gandara and Lopez (1998) have highlighted some particular difficulties with the SAT and ACT for Latino students. In their study of Latino students who had attained good GPAs from high school, they found that almost half of these performed poorly on the SAT or ACT. Some students removed themselves from attempts to attend competitive colleges because of their low test scores, despite their GPAs suggesting this decision was unwarranted. A further finding was that although hard work in school appeared to pay off for these students in terms of GPA, similar efforts for the entrance tests did not. This highlights the need for students and those involved in selection to more fully understand what scores on different tests indicate and, more importantly, what they really mean in terms of life chances.

One of the major points of Neisser *et al.*'s review of intelligence was that there is an emerging consensus that test scores are far too narrow a source of information on which to base the whole concept of intelligence. A very similar argument could be put forward for aptitude tests such as the SAT and ACT; the abilities they assess are far to narrow to provide an adequate indication of college performance. Work by researchers such as Gardner (1993) and Sternberg (1999) has started to recognise a broader basis of intelligence. Within this is the acknowledgement that people can behave in very 'intelligent ways', but may not able to reproduce this on paper-based tests, and also that apparently intelligent behaviour can be unrelated to traditional IQ tests. In academic terms, the narrower, test-driven concept of intelligence prevails. However, relatively little attention has been given to the consideration of what other forms of intelligence may play a role in determining academic success.

3.4 Do test scores predict fairly for all groups?

It is accepted that not all groups of students attain comparable scores on tests such as the SAT and ACT. These differences are not unique to the tests in question, as similar performance differences have been observed on a range of tests, despite sometimes overt attempts to remove them (Neisser *et al.*, 1996). However, the SAT in particular has shown some unusual patterns of associations with background factors and trends in scores over time (e.g. Hyde *et*

al., 1990). Although scores on tests used for college entrance show some ability to predict grades, the limited power of this prediction may account for the discrepancies observed between the SAT and GPA for groups of college students. Finally, the associations between SAT scores and college grades have been seen to vary according to a range of factors and, after allowing for high school records, the predictive power of the SAT is reduced considerably.

Much of this evidence is brought together when the ability of high school records and SAT scores to predict actual college grades is considered. Up to this point, group differences and the associations between SAT scores and GPA have been considered largely in isolation from each other. However, as the major goal of the admissions process is to select those students who will do well academically, it is necessary to extrapolate from information on prospective students to probable future performance. This is achieved through constructing regression lines showing the association between the entrance qualifications and subsequent attainment of students who have already attended college. These regression models can then be used with new applicants to determine what their first-year or final GPA is likely to be, given their attainment at the time of college entry.

Differential predictions for males and females have attracted considerable interest. Reviewing this work, Rosser (1989) contrasted females' consistently lower scores on the SAT with data from the College Board's validity studies, which showed female first-year college attainment to be as good as or better than that of males. It follows from this that the SAT was also generally over-predicting the performance of male students. This data suggested that 'the SAT is not fulfilling its primary purpose - to predict first-year college performance' (Rosser, 1989, p. 23). Similar evidence is available for the ACT, with males outperforming females on each of the four areas except English usage, despite females going on to earn higher GPAs in the areas covered by the ACT.

One of the arguments used to defend the differential prediction observed from the SAT is that females tend to choose easier courses which lead to higher GPAs. Studies which have controlled for grade distributions have shown this to reduce the differential prediction between males and females, but not eliminate it (e.g. Bridgeman and Wendler, 1991). Using data from over 47,000 students collected by ETS, Wainer and Steinberg (1992) set out to examine the predictive validity of the math section of the SAT for males and females. The first approach they used matched students on their attainment in college-level math courses, and then looked back at what the SAT scores of males and females who had obtained comparable grades were.

Both males and females in higher-level courses had higher SAT scores, and within these courses an association was seen between SAT scores and course grades. However, females were found to score between 21 and 55 points lower on the SAT math section than males, when matched for grades and course type. The performance of females was seen to be consistently under-predicted by SAT scores, and it was concluded that 'the SAT-M used alone, is mismeasuring the profile of proficiencies that contribute to success in college' (Wainer and Steinberg, 1992, p. 330). Although the College Board state that the SAT should never be used alone, the authors noted that some universities and scholarship programs were setting cut-scores for qualification based solely on the SAT.

Wainer and Steinberg went on to analyse the data in the opposite direction, that is to predict college math grades from SAT scores - the way SAT scores are actually used in admission. Worryingly it was found that the 'prospective estimates of the sex differences are...

considerably larger than those obtained from the retrospective analysis' (p. 329), with these again under-predicting female attainment. That is, in an actual admissions context, SAT math scores would substantially under-predict attainment on math courses for females. Although this study did not allow the underlying causes of this prediction bias to be determined, three possible causes were suggested: firstly, that there were different selection mechanisms affecting females' choice to take math courses that were not measured in the study; secondly, that the SAT math section favours males; and thirdly, that college grading practices favour females.

In reviewing research on the Israeli PET, Beller (1994) concluded that there was no consistent evidence of under-prediction for males or females, as the links between PET scores and GPA tended to be quite consistent. However, as with the SAT, where evidence of under-prediction had been seen, the suggestion was that the PET was under-predicting for females.

Whilst there is quite consistent evidence that the SAT under-predicts the college performance of females, the pattern is less clear when ethnic status is considered. A recent analysis of the predictive validity of the SAT for Black and White students has been presented by Vars and Bowen (1998). Students from the 1989 intake of 11 institutions were studied, with CGPA as the outcome measure. These institutions were categorised as being 'highly selective' and studied because, as the authors note, most controversy surrounds those institutions which have highly selective admissions criteria. Vars and Bowen found 'African-American students have lower GPAs than one would predict on the basis of their SAT scores and high school grades', with this being present for 'both males and females and... in all major fields of study' (p. 466). These authors also studied the ability of a range of socio-economic factors to account for the observed score differences, but none substantially explained the differential prediction. Rather, they suggested that it was the college experiences of Black students that accounted for these effects - experiences that may be particularly influential in highly competitive institutions.

Somewhat contrary evidence was presented by Lawlor *et al.* (1997), who studied the predictive validity of the SAT for African and White American students. The findings indicated that 'If decisions regarding college admittance and scholarships are based on SAT scores, a disproportionate advantage would be given to white relative to black students. Furthermore, if SAT scores were used to predict college graduation GPA, black students' college GPAs would be underestimated' (p. 510).

In comparing the predictive validity of the SAT and HSGPA for Black and White students, Hand and Prather (1985) studied over 45,000 students from 31 institutions. Their analyses supported the view that GPA is less predicable for Black males than the Black females and Whites. This was at least partly due to the weaker association between HSGPA, SAT verbal scores and college grades in Black males. The authors concluded that as Black males had the lowest HSGPA in this study, it was this that led to the prediction of lower grades in this group, rather than the SAT.

Comparisons of the predictive validity for other social groups have also been examined. For example, Zeidner (1987) studied the predictive validity of the Israeli PET according to students' age. The PET was found to be a less valid predictor of FGPA in older students, as it slightly under-predicted older students' performance. However, using a common regression line for all students was considered to provide an acceptable degree of predictive validity. It was argued that the extended time since these students had been in formal education, and their different educational experiences and motivational factors, may explain these differences,

although as the study was cross-sectional, differences between the cohorts could also have accounted for the findings.

Various adaptations to the administration of the SAT can be made to accommodate the needs of test takers with disabilities. The effects of these on the predictive validity of the SAT have been studied by Braun *et al.* (1986) and Ragosta *et al.* (1991). These studies looked at the SAT under standard and modified testing conditions with students who had learning difficulties and visual, physical or hearing impairments. For those with learning difficulties and physical and visual impairments, a combination of the SAT and HSGPA was judged to be a good predictor of college performance. Students with hearing impairments performed quite poorly on the SAT, and when combined with HSGPA this tended to under-predict the performance of these students. Prediction was generally seen to be not as good for handicapped compared to non-handicapped students, and special test conditions also reduced the predictive validity of the SAT.

A final study on differences in predictive validity concerns the policies of some colleges which have made the reporting of SAT scores optional. Schaffner (1985) compared the predictive validity of SAT scores according to whether students offered their SAT scores on their application forms or withheld them (SAT scores from all students were collected on acceptance to the college). No differences were seen in the prediction of GPA between those students who submitted the SAT and those who chose to withhold it. Submitters performed slightly higher than withholders, but this was to the degree that would have been expected given the differences in the SAT scores of these groups. From this Schaffner concluded that the predictive validity of the SAT appeared to be unaffected by college admission policies.

3.5. Discussion

This section has brought together research on the ability of academic aptitude tests to predict college performance, and the extent to which these tests provide fair assessments of academic potential. In terms of the second of these points, average SAT and ACT scores have been observed to differ considerably according to a range of background factors, particularly sex and ethnic status. Whether these differences reflect actual test bias or not is often a matter of judgement following the collection of relevant evidence. In terms of the higher performance of males on these tests, there is some evidence that the SAT shows discrepant results when compared with tests which measure similar constructs and college attainment. Although this difference is suggestive of bias in favour of males, the possibility that female SAT takers are less selected and less prepared for the tests than males cannot be ruled out.

The evidence on ethnic bias is less clear. Most attention has focused on the considerably lower scores of Blacks on the SAT and ACT, but this is not exclusive to these tests as Blacks score approximately one standard deviation lower than Whites on a range of intelligence tests (Neisser *et al.*, 1996). Despite overt attempts to control for aspects of the test which may cause these differences (e.g. language), score differences have remained. There is little evidence that genetic differences between Blacks and Whites are responsible for this finding, but social conditions and expectations may play a role in the lower scores of Blacks. Currently there is no wholly adequate explanation for the lower scores of Black students or those from other ethnic minorities.

This difference in test-level scores remains despite attempts to remove items that show bias towards different groups from the SAT. Since 1989, the item pools from which SATs are constructed have been screened for differential item functioning (*dif*). Items which show

significant bias have generally been excluded from SATs, but there is little evidence to show that this strategy has significantly reduced test-level score differences.

In terms of prediction, both aptitude tests and high school record are able to predict performance to a modest degree. High school record has generally been found to be a better predictor than the SAT, although some studies suggest each predicts college GPA to approximately the same degree. When considered together, the SAT tends to improve prediction over the use of high school record alone, although this improvement is variable and in some cases quite small. Much debate has surrounded the prediction of attainment for different groups, and although prediction across all students is generally seen as adequate, various group differences have been observed. The most consistent finding is that female students generally attain higher college GPAs than predicted, particularly when the SAT is considered in the admissions process. The evidence for ethnic groups is less clear, with studies showing both under-prediction and over-prediction.

Overall two clear findings emerge from the work on predictive validity. First, the ability to predict college performance is limited - at best all available predictors are able to account for about 30 per cent in the variation in GPA, but this is often considerably less - and second, prediction is highly variable between institutions. It is likely that the same factors underlie both of these findings. A further point worth considering is that the SAT appears to have most incremental validity in predicting performance under highly selective conditions, but less utility as a predictor of which students will actually complete a course. Whilst the more selective colleges may want to identify students with high academic potential, for many others it may be more important to select students likely to complete their course of study, even if their final attainment is only modest. Under these conditions, sufficient information may be obtained from high school record, so making the SAT redundant.

The limited ability to predict attainment is not exclusive to the area of education. For example, ability tests used for occupational selection at best account for between 25 to 30 per cent of variation in job performance (Schmidt and Hunter, 1998). A considerable problem in making any prediction for a selected group is range restriction. Range restriction refers to a limited range of scores in either the predictor (e.g. SAT scores) or criterion (e.g. college GPA) measure, and is clearly a problem when more selective colleges are considered. Although statistical methods for correcting range restriction exist, they provide only a partial solution and introduce a further source of error into the statistics.

A second statistical issue concerns the reliability of the admissions tests and attainment indicators such as GPA. Reliability indicates the extent to which the items on a test or assessment all measure the same underlying construct. For tests such as the SAT, the reliability of the math and that of the verbal sections reported, and show that each is measuring a distinct construct - verbal and numerical reasoning. The reliability of the total SAT score also suggests this is a coherent assessment of reasoning. The reliability of the grades which go to make up a student's overall GPA is more debatable. For example, a geography course may involve essay writing, independent project work, field work and also a reasonable fluency with statistics. Although all encompassed under a course labelled

'geography', these different aspects are clearly not measuring the same skills. Because of this, grading for courses such as geography may show low reliability, whereas this may be much higher in areas such as maths. This variation adds a further problem to prediction which is difficult to address without information on the reliability of course grading.

Whilst it has been shown that a general intelligence factor (often labelled 'IQ' or '*g*') underlies scores on tests such as the SAT and ACT, and academic attainment (e.g. Brodnick and Ree, 1995; Neisser *et al.*, 1996), this can be broken down into different components or skills which are relatively independent of each other. In terms of predicting academic success, this means that the predictors (e.g. SAT, ACT and high school record) may be measuring one set of skills and college GPA measuring another, quite diverse set. In this situation, consistent high levels of predictive validity are unlikely to be seen. This problem has recently been highlighted by Wolming (1999), who showed that the SweSAT and high school GPA could be reduced to verbal and numerical abilities. However, the definition of academic achievement differed between university courses, suggesting that predictor and criterion were not always assessing the same underlying construct. Under these circumstances high levels of prediction would not be expected.

Predictive validity has been seen to vary considerably between colleges and different groups of students. This may be partly due to statistical reasons, but researchers have been keen to speculate on other reasons for this. The ability of students to adjust to college life has been argued to be an important factor in academic success (e.g. Fleming and Garcia, 1998). Studies which have shown predictive validity to vary according to ethnic group, sex and type of college (e.g. predominantly Black or White students) have provided suggestive evidence that adjustment may be important (e.g. Hand and Prather, 1985).

Some attempts have been made to look at personality factors that may explain college performance over and above admissions tests and high school record. For example, Kanoy *et al.* (1989) found evidence that a more positive academic self-concept was related to college GPA in addition to SAT scores, as was a high level of internal motivation. Similar findings have also been presented by Fuertes *et al.* (1994), who studied Asian students, and Wolfe and Johnson (1995) identified self-control as an important factor for some students, as given the potential freedom college may offer, there will be a need to stay adequately focused on academic work.

Research has suggested that personality factors may be useful in predicting college attainment, but far more work is needed before there is sufficient evidence for them to be used in the admissions process. A final area to consider is the apparent decreasing ability to predict college attainment. Colleges are increasingly looking at individual students' needs, and providing greater support for them as they make their way through college. The provision of support such as remedial teaching effectively seeks to break the link between attainment on entry to college and final GPA, through assisting students in areas where they are weakest. As not all colleges are likely to offer the same level of support, a fact that may be reflected in their admissions policies, these differences may account for much of the variation in predictive validity observed between institutions.

4. Consequences of aptitude testing

4.1 Introduction

In any complex system such as that for education, the various parts that go to form the whole will not be independent from each other. In some cases this relationship may be quite obvious, for example the relationship between a curriculum and a test designed to assess attainment in that curriculum. In this example, changes to the curriculum would be expected to have a direct impact on the test designed to assess curriculum learning. If this did not happen, the validity of the test would be reduced, due to it not adequately reflecting the curriculum content. Although revisions to the test may be one of the more obvious effects of curriculum change, the impact of changes may spread beyond this in ways that are not always immediately apparent. If the revised curriculum covers a greater number of subjects or covers established areas in more depth, time may need to be taken from other school activities to provide necessary coverage. Equally, if the importance of attainment in certain subjects is raised, as has been the case since the publication of school league tables, greater emphasis may be placed on these at the expense of other subjects. The increased emphasis on certain areas may also result in 'curriculum alignment', where the curriculum narrows to focus on the test content, resulting in 'measurement-driven instruction' (Hamp-Lyons, 1997).

In the case of a high-stakes test such as the SAT or ACT, where the results can have a significant impact on a student's life chances, the influence it has on the education system are likely to be considerable. These effects are what Messick (1989; 1995) has termed 'consequential validity'. This concept recognises test use has an impact at a societal level, and it is through an examination of these effects, whether positive or negative, that the consequential validity of a test or testing programme is determined. The magnitude of the SAT testing programme, and the amount of controversy that has surrounded it, clearly show that the impact of the SAT has not been insignificant. Whilst this is not in doubt, the extent to which its effects have had positive or negative consequences for students, colleges and broader society is less immediately obvious.

This section of the review discusses the effects of tests such as the SAT and also considers what the effects might be if they were no longer used as part of the admissions process. Although much of the evidence in this area is necessarily speculative, it provides insights into possible changes that may occur in the British education system if aptitude testing was introduced for university entrance.

4.2 Test preparation and coaching

Whilst tests such as the SAT and ACT are not directly related to school curricula, they can play an important role in determining the future academic progress of high school students. Because of this, high school teachers undoubtedly feel that it is part of their task to prepare students for the tests. Results from tests such as the SAT are also used for monitoring purposes at a number of levels (e.g. school and state; Powell and Steelman, 1996), so placing an additional incentive on teachers to ensure that their students perform well on them.

It is hard to quantify the extent of preparation that students put into the SAT, and even more difficult to determine what detrimental effects, if any, school-based preparation has on other

subject areas and school activities. However, it is clear that a considerable amount of effort is devoted to test preparation, effort which in many cases extends to students purchasing additional study materials and often attending courses outside of school specifically targeted at raising test performance. As Linn has observed, 'As greater weight is placed on the results of the test, the links to and impact on instruction become stronger and stronger' (1983, p. 181).

The effects of the SAT on the curriculum in American high schools was explicitly acknowledged by Nancy Burton, when she was Director of Research and Development for the SAT programme. As Burton states, 'The College Board and ETS acknowledged that tests such as the SAT influence high school instruction despite repeated statements and research studies warning against teaching to the test' (1996, p. 5). One of the aims of the redevelopment of the SAT which took place in 1994 was 'better alignment with the curriculum standards being developed' to influence the curriculum in 'positive ways'. However, the relatively modest changes to the SAT, and its continuing reliance on predominantly multiple-choice answers, suggest that its 'positive influence' on the curriculum may be questionable.

In terms of preparation activities, Powers and Rock (1999) found that in a sample of over 2,700 students who had not attended formal coaching programmes outside their schools, most had undertaken some preparation for the SAT. This included 80 per cent who had taken the PSAT, 58 per cent who had read the guidance booklet *Taking the SAT* and 54 per cent who had previously taken the SAT.

Some critics of the SAT have targeted the time spent on preparation, arguing that abolishing the SAT would allow greater time for academic subjects (e.g. Wilmouth, 1991). Rooney with Schaeffer (1998) also argued that the effects of time spent on test preparation may adversely disadvantage low income and minority students, as the test scores of these groups are often not a fair indicator of their academic potential. Wilmouth (1991) further criticised the SAT for rewarding only partial knowledge, as guessing or elimination can be used to answer the majority of questions due to them being multiple-choice. He argued that this is exploited by coaching companies, as they tend not to teach reasoning skills but teach about how ETS designs the questions, essentially teaching 'test-wiseness' – the skill to reach the answer without necessarily knowing what it is. In support of this, Wilmouth (1991) quoted a study by Katz where students were given retired SAT verbal questions with and without the stimulus passages. When the passages were present, 70 per cent of the questions were answered correctly, but even without these there was a 46 per cent success rate, considerably higher than the 20 per cent which would have been expected. Although this 'guessability' of the SAT is seen as a major weakness by its critics, these findings are surprising and should not be over-interpreted without further evidence.

In 1978 the College Board introduced a booklet called *Taking the SAT* (College Entrance Examination Board, 1978), which explained about the test, gave candidates tips and included a full-length practice test. It is informative to note that this was introduced as a result of requests from secondary school teachers and students (Powers and Alderman, 1983), suggesting that teachers felt they needed more support and advice in preparing their students for the SAT, and that some students felt under-prepared. Powers and Alderman (1983) examined the effects which the introduction of this booklet had on students' SAT performance. In a randomised study of over 2,000 students, half were sent the new booklet with others receiving the existing preparation materials. Most students who had received the booklet had at least skimmed it or read parts of it, generally to supplement their test preparation rather than as a substitute for other methods. Regression analyses controlling for

class rank and PSAT score showed a slight effect of using the booklet on SAT math scores. There was some evidence that those who had the preparation booklet received a smaller penalty for guessing than those who did not, which translated into two or three scale points. The booklet also had a generally positive effect on confidence, and students who received it were more likely to feel that it had a beneficial effect on their performance compared to the older-style information.

Powers and Alderman concluded that despite favourable reactions 'a test familiarisation booklet like *Taking the SAT* is likely to have little, if any, effect on SAT performance' (1983, p. 77). It was also noted that many students had already received some practice, and that the older materials may have been adequate for the purpose of preparation. However, there was some evidence that the booklet had modified behaviours such as omitting questions, and its major benefit appeared to lie in the greater confidence it inspired in test takers.

One of the most publicly visible consequences of aptitude testing for college entrance in the United States has been the growth of the test coaching industry. Many companies offer students a wide variety of courses, all of which claim to enhance their performance on the SAT or ACT. Despite the variations in admissions procedures between colleges (see Section 2.2.3), these tests are still important to the majority of students, particularly if they hope to go to more prestigious and selective institutions. Coaching therefore has a considerable appeal, especially if the claims made by coaching companies are accepted without question.

For example, coaching companies have claimed to produce score improvements of up to 120 or 140 points, an increase of over one standard deviation (Powers and Rock, 1999). Although the profits of the coaching companies may depend on these claims being accepted by many students and their parents, Powers (1993) has criticised the ways in which these claims are advertised. The major limitation is that they fail to include control groups. To determine the true effects of any intervention such as test coaching, it is necessary to know what would happen to a comparable group who retook the test, but who did not receive the intervention. Without this control it could be equally argued that any increases in scores simply resulted from practice and familiarity effects due to taking the test twice. Indeed, it is known that simply retaking the SAT will improve scores on average (e.g. Nathan and Camara, 1998).

Improvements in scores can also arise through learning and development over time and measurement error. For example, a study of young, able students has suggested that age may account for improvements of about 50 points per year (Wilder *et al.*, 1988, cited in Powers, 1993). Measurement error from one test to the next also means that typically 1 in 25 test takers gain 100 points or more and 1 in 110 will loose 100 points or more, depending on initial score (Powers, 1993).

ETS have been keen to determine the extent to which the SAT is coachable, particularly following claims that the revised SAT I: Reasoning Test is more susceptible to coaching that its predecessor (Powers and Rock, 1999). In 1993, Powers provided a summary of the research on the effects of coaching. Overall, effects were somewhat greater for the more curriculum-related area of math than for the verbal section. There was also some evidence that longer coaching courses produced greater score gains. In this review the SAT was seen to compare favourably with many other aptitude tests, generally being less susceptible to the effects of coaching. This was argued to be largely due to the relatively simple format of SAT questions, as more complex question formats were argued to be more coachable. However, it was noted that the results were difficult to interpret, as students who take coaching programs

are also likely to use other preparation methods (e.g. books), so making it difficult to attribute effects specifically to coaching courses.

A recent stratified survey of 4,200 SAT takers revealed that 12 per cent had attended coaching programmes not offered by their schools (Powers and Rock, 1999). The effects of coaching were seen to be between four and 14 scale points for the verbal test and 12 to 22 for the math test, with a 95 per cent confidence band. According to established effect sizes, these can be considered small, being around 0.1 and 0.2 standard deviations respectively. When examined by background variables, there was some evidence that students who had good high school grades benefited more, and improvements on the verbal test were associated with number of years of English taken, English grades, English being reported as a student's best language, and parental education.

Demographic differences were also seen between those who sought coaching, and those who did not. Specifically, students who attended additional test preparation classes were more likely to have more affluent and highly educated parents, higher high school grades and higher degree aspirations, and to choose colleges which required higher SAT scores. Those who attended coaching sessions were also more likely to have prepared for the SAT in a variety of other ways (e.g. books, study aids), although 'uncoached' test takers undertook preparation of various sorts, and so were not totally unprepared.

This study concluded that the effects of coaching were small and much less than those claimed by coaching providers, although the effects were larger for the math than the verbal section of the SAT. The results were seen to be highly comparable to those from previous studies on coaching, and provide no evidence that the revised SAT was more coachable than the previous version.

Coaching may only have modest effects on SAT scores, but an important question is whether these changes have a significant impact when scores are used in the prediction of college grades, and so affect which students are offered places. Evidence relevant to this has been offered by Baydar (1990), who conducted a simulation study on the effects of coaching in four colleges which varied in selectivity and student characteristics.

Using previous data on coaching, the probability of each student receiving coaching was calculated, and the estimated effects of coaching were added to the selected students' SAT scores. The effects of this on predictive validity were then studied. The effects of coaching were least in the less selective colleges, due to the generally larger variations in SAT scores of students in these institutions. Students who attended these institutions were also less likely to receive extra coaching. When students in the least selective colleges did receive coaching, it was argued that their score increases might reflect background variables known to be associated with first-year grades (e.g. parental income and education). Predictive validity in more selective colleges was affected to a greater extent by coaching, as the spread of SAT scores was generally less in these.

The sometimes intensive preparation that students undertake before sitting entrance tests for higher education is not limited to the SAT. Allalouf and Ben-Shakhar (1998) reported that in Israel the number of students taking coaching courses for the PET rose from one per cent in 1984 to 77 per cent in 1996. In a study which employed two randomly assigned groups of students, Allalouf and Ben-Shakhar studied the effects of coaching on score gains and predictive validity. Coaching was seen to have a significant effect on test scores, with larger effects being seen for the numerical than verbal section, in accordance with findings for the

SAT. In all cases, scores from when students took the PET a second time were more highly correlated with high school GPA, whether they had received coaching or not.

Coaching for the PET was therefore concluded not to have a significant impact on its predictive validity. When regression lines were studied, a modest effect was observed for coaching, with predictions for those who were coached being slightly more accurate than for those who were not. Allalouf and Ben-Shakhar argue that coaching may have a positive effect, allowing students with poor test-taking skills to show their true abilities. Although apparently contradictory to Baydar's work, in interpreting this study it should be noted that the effect of college selectivity was not studied, and the criterion used by Allalouf and Ben-Shakhar was high school matriculation grade, not actual college performance.

Although preparation for tests such as the SAT has beneficial effects on test performance, simply increasing the amount of time spent on preparation or coaching will not result in correspondingly higher scores. As Powers (1993) has observed, although longer coaching courses produce greater score gains, 'simply doubling the effort, for example, does not double the effect. Diminishing returns set in rather quickly, and the time needed to achieve average score increases that are much larger than the relatively small increases observed in typical programs rapidly approaches that of full-time schooling' (p. 26).

A previous review on the effects of coaching by Messick and Jungeblut (1981) attempted to relate actual score increases to coaching time. Through combining the findings from a number of studies, they estimated that an increase of ten scale points on the verbal section of the SAT could be expected from 12 hours coaching, but that 20 points would require 57 hours, and 30 points 260 hours. The number of coaching hours for comparable increases on the math section was slightly lower, probably due to this section being more coachable than the verbal (e.g. Powers and Rock, 1999), but still considerable. On the basis of these findings, Messick and Jungeblut argued that SAT preparation should be integrated into students' broader education: 'the soundest long-range mode of preparation for the SAT would appear to be a secondary school program that integrates the development of thought with the development of knowledge' (1981, p. 216).

4.3 Aptitude testing and educational opportunities

Evidence has been presented that the SAT may not be a fair reflection of the academic potential of certain groups of test takers, both in terms of the overall scores that are derived from it and in its prediction of college attainment (see Section 3). The potential consequences of this for students are discussed here.

Inhibiting access to college is the most obvious way in which tests like the SAT can restrict the educational opportunities of low-scoring groups. This is likely to be most noticeable in the more prestigious, selective colleges, where SAT scores may be given more weight and used to establish minimum requirements for students. Crouse and Trusheim (1988) have provided a detailed analysis of the effects of using the SAT on the admissions of Black and White students. Using data from a national longitudinal study of high school students, they found that colleges which used class rank and SAT scores together rejected approximately 11 per cent more Black applicants than Whites, compared with using class rank alone. This occurred because using the SAT as part of the prediction of first-year college grades reduced predicted grades for Black students but not for Whites. Even when colleges lowered their admissions criteria for Black students, the addition of the SAT to high school class rank had very little effect on grade prediction over high school class rank alone.

In order to attain a more diverse student population and to ensure that they do not discriminate against certain social and ethnic groups, some institutions have adopted affirmative action policies. These have tended to target specific groups, most often the ethnic minorities, with the goal of increasing their representation. Whilst these programmes seem to have had some positive effects, recent studies have suggested these have not been sustained (e.g. Carnoy, 1995; Tekian, 2000). In terms of GPA and selection tests used for medical schools, Tekian observed that quantifiable factors derived from such tests have proved poor predictors of performance, although they still predominate in the admissions process. In order to recruit more students from under-represented minorities, it was suggested that medical schools need to broaden their conceptualisations of intelligence, and look at a wider range of qualitative factors in admissions generally.

Carnoy (1995) took a broader perspective on the falling proportion of Blacks attending college. He argued that the American government changed its policy from one of reducing discrimination, to a view that this goal had been achieved and discrimination no longer existed. This was accompanied by a reduction in funding for financial aid, coupled with increased poverty among minority groups.

Evidence for the shift in opinions about discrimination at a national level described by Carnoy can be seen in changing policies and use of test scores in the United States. Tekian (2000) viewed affirmative action policies as being 'increasingly under attack' and Rooney with Schaeffer (1998) reported that some states have banned the use of racial preferences in selection decisions. In discussing the effects of this, Rooney reported that some universities have had to reconsider their use of SAT and ACT scores in conjunction with affirmative action policies. In some cases, colleges have dropped the use of admissions tests or made reporting optional. Where this has happened, no fall in the quality of applicants or college performance has been observed, but colleges have benefited from the student pool becoming more diverse (e.g. Schaffner, 1985).

Evidence suggests that SAT scores represent a direct barrier to college access for some students, but their effects may also be felt in other ways. For example, SAT scores can play a significant role in the awarding of state scholarships. Rosser (1989) argued that the generally lower scores of females may explain why they often do less well in state scholarship competitions. Survey results indicated that in states which use class rank and SATs, or where SATs are not used at all in scholarship competitions, females tend to do better. The average scores of many ethnic groups have also shown they would be at even more of a disadvantage if only SAT scores were relied on. The use of tests for awarding scholarships shows that admissions tests can affect educational opportunities through indirect means. In the case of ethnic minorities, they may work against groups who need greatest support – even if students are able to obtain places, they may be denied the financial support they need to be able to take them up. In some states it is now accepted that aptitude tests are biased against certain groups, with this being reflected in law. For example, Rosser (1989) reported a case in New York when a judge ruled that using the SAT as the sole basis for awarding scholarships was unlawful.

Children from more affluent backgrounds may also be advantaged in other ways when it comes to college admissions tests. Students have to pay to take the SAT and ACT, and although the cost of this is not great, it may be a burden on the poorest families. (At the time of writing, the SAT I: Reasoning test costs $24 to take and the ACT $26, although financial support for this is available in some cases.) However, there are many more potential costs associated with admissions tests in the form of preparation materials, coaching courses and

retakes. Powers and Rock (1999) showed that a large proportion of students used a range of preparation materials. Although some of these may have been supplied by the school, the amount of preparation materials available from publishers other than the College Board suggests that there is a considerable market for these. Actual coaching courses can also incur additional expenses, and the substantial rise in coaching in Israel reported by Allalouf and Ben-Shakhar (1998) has shown this to be very popular in countries which have comparatively recently introduced aptitude testing.

Additionally, the increases in scores that are often seen when retaking the SAT may encourage students to do this: 'As a counselor at a college preparatory school, I saw 80 percent of seniors in 1993 increase on either verbal or math SAT their second time, and 56 percent had an increase on their second Composite ACT' (Smyth, 1995, p. 30). Smyth went on to advise high school counsellors that 'Students interested in selective colleges are well-advised to take both the ACT and the SAT twice …This may sound like a lot of testing, but I think the benefits justify four Saturday mornings over eight months' (p. 30).

The preparation, testing and likely retesting may all add up to a significant financial outlay for some of the poorest students. As with the scholarships discussed above, this may further disadvantage those students who have already been put at a disadvantage by the nature of the SAT or ACT. Score increases for retaking and coaching are modest, but nevertheless exist, indicating that coaching and resitting the tests will often pay off to a limited degree. This may be particularly the case in admissions to selective colleges, where coaching has been shown to have a positive effect on predicted grades (Baydar, 1990).

4.4 What would happen if the SAT were abolished?

Both critics and supporters of admissions tests have been keen to speculate on what would happen if these tests were no longer used. This would clearly have an impact on the considerable business in test preparation and coaching that has built up around the SAT and ACT, and on the work of ETS, the College Board and the ACT Program, but what would the broader effects on the high school and college system be?

Due to admissions testing being well established in the United States, and since its establishment there having been a considerable growth in the student population, it is difficult to determine what would happen if this system was no longer in place. If admissions tests were dropped, colleges would have to make their selection decisions primarily on the basis of high school class rank or GPA.

Crouse and Trusheim (1988) reported a number of arguments put forward by senior staff at ETS against the reliance solely on high school record. First, similar GPAs from very different schools may not mean that same thing - whilst representing modest achievement from a generally high-attaining school, they may indicate a student of greater ability and motivation from a low-attaining school in a deprived area. Second, it is has been argued that without results from a standardised test such as the SAT, decisions may be made on the basis of the prestige of the high school attended and judgements as to whether a student is likely to 'fit in' to the college. Third, the increased emphasis on high school grades may mean that students opt for easier courses to boost their GPA, and their tutors may adopt more lenient grading procedures, leading to rampant grade inflation. If this happened, high school GPA would soon lose all its predictive validity. In a related argument, Wilmouth (1991) argued that the SAT effectively provides a check on high school grading, so preventing this becoming more lenient.

In discussing these claims, Crouse and Trusheim have argued that many selective colleges have relatively small student intakes. Admissions staff would therefore be able to read each application and in the majority of cases should have knowledge of an applicant's high school. Even when they did not have this information, factual information on the high school that would allow admissions tutors to make an informed decision should be readily available. They go on to report an incident where SAT scores were late arriving at a college for some students, and admissions tutors had to make decisions in the absence of reports from ETS. When the ETS reports did arrive, it was determined that no changes in the admissions decisions would have been made had they been available at the time. In terms of the third claim, that students would choose easier courses and high school grades would become inflated, Crouse and Trusheim argued that there is no evidence that high school grades had less predictive power in the 1960s when testing was less prevalent. Indeed, there is some evidence that the opposite is the case (e.g. Fincher, 1990), although it is not possible to prove conclusively that this scenario would not occur.

It is difficult to provide an unbiased judgement of what may happen if admissions testing was dropped - the majority of the evidence surrounding this is opinion and speculation. However, some more concrete evidence is available from colleges where SAT scores have been made optional or abandoned altogether. Indeed, this pool of evidence is likely to be quite large, as Rooney with Schaeffer (1998) reported that more than 275 four-year American colleges did not use SAT or ACT scores to make selection decisions about some or all of their students. Evidence on the effects of this have already been presented (see Section 2.2.3), but overall it appears that where reporting of scores has been made optional, this has not led to noticeable decline in academic standards (e.g. Schaffner, 1985).

Whilst evidence from colleges which have made the SAT optional is suggestive, it does not address all the issues raised above. The possibility of grade inflation in high schools and changes in subject choices, for example, cannot be determined. Whilst the SAT is still in place and plays a role in college admission for the majority of students in any one high school, grading will not be affected by the policies of a limited number of colleges.

4.5 Discussion

Evidence on the effects of admissions testing has focused primarily on work from the United States, as it is there that testing is most firmly established and has been the subject of greatest discussion. This system clearly has significant consequences for both those directly involved with the testing process and those such as policy makers and legislators who have to consider the role of testing within broader society. It is easy to be critical of the American system, particularly in the areas of equality in college access and the commercial test preparation programmes. However, the debate which has surrounded the SAT has been open and largely in the public domain, whereas until recently there has been very little public debate on the A-level system.

In commenting on Messick's (1989) work on consequential validity, Reckase (1998) argued that there is a logical error in this for test developers and legislators considering the introduction of new assessment systems. This occurs as the consequences of a testing system such as the SAT or ACT cannot be known at the time of test development. Despite this, Reckase acknowledged that some anticipation of consequences may be possible, and experiences from similar initiatives called on to inform judgements. In terms of the SAT, consequential validity has only really been addressed since Messick's influential work in the 1980s, a long time after the introduction of the SAT at the start of the last century. However,

it is worth considering that the goal of the College Board at this time was to standardise and streamline the admissions process to American colleges, and to a large extent this goal was achieved.

Aptitude tests have been introduced more recently in Sweden and Israel. In both cases the increasing number of students applying to study at universities has meant that selection into higher education has been necessary. Although the tests used in these countries may not provide a perfect solution to this problem, they do provide a defensible answer to the problem of selection. In Sweden, the SweSAT was originally introduced to open access to higher education to older students, and encourage greater participation from this group. To the extent that aptitude testing offers another possible route into higher education, particularly if this was not previously open to certain groups, at least the intentions of introducing such a system are broadly positive, even if its fuller consequences are more difficult to foresee.

Literature on the SAT and other admissions tests raises some interesting possibilities on the potential consequences of introducing aptitude testing for university entrance in Britain. Experience indicates that certain consequences are almost inevitable. Preparation for any such high-stakes test will detract from other courses of study, although the extent to which this may reduce subject knowledge, lower A-level attainment and mean that students are less prepared for university study is unclear. Pressures on students to perform well on any selection test are also bound to create a market for test preparation courses, but the costs associated with these may mean that they are only accessible to more affluent families. Finally, any such test will attract considerable attention, and so will need to be defensible both in the eyes of assessment experts and the wider public.

Ultimately, if such a selection instrument was more successful in identifying talented students than the current A-level system, this could be used by universities to increase the selectivity of their intake, and so their prestige. Whether the creation of even more elite universities would have desirable effects on the wider British education system and society in general needs to be carefully considered. These issues are further discussed in the final section of this review.

5. Research in Britain

5.1 Introduction

This section of the report summarises research that has been conducted in Britain on the prediction of success in higher education. A detailed account of the work conducted by Bruce Choppin and colleagues at the NFER is given, as this represents probably the most in-depth investigation into predicting performance in British universities available. Following this, an overview of the more limited studies which have been conducted in this area is presented.

The recent debate on bias in access to British universities has highlighted that a number of institutions are considering the issue of access. Most relevant to this review are those institutions looking at ways of identifying students who have the potential for study at higher education, but whose personal circumstances may prevent them from fully demonstrating this through the A-level system. To gain an understanding of this work, interviews were conducted with staff from three institutions and the findings from these are also reported in this section.

5.2 Predicting success in British universities

5.2.1 Previous research conducted by the NFER

Background: Probably the most detailed investigation into predicting academic success in British universities was conducted during the late 1960s and early 1970s by Bruce Choppin and colleagues at the NFER (Choppin *et al.*, 1973; Choppin and Orr, 1976). This programme of research stemmed from the Robbins Committee on Higher Education, which was set up in 1961 to look at what developments in the higher education sector were needed. At the time of its inception, many universities had their own entrance exams and the system of entrance to higher education was described as being 'chaotic'. However, whilst the Robbins Committee were working, the University Central Council on Admissions (UCCA) was established, which did much to standardise the admissions procedure.

Despite the increase in the number of places at universities and polytechnics, it was still clear that some form of selection into higher education would be necessary, particularly as there was a noticeable swing away from natural sciences and towards humanities. The A-level system was seen as being a poor predictor of university success, and a need was identified for an assessment which could supplement A-levels and predict performance over the duration of a higher education course. The Robbins Committee report stated that the SAT should be further investigated as a tool for selection, but that any such test should not be viewed as a replacement for academic examinations (Robbins Report, 1963). The recommendations of the Robbins Committee led to the research conducted by Bruce Choppin and colleagues at the NFER. This work involved the development of an aptitude test for higher education and an evaluation of this. In total this research lasted six years, due to the necessity of following students throughout their time at university to provide an adequate evaluation of the aptitude test.

The Test of Academic Aptitude: The test which was developed for this work was called the Test of Academic Aptitude (TAA). The structure of the TAA was based on the SAT, with it having a verbal and a numerical section. The verbal section consisted of five types of multiple-choice questions: sentence completion, antonyms, analogies, reading comprehension

and verbal discrimination. The first four of these were present in the SAT at the time the TAA was developed. The fifth type, verbal discrimination or 'odd man out', was introduced by the test working party. It was reported that the verbal questions covered a wide range of subject areas. The numerical section of the TAA also mirrored the SAT by having two item types: '(a) general mathematics problems similar in type to those employed in multiple-choice tests of achievement, but with an emphasis on powers of reasoning rather than factual knowledge, and (b) data sufficiency items in which the candidate was asked to judge the logical completeness of a set of information' (Choppin and Orr, 1976, p. 26).

A number of versions of the TAA were developed and used for the research programme. All had 90 verbal items and 60 numerical items, with the exception of one version which had only 54 numerical items. Factor analyses clearly identified the verbal and numerical factors from the subtests, and both sections had acceptable reliability (0.82 or greater for the numerical section and 0.90 or greater for the verbal).

Research with the TAA: The first trial of the TAA was conducted in October 1967 and the findings from this reported by Choppin *et al.* (1973). For this the TAA was taken by over 27,000 sixth-form students, of whom just over 7,000 entered universities in Autumn 1968.

As predictors of first-year degree results and final degree grade, Choppin *et al.* considered the following variables: scores from the TAA, number of O-level and A-level passes, mean A-level pass grade and school assessment of suitability for higher education. Results were analysed by subject where possible, but the limited number of students studying some subjects prevented this being done for all areas. This was noted as a limitation due to the interesting predictive differences observed between subjects, although sufficient numbers were attained for the analyses in all major subject areas.

In terms of simple correlations, mean A-level grade was observed to be the best predictor of first-year degree results in most courses, followed by school assessment of university aptitude. Mean A-level correlations varied between 0.49 for mechanical engineering to 0.17 for history. Correlations with TAA maths score varied between -0.07 for economics to 0.30 for psychology, and verbal score correlations varied from -0.13 for economics to 0.22 for history. This showed that at best, A-levels accounted for 24 per cent of first-year degree results, and the TAA maths and verbal sections nine and five per cent respectively.

Further analysis of the data involved the use of multiple correlations, which allow a number of predictors to be considered simultaneously and indicate the unique predictive power of each. When considered together, TAA maths and verbal scores were associated between 0.15 and 0.20 with first-year degree results for most courses. It was observed that these multiple correlations were usually only slightly higher than the larger of the correlations when TAA maths and verbal scores were examined separately. Maths and verbal scores added very little to the prediction of first-year degree results obtained from school assessment and number of O-levels achieved. Similarly, the TAA added very little to the prediction after A-levels had also been entered with school assessment and O-levels.

Sex differences in prediction were noted, particularly for science courses. Both TAA maths score and mean A-level grade predicted first-year science grades better for females than for males, and it was noted that this replicated previous findings from America (e.g., Seashore, 1962). O-levels and school assessment were seen to predict science grades comparably for males and females, but when TAA maths was considered in addition to these, the superior prediction for females again emerged. Choppin *et al.* (1973) argued this showed that the TAA maths score predicted well for females, and that it was able to do this independently of the other predictors studied. When A-level results were added as predictors, the strength of the

correlations increased considerably, but TAA maths was still seen to raise the correlation between predictors and first-year grades in science more for females than males.

When looking at the prediction of final degree result, patterns varied considerably according to subject area. Overall, mean A-level grade was the best predictor of degree performance followed by school assessment with average correlations of 0.36 and 0.26 respectively. Prediction was highest in the areas of science and technology, but was quite variable in arts and social sciences. The two TAA scores again showed wide variation in their relationship with degree results, with correlations rarely exceeding 0.2 and some being negative.

Multiple correlations for the TAA and final degree results were similar to first-year results, with combined TAA maths and verbal scores generally not exceeding the larger of the two when considered separately. When looking at O-levels and school assessment, the information that most admissions tutors have when making their decisions, adding TAA scores generally increased prediction by only 0.02 to 0.03 and rarely above 0.05. When considering all normally available information, including A-level grades, prediction of science and technology courses was around 0.5, language and arts around 0.4 and social science courses between 0.2 and 0.5. TAA scores added little to any of these predictions, with increments ranging from 0.01 to 0.03.

The effects of applying minimum A-level and TAA scores to help identify those students who were likely to fail the first year of their degree course was also studied. It was found that strictly applying an A-level cut-off could reduce the number of failures, whereas using the TAA did not.

Choppin and Orr (1976) present a summary of the work from the three administrations of the TAA, including the work presented above. One of the additional samples was obtained in 1968 and comprised sixth-form students regarded as 'likely university applicants'. This was later reduced to include only those who actually applied for university, giving a sample of 10,561 students. The TAA was further administered to a one-in-seven sample of schools with students who applied to UCCA in the previous year. Likely university applicants, regardless of being in the upper or lower sixth were tested in October 1969, with there being 11,615 students in total.

The results from the three administrations of the TAA were combined for the majority of the analyses reported by Choppin and Orr (1976). This report with the combined data largely replicated the findings from the first administration of the TAA described above. Overall, performance on the TAA was not seen to be independent of A-level subjects. Science students scored higher on the TAA maths section than non-science students, and the high scores of A-level maths candidates particularly stood out on this part of the TAA. However, the overlap between TAA and A-level results was modest, suggesting that TAA scores were not totally redundant.

Across the three samples, mean TAA scores also differed between those who went on to university and those who did not, but the individual overlap within these groups was too great to identify individuals likely to attend university simply on the basis of the TAA. Those who were successful in their university applications also had better school assessments and GCE results.

Combining the findings from the three studies, TAA scores were seen to be far weaker predictors of first-year degree performance than school assessment or A-levels. Associations

varied from 0.19 (maths and history) to 0.07 (mechanical engineering) for the TAA maths section, and from 0.23 (sociology) to -0.04 (civil engineering) for the verbal section. When regression analyses looked at the incremental validity of the TAA over school assessment and number of O-levels attained, the typical value for this was 0.04. Prediction of overall degree result was also considered, and individual results were found to be quite similar to those obtained for first-year performance. The TAA typically added about 0.03 to the prediction of degree performance over teacher assessment and O-levels, the typical information available when selection decisions are made. That is, using the TAA enhances the prediction of university attainment by less than one per cent over the information normally considered when admitting students.

By pooling students from all three administrations of the TAA, sufficient numbers were attained to compare predictive validity for maths and medicine courses between universities. Considerable variations between universities were seen when predictors were examined separately, with the exception of mean A-level grade, which consistently appeared to be a good predictor. When all predictors were combined, a reasonable level of consistency in prediction between universities was observed, probably due to the substantial influence of A-levels.

Overall it was noted that the level of prediction obtained for the TAA was much lower than that often quoted for the SAT combined with measures such as high school grades. In summarising the findings, the authors considered issues behind the lack of predictive power. As the TAA was administered under research conditions, one important factor may have been the motivation of the test takers. However, comments from test takers and reliability analyses suggested that random responding was not occurring, and students saw it as a valuable exercise and so were motivated to perform well on the test. The lack of predictive power was also not seen to be due to technical limitations of the TAA, as it was adequately developed, and there were clear parallels between the TAA and the SAT.

The different structure of the school system in Britain was seen as one reason why the TAA may not have shown the predictive power expected. The sixth-form students had already been selected at a number of stages by the time they took the TAA (11+, O-levels and satisfactory A-level progress so far). They would therefore have been a more highly-selected sample than students who take the SAT as there is much less screening before college in the United States. In terms of predicting drop-out, it is also noted that failure for academic reasons at university is comparatively rare, and that many apparent ‘academic’ failures may be due to personal or motivational problems.

Since this work was conducted, it is worth considering how the education system in Britain has changed. In some ways it may have moved closer to the less selective system in America, as the increasing proportion of students attaining high grades which make them eligible for university places suggests that the British system has become less selective. Additionally, although selection still occurs through exams such as GCSEs and A-levels, there are now more diverse routes through which students can access higher education (e.g. vocational courses, access courses). It is interesting to speculate whether a replication of Choppin’s work today would provide greater support for using a test such as the TAA in university selection.

5.2.2 Research on A-levels and other tests

Research on the prediction of success in higher education has been conducted both before the work with the TAA, and since. This work has primarily focused on attainment exams (e.g. O-levels, A-level, Scottish Highers) as predictors of success at university, although intelligence tests have also been studied.

A-levels and Scottish Highers are the major source of information that admissions tutors use when making decisions (Smithers and Robinson, 1991). In order to determine whether admissions decisions are valid, researchers have sought to identify the link between attainment at the end of sixth form, typically the time when A-levels are taken, and subsequent degree performance.

An early study in this area was conducted by Williams (1950), who investigated the association between subject marks on the Northern Universities' Joint Matriculation Board exam and first-year university performance. Overall, prediction in science subjects was seen to be better than in arts. High positive correlations were seen between some subjects at matriculation and first-year degree attainment (e.g. biology (0.77), Latin (0.79)), but many commonly studied subjects were predicted poorly from matriculation results (e.g. English (0.33), physics (0.33)). Degree subjects such as economics were considered to be very poorly predicted, possibly due to this not being an area covered by the matriculation exam at that time. It is worth noting that although Williams considered performance in most subjects to be poorly predicted, many of the correlations he reports are comparable to, or higher than those seen in subsequent work. In studying the syllabuses for different examinations, it was observed that where associations between matriculation exams and university grades were high, there was considerable overlap in course requirements. The correlations reported for subjects such as biology and Latin suggests this may have been particularly so, and also raises the possibility that the differences between the education processes of sixth forms and universities were less at the time of this work than they are now.

Around the same time, the ability of Scottish Highers to predict university performance was also being examined. For example, Nisbet and Welsh (1966) studied the performance of 303 arts and 198 science students at Aberdeen University between 1961 and 1964. Correlational analyses were not conducted on the data, but when degree performance was broken down by number of Scottish Highers achieved, clear associations between the two were seen. Similar results had been previously obtained by Gould and M'Comisky (1958), who had looked at 674 arts students who entered Edinburgh University between 1949 and 1951. Again a clear link was seen between Scottish Highers and degree performance, and this was particularly noticeable at the top and bottom ends of the high school qualifications. It should be noted that this study is interesting as providing students met minimum standards, all students who applied were admitted to the Arts Faculty. This means that the sample was far less selected than students in many comparable studies.

Both of these studies also examined wastage, that is, students who failed to successfully complete their degree courses. Nisbet and Welsh (1966) found first-year degree progress to be crucial, in accordance with previous work. Various analyses were conducted to try and identify the best indicators of subsequent failure, and although poor performance in two or more subjects identified the majority of science students who would fail, no such criteria could be established for arts students. Nisbet and Welsh report that this research led to the implementation of a system at Aberdeen University which monitored first-year exam results and offered support to students where necessary. Evidence suggested that this had a positive

effect on dropout rates. In their work, Gould and M'Comisky (1958) found the highest wastage in the least qualified group. However, due to considerable variation in the performance of students with similar qualifications, they concluded that Scottish Highers could not be used to indicate potential wastage and an alternative indicator of this was needed.

Kelsall (1963) provided an early review of research into university selection, which concluded that the predictive power of A-levels had been seen to vary considerably between studies - some identifying weak prediction whilst others concluding this was much stronger. Taking a broader view, prediction was seen to be better in the area of science than arts. This review also found that the links between specific A-levels and

corresponding degree subjects were often not as strong as may be assumed. Overall, what did emerge from the literature was that A-levels, although far from perfect, were one of the best predictors available.

More recent summaries of the research have provided comparable conclusions. Peers and Johnston (1994) conducted a meta-analysis of 20 prior studies which included 60 analyses of the relationship between A-levels and degree performance. Overall, the association was seen to be 0.28, which, after taking possible sources of error into account, was significantly different from zero. This indicates that A-level grades accounted for just under eight per cent of variation in degree performance on average. As meta-analysis provides one of the most powerful methods available for synthesising research results, this study provided convincing evidence that A-levels were modest predictors of degree performance.

Peers and Johnston (1994) were able to further refine their findings by examining the effects for type of institution (university versus polytechnic) and academic discipline. Across comparable disciplines, the predictive power of A-levels was seen to be greater in universities than polytechnics. Differences in the extent to which A-levels predicted science performance were particularly noticeable, with prediction being 0.43 of a standard deviation weaker in polytechnics. Where data was available for subjects studied in both universities and polytechnics, some consistency in the rank order of prediction for subjects was seen. In both universities and polytechnics, A-levels were weakest predictors for social sciences courses, but stronger predictors of arts and languages. In accordance with previous findings, the best prediction was seen for science courses in universities, but A-levels predicted performance on these courses very poorly in polytechnics. Whether these differences would still emerge now that polytechnics have changed their status to that of universities is unclear.

As a result of their work, Peers and Johnston questioned the reliance on A-levels as entrance criteria to universities, as degree success appeared to be influenced by other factors. They suggested that A-levels functioned least well as predictors in 'contexts where a mature learning approach based on personal understanding is encouraged' (1994, p. 13), although no evidence on differences in learning environments was presented to support this view. The differences between effect sizes for disciplines were consistent with the idea that interactions between learning approach and environment mediates the predictive power of A-levels. For example, they argued that social science courses may provide students with different information and methods of obtaining this than they have been previously exposed to, whereas other subjects may be more similar to A-levels in requiring an accumulation of facts. This would account for the small effect sizes in social

sciences and larger ones in sciences, technology and medicine. Alternatively, differences in the reliability with which attainment in these subjects is assessed could account for these findings.

The differential prediction of subject areas observed by Peers and Johnston has been replicated by a number of other authors. For example, Chapman (1996) explored the link between A-levels and degree results in eight subjects over 21 years. The strongest links were seen for biology (0.47) and weakest for politics (0.23). When examining variations between the number of 'good degrees' (firsts and upper seconds) awarded by universities, A-levels accounted for only 5.3 per cent of the variation in politics but 23.5 per cent for maths. Previously, Pilkington and Harrison (1967) had shown mean A-level grade to correlate 0.24 with first-year degree performance in psychology, and 0.30 with final degree grade. A-levels therefore accounted for approximately nine per cent of the variance in degree performance. A similar figure for mean A-level performance was also reported by Richardson *et al.* (1998) for prediction of pre-clinical exam results in medical students.

The extent to which the SAT predicts college success for different social groups has been the focus of much controversy in America. A-levels have received relatively little attention in this area, although available evidence suggests this needs further investigation. In a study of students in two science and two non-science disciplines at Manchester University, Peers (1994) found A-levels under-predicted degree performance more for males than females in social sciences, and that greater under-prediction occurred for older students in engineering and technology. From this it was concluded that A-levels should not be treated as valid predictors if subgroup membership is ignored.

Differences between the success of males and females at Oxford and Cambridge have also been identified (McCrum, 1996). In most subjects it was found that to have equal final degree grades, females needed considerably better A-level results on entry than males. Subjects in which this was particularly noticeable included maths and physics. It was observed that although A-level results for females were slightly lower than for males, this difference was not sufficient to account for their considerably worse degree performance.

Possible causes of the gender gap in performance at Oxford University have been investigated by Mellanby and Rawlins (1997). They found that in a psychology, philosophy and physiology course, males did better at philosophy but not psychology. This suggested that a difference in general ability was unlikely to account for the observed sex differences. Subsequent work by Mellanby *et al.* (2000) involved assessing students on high-level tests of verbal reasoning and a range of individual difference measures.

These measures were examined for sex differences and the association between them and degree performance was studied. Scores on the verbal reasoning test predicted final degree class (0.33), but as no sex differences in scores were observed it was concluded this could not account for the differences in degree performance. Some sex differences were noted in the variables studied, but this only occurred in variables unrelated to degree performance. The causes behind sex differences in degree performance at Oxford therefore remain unknown.

Mellanby *et al.* (2000) showed that verbal reasoning test scores were a moderate predictor of degree class. Although these students took the verbal reasoning test towards the end of their degree course, so possibly inflating the concordance between the two, previous work has obtained test scores from students earlier in their degree studies. For example, Pilkington and Harrison (1967) gave 252 students on a first-year psychology course at the University of Sheffield two high-level reasoning tests. The tests correlated 0.28 and 0.18 with final degree

result, but regression analyses showed that these tests did not contribute to the prediction of final degree class after first-year marks and A-levels had been allowed for.

5.2.3 Summary and issues for consideration

Overall, available evidence indicates that A-levels, at best, are quite limited in their ability to predict success in higher education. Peers and Johnston's (1994) figure of 0.28 indicates that A-levels accounted for slightly less than eight per cent of the variance in degree performance. Choppin's work suggested that A-levels may have had slightly more predictive power, with average correlations for A-levels and first-year university performance being 0.36 (Choppin *et al.*, 1973), but this still showed that no more than 13 per cent of variance is explained. In light of this, the conclusion presented by Nisbet and Welsh over 30 years ago still seems to apply: 'It would appear that a substantial part of the variation in students' performance in university is basically unpredictable from evidence available at the time of entry to university' (1966, p. 477).

However, it needs to be remembered that Peers and Johnston's and Choppin *et al.*'s findings do not strictly reflect the information typically available at the time admissions tutors have to make their decisions, as they usually have to work with predicted A-levels, not actual results. Considering the unreliability in predicted A-level grades (Delap, 1994; 1995), the ability of these to predict which students are likely to be successful at university is probably even lower.

Although prediction of degree attainment across all subjects is modest, this masks a wide variation at subject level. Evidence has been presented showing prediction of science subjects to be generally more accurate than of arts, which in turn is more accurate than of social sciences. Peers and Johnston (1994) have suggested that this may be due to the greater similarity between A-level and degree courses in some disciplines. Also, many science-related courses rely on A-levels to provide students with the fundamental knowledge that the degree course will build on. If they have not attained this basic knowledge, as reflected in their A-level results, they may be less likely to succeed when required to further it. Similar arguments could be applied to subjects such as languages, but it would be unwise to draw global conclusions without looking at the requirements of individual courses.

In many social science degrees, which were the least well predicted, the subject matter may be less familiar to students, as they are often not required to have studied the subject at A-level. As Peers and Johnston (1994) have observed, the type of information which students are required to learn and the methods for obtaining this may present social science students with new experiences. In the absence of prior indicators of their aptitude for this type of work, it is understandable that the performance of students may vary considerably.

A further reason why A-levels may not discriminate between students in some subjects or at some institutions is the increasingly high proportion of students who now attain A grades. This is particularly a problem for the more selective universities, where applicants may be predicted As in all their A-levels (Clare, 1999). In this situation, A-levels will have absolutely no predictive power.

A number of proposals have been made and developments undertaken to address this issue. One of these is to release actual A-level scores to universities, instead of just grades (Tate, 2000). This possibility, suggested by Nick Tate, then Chief Executive of the Qualifications and Curriculum Authority, acknowledges that A-level grades may be too broad to adequately discriminate between candidates. If A-level scores were released, it would be necessary to

acknowledge the degree of error in A-levels, so that this could be taken into account. Admissions tutors would also need to understand this concept and make considered judgements in light of the inherent error in any test or assessment. Releasing actual scores would also raise the issue of comparability in scores between exam boards, and between different sittings of an exam. From 2002, the 'Advanced Extension Awards' will be available, and will provide an alternative way of distinguishing between very able students. These are 'qualifications based on A-level subject criteria but testing conceptual understanding and critical thinking to a higher level' (Tate, 2000). The extent to which these proposals can resolve the current difficulties faced by admissions tutors in British universities remains to be seen.

Although this discussion has focused on the ability of A-levels to predict university performance, a distinction needs to be drawn between prediction and selection. Prediction is essentially a statistical issue, and as has been seen in Section 3.5, problems such as low reliability and restriction of score range may account for the low predictive validities seen in much research. If the function of A-levels is as a selection tool, then maybe they perform somewhat better. For example they may screen out low-attaining students, suggesting that they are not suitable for studying at degree level. A-levels can also act as a guide to students, suggesting which universities are most appropriate for them given their predicted or attained grades, and universities also use them as the basis for selecting students. As no adequate investigation of A-levels as selective, rather than predictive, tools has been conducted, their true power as selection instruments is not known. However, by effectively increasing the discrimination provided by A-levels and related exams, the proposed changes discussed above could help to increase the utility of A-levels as selection tools.

Within the American system, there is far less standardisation of assessment than there is in Britain, as there are no national assessments in high school. Students are also far less selected by the time they apply to higher education institutions. The SAT is viewed as one way of overcoming the lack of national assessment, as it provides a common measure against which all university and college applicants can be assessed. These differences between the American and British systems may also explain the low utility of the TAA identified by Choppin and Orr (1976). Findings from the latest SAT data have shown this adds an additional 6.4 per cent to the prediction of college grades over HSGPA (Bridgeman *et al.*, 2000). Choppin *et al.*'s (1973) work with the TAA, which was closely modelled on the SAT, showed the incremental validity of this to be negligible.

If British students are generally more selected than their American counterparts by the time they apply to university, this could explain these findings, as much of their initial screening has already been done through the education system. This would also explain why the simple correlations between the TAA and degree performance seen by Choppin were less than comparable figures for the SAT. A final point is that although students in Britain may in the past have been more selected, the increasing number of students going through the education system, and the variety of routes through which higher education can now be accessed, suggests this is changing. This increase in variety of prior experiences and qualifications, will make it more difficult to select which students have the greatest potential for higher education. Under these circumstances a generally applicable method of assessing aptitude may be desirable, although it does not follow that the SAT is the appropriate model to base this on.

5.3 Current developments in British universities

Immediately prior to this review being conducted, considerable media attention was focused on the university admissions system in Britain. Much of this attention concerned the under-representation of students from state schools in what had been judged to be the 'top' universities on the basis of league tables published in the national press (The Sutton Trust, 2000). This discussion also highlighted the difficulties admissions tutors faced when making decisions, due to the increasing proportion of students attaining A or B grades at A-level. Because of the insufficient discrimination between university applicants given by A-levels, particularly for applicants to the most competitive institutions, the need for a way of identifying particularly able students was highlighted. Moreover, whatever this method was, it was clear that it should be able to identify students' ability to succeed at university irrespective of their previous educational experiences and social circumstances.

In response to this, a number of newspaper articles cited the use of the SAT in the United States as a potential solution (e.g. Clare, 1999; Wolchover, 2000). It was argued in these newspaper articles that the SAT was able to identify potential for study at university level irrespective of social background. In doing so it would give admission tutors the additional information they required to discriminate between students with comparable predicted A-levels. A further advantage with the SAT was that it could be taken and scored before students applied to universities, so reducing the reliance on predicted grades.

Although subsequently media opinion turned somewhat against the SAT (e.g. Richardson, 2000; Lewis, 2000), and a review by QCA recommend alternatives to introducing the SAT (Stobart, 2000), this debate revealed that a number of institutions were already taking steps to address the admissions problem. As part of this review, interviews with people involved with the admissions process at three higher education institutions were conducted, to obtain a picture of developments in this area.

5.3.1 Dr Jane Mellanby, Department of Experimental Psychology, University of Oxford

5.3.1.1 Background

The current work arose from the report of the Vice Chancellor's working party on Access. As part of this, Dr Jane Mellanby and Professor John Stein were asked to give their views of widening access to the University of Oxford. Professor Stein spoke about testing for innate ability, whilst Dr Mellanby focused more on the need to target children far earlier in their education, for example around the age of 11, if a large number of able children are not to be lost from higher education. One of the outcomes of the Vice Chancellor's report was to recommend the development of a pilot assessment which could be used to identify able students who may not be identified through current procedures. This work has subsequently been undertaken by Dr Mellanby and Professor Stein.

5.3.1.2 Trial test and initial findings

The test that was developed for the pilot study consisted of stimulus material adapted from an article in *Nature*, followed by five open-response questions. The questions were designed to assess ability to interpret the material presented and to think beyond it by considering ways in which it could be further developed and the findings additionally tested. The goal was to

develop an assessment of fluid rather than crystallised intelligence, which would measure a student's abilities to think flexibly and creatively when faced with a problem.

The test required a basic level of comprehension, but not essay writing skills, which may be associated with factors such as school experience. Through assessing problem-solving and abilities to think beyond what is immediately given, the test was seen to be assessing qualities generally viewed as important for success at Oxford across a range of courses (with the possible exception of maths and physics, where mathematical ability is paramount). The underlying belief behind the test is that it is possible to assess 'aptitude' for succeeding in a degree course, independently of an applicant's prior experiences. It was reported that the use of existing IQ tests to do this was rejected, due to the tendency for them to be too culturally loaded.

The test was trialled initially on approximately 150 applicants to four Oxford colleges. A fifth college was due to participate but withdrew due to concerns over an unexpected test causing distress to some students, despite it being stressed that participation in the study was voluntary and unrelated to the admissions process. The test was untimed, and it was reported than most students completed it in around 20 minutes. Applicants were also asked to complete a questionnaire which assessed their learning styles. This questionnaire was not intended to be used in conjunction with the test described above as part of the selection process, but was used at this stage only to gather validity evidence for the newly developed test.

Data from the test was still being analysed at the time the interview was conducted, although some initial results were available. These indicated that the scores of pupils who came from comprehensive schools were not significantly different from those from independent schools, whereas the GCSE results were lower in pupils from comprehensive schools. Test scores were also generally unrelated to GCSE results, suggesting that the aptitude test was measuring something distinct from academic attainment. Dr Mellanby reported that the A-level results from the students who took part in the trial were also going to be collected, so that the links between these and test scores could be examined.

The questionnaire on learning styles provided some initial validity information for the newly developed test. Total test score did not correlate with a deep learning style - that is, the tendency to make connections between various elements of a subject and obtain a fuller understanding of it - but two test questions did. The first of these questions required test takers to describe an experiment they could conduct to test the validity of the findings described, and the second was about the generalisability of the conclusions that could be drawn from the stimulus material. Surface learning - the tendency to just learn the necessary facts - was unrelated to test scores.

The results from this trial of the test were not used in the actual admission process, but it was possible to relate findings to admission outcomes. It was found that of the nine highest scorers, eight were accepted to the university. When test scores were combined with learning style, it was seen that 80 per cent of those who had high test scores and a deep learning style were accepted, providing further evidence of the potential value of the test.

Overall, this initial trial was seen as promising. The next stage is to extend the trialling of the test and to obtain evidence on its predictive validity.

5.3.1.3 Planned research

During the autumn term of 2000, it is planned to test 1,000 students in the second year of their A-levels. Test takers will be applicants to the University of Oxford, and the test will be overseen by an administrator employed to work on this research. The co-operation of individual departments and/or colleges within the University of Oxford will be needed for this work, and so only applicants to certain subjects will be studied. The test used in this research will be similar to the one already piloted and will be designed to measure the same constructs, but the subject matter will be changed.

It is intended to follow all students who take the test over the next four years of their education, whether or not they are successful in their applications to Oxford. Associations between test scores and a number of factors will be studied, including GCSEs and A-levels, and academic performance throughout their time at university, as will the effects of gender and prior schooling (e.g. state versus independent school). A range of additional, non-academic measures will also be obtained from those students who gain places at Oxford, and analysed in conjunction with the test data.

5.3.1.4 Potential use of the test and other issues

The intended purpose of the test is to provide information to supplement that which is regularly given by applicants to the University of Oxford. It was suggested that this information would not be routinely used for admissions, although all applicants would probably be asked to take the test. For example, test results could be useful when used in conjunction with school league tables, and could lead to the identification of students with high potential from low-performing schools. An example was also given of an applicant who was extremely nervous and tearful during an admissions interview, so making it very difficult to obtain the necessary information. In situations such as this, and less extreme situations where interviewees are nervous and so may come across poorly in an interview, having information from an additional test could be particularly useful.

Although it was stated that test results would probably not be used for all applicants, some admissions tutors at Oxford feel that additional information on students is needed, as it is generally not possible to discriminate between applicants on the basis of their predicted A-level grades alone. It is currently unclear whether test results could be seen as a way of discriminating between applicants by some admissions tutors, so leading them to apply test results in inappropriate ways.

If testing of all applicants was adopted, a number of issues would need to be considered. Firstly, it was suggested during the interview that the constructs the test measures are appropriate for study in an institution such as the University of Oxford, and the test has been developed specifically with Oxford's needs in mind. The extent to which it is also appropriate for other institutions is not yet clear. If not appropriate for other institutions, this could lead to a system where each university has its own tests, and students have to complete a different admissions test for each university they apply to. It is just such a system that led to the establishment of the SAT over 70 years ago, so co-ordination between universities would be necessary to avoid overloading students with admissions tests. The existence of multiple tests would also undoubtedly raise issues over equivalence, both between universities and within universities over time.

Secondly, it was reported that the marking of the 150 tests used in the trial took one person around three days. This suggests that testing all applicants would incur considerable costs in marking, let alone costs for test development, administration, etc. If multiple-choice questions could be used, they would allow computer marking and so reduce costs, but it is currently unclear whether this format is capable of assessing the necessary construct(s).

5.3.2 Professor Michael Worton, Vice-Provost, University College London

5.3.2.1 Background

The work being undertaken at University College London (UCL) intends to develop a series of tests to supplement the information currently obtained from applications and interviews. UCL routinely interview all British students before offering them places, unless they would have exceptional difficulties in attending. The interview is seen as very important because it not only provides admissions tutors with valuable information about applicants, but also allows applicants to make a judgement as to whether the course and wider university is suitable for them. Through the information gained from interviews, UCL is able to make a more informed selection of students. The drawback of this process is that it is labour-intensive and places an additional burden on admissions tutors. Due to this it was reported that admissions tutors have questioned whether there is a need to interview all applicants, although efforts are being made to ensure interviews continue.

It is recognised that even with appropriate staff training, interviews are often quite subjective. The tests currently being developed offer a potential way of obtaining information on the personal characteristics of applicants that is more resistant to bias. In addition to the test development work, UCL is also looking at other methods that could be valuable in furthering equality in admissions. For example, it was reported that the use of postcodes as an indicator of socio-economic status was being explored, and the needs of different ethnic minorities considered. The importance of early educational opportunities, long before students may consider attending university, is also recognised.

In discussing admissions policies more generally, it was felt that current Government initiatives had created a degree of conflict for universities. Specifically, whilst the emphasis on widening access could lead to average A-level grades of those admitted falling, this was seen as being in conflict with the publication of university league tables. It was felt that without careful explanation and publication of policies regarding access, broadening access could lead to league tables being interpreted as indicating a lowering of standards.

5.3.2.2 Trial tests and planned research

At the time of the interview, the tests UCL plan to pilot were still under development, and so only an overview of them can be given. The tests are being developed by a company specialising in psychological testing, and are primarily designed to assess applicants' personality, although they will also include a test of ability. The personality tests are designed to measure factors such as perseverance, boredom threshold, team-working and other aspects of personality important to the successful completion of degrees at UCL. Due to their importance, admissions tutors tend to look for evidence of these characteristics during interviews with potential students, but the tests are seen as a way of increasing the objectivity of this assessment.

The tests are computerised and will take about 45 minutes to complete. It is intended that the tests will require no specific preparation, so as to deter students being tutored specifically for them, and there will be no pass/fail criteria.

During the autumn term of 2000, it is planned to trial the tests on approximately 40 medical students, 40 engineering students and 40 humanities/arts students.

5.3.2.3 Application of the tests

The tests will be used in conjunction with predicted or attained A/AS-level grades and interviews, effectively giving admissions tutors a third source of information. It is believed that many of the personal characteristics which the test battery assesses are already assessed informally during the admissions interview, but the tests will make this process more objective. The tests should also give a greater degree of consistency across applicants than interviews can.

It is recognised that the test results need to be set in the context of different courses - there is no personality profile for the 'ideal' student. Instead, the different methods of teaching and learning within each subject need to be made explicit, and matches made between the styles and demands of courses and the personal characteristics of students. Currently, Departments within UCL are required to produce academic strategies, and these should provide valuable information that can be used in this process. However, it was emphasised that each student must still be assessed on their individual merits, and personal characteristics interpreted liberally, so as to prevent a situation which filters applicants according to personality and prevents diversity.

5.3.2.4 Outstanding issues

Due to UCL's tests being in an early stage of development, there are many outstanding issues. The initial research which is planned for autumn 2000 will answer some of the most fundamental questions, primarily by looking at the associations between test results and course performance. If the test trials prove successful and the tests are seen to possess sufficient predictive validity, the following points will need to be addressed before they could be used as part of the admissions process.

The first of these involves the timing of the testing and when the results would be combined with the other sources of information on each applicant. For example, should the information be available to admissions tutors so that it can be brought into interviews and further explored if necessary? Alternatively, should the interview be conducted without knowledge of the test results? Regardless of when the different sources of information on each applicant are integrated, there would be training implications for admissions tutors, and dealing with the tests results would inevitably place a further burden on them. It would be particularly necessary to consider how the personality data is interpreted and what skills are needed to do this competently.

A second issue concerns the need to give feedback to students. Feedback would need to be handled particularly sensitively in the case of rejections, as students should never be made to feel that they have been rejected on the basis of their personality. Best practice would suggest that all feedback of test results is done face-to-face. At the very least applicants should be given the opportunity to talk to someone about their results via telephone, if they feel this is necessary. This would again place further burden on university staff.

A final point concerns the need to match aspects of each course's teaching and learning environment to the personal characteristics of potential students. As courses develop, methods of teaching are likely to change. This is particularly the case at UCL where considerable emphasis is placed on innovative teaching. If major changes in teaching occur, it would be necessary to re-evaluate the link between the teaching environment and student characteristics, to determine if the two still match adequately.

5.3.3 Professor Dylan Wiliam, Head of the School of Education, King's College London

5.3.3.1 Background

Professor Wiliam's current work on university entrance focuses on access to medicine. One of the concerns of King's College Medical School is the low proportion of applicants from some ethnic groups, particularly Afro-Caribbeans. This is of particular concern as there is a centre for Afro-Caribbean medicine at the college, and evidence suggests that this group may have specific needs in terms of medical care.

Work is currently being planned in an attempt to improve recruitment to the Medical School. This will currently target the geographical area around King's College, which has a high ethnic minority population. As part of this work, medical students will be involved in raising the profile of medicine and providing information about the necessary entrance requirements, through visits to local schools. Guidance on the selection of A-level subjects is particularly important, as students cannot be admitted to medicine without A-level chemistry.

It is hoped that in the longer term this type of approach will redress the recruitment problem, but it is recognised that this will take time to have the desired impact. In the mean time, alternative methods of selecting students are being studied.

5.3.3.2 Testing for Medical School and enhancing access to medicine

As part of the 'Access to Medicine' programme, King's College is planning to extend the initial part of their medical training, which currently involves two years' academic study, primarily in the area of science. From 2001, additional places have been made available for students admitted through the Access to Medicine programme. For those students admitted through this route, the science course will be extended from two to three years, and will include more support for students during this time. This is being done as difficulties with this part of the course have been identified as one of the barriers to greater diversity in the student population. Students who are not selected through the Access to Medicine programme will take the standard two years of initial training.

Whilst it is hoped that using students to promote medicine in schools, coupled with the changes to the science programme, will improve recruitment from currently under-represented groups, it will still be necessary for selection of these students to take place. A-levels are not seen as appropriate for this purpose due to the problems of range restriction, and instruments such as the SAT have been rejected as they are considered to be too culturally loaded. In line with the initial emphasis on science in the first years of medical training, an appropriate test would be one which could assess candidates' potential for science learning. It is emphasised

that any chosen test should be assessing 'potential' rather than current academic attainment, as this will tend to be confounded with educational background and so fail to redress the balance in the selection of applicants from minority groups. However, it is acknowledged that such a test will not be completely independent of prior education.

The tests which have been identified as suitable for this purpose are based on the science reasoning tests developed by Shayer and Adey (1981). Although initially developed for children in secondary schools, two of the most demanding tasks are of the appropriate difficulty to discriminate between applicants to the medical school. These tests are dynamic, that is, an administrator provides demonstrations to the test taker with the test materials, and the test taker then has to use the knowledge they have gained from this to answer subsequent questions. Validity work has been conducted on these tests, and they are known to be a good indicator of general intellectual functioning and to predict academic attainment. The science reasoning tests can be administered to small groups of students, and each of the two tests takes around 50 minutes to complete.

5.3.3.3 Planned research and developments in access to medicine

The first stage of the research, planned by King's for autumn 2000, involves the calibration of the two science reasoning tests on existing first- and second-year medical students. Examining the association between test scores and exam performance in these students will provide further validity data on the tests and will also allow a minimum threshold for the selection of subsequent students to be set.

It is planned to use the science reasoning tests in the admissions process to the Medical School for the first time with the autumn 2001 intake. These will be used as part of the admissions process only with students from areas surrounding King's College, who can apply through the Access to Medicine programme. All other students will follow the established route. Students who apply through the Access to Medicine programme will still be screened through predicted A-level grades, and all those who are successful will be asked to take the science reasoning tests. Keeping the A-level criteria is not seen to conflict with the goal of widening admissions, due to the increasingly high general A-level performance.

For the course starting in 2001, it is planned to make a limited number of additional places available in the Medical School for students on the Access to Medicine programme. Admission to these places will be influenced by performance on the science reasoning tests. However, it was emphasised that these tests will not be the sole determinant of selection, but that they have the potential to provide further information on students which can be used in conjunction with predicted A-levels and interviews. Students selected through this route will study an extra year of science, as outlined above. At the end of their first-year, the performance of those students admitted partially on the basis of their science reasoning test scores will be compared with those of students accepted through the traditional route. This will provide the first full test of the validity of the science reasoning tests for selection into medicine.

Further developments include an investigation into the utility of non-cognitive predictors. This work will involve a literature review and research using data which has already been collected by King's College. The data, which includes factors such as family background, motivation and personality, has been collected since 1994, but has not been tied up to students' performance in medical school. By doing this, the potential for developing further assessments which may be of value in selecting medical students will be explored.

Although it was felt that the predictive power of any non-cognitive factors is likely to be small, one area in which they may be useful is in reducing the dropout when students start their clinical training. One relatively common finding is that although medical students may be successful in their initial academic training, they encounter difficulties when they move into the clinical part of their course, due to the very different nature of this work. It was suggested that this might be due to some students not having a complete understanding of what being a doctor will actually involve. Potential indicators of how students will adapt to their clinical experiences could therefore be useful in predicting retention.

As the work at King's College is still at an early stage, a number of issues remain to be resolved. The most significant of these is the extent to which the science reasoning tests are able to predict success in medical training, and the extent to which they can do this independently of the information that is routinely available on students. It is expected that test results will show some association with A-levels, particularly those in science disciplines, but they will also need to be moderately distinct from these if they are to be of use in selection. This is a further issue that remains to be determined, as is whether these tests provide a fair assessment of potential in all groups. As the exploration of non-cognitive factors is also to be completed, the potential use of these for selection also remains to be determined.

The use of the science reasoning tests is part of a broader programme to increase access to medicine for those groups currently under-represented in the Medical School. It is hoped that this can be achieved through using medical students to raise the profile of medicine and to provide advice and guidance to potential applicants. The success of this initiative will take a number of years to evaluate fully.

6. Conclusions

6.1 Reasons for interest in the SAT

This final part of the review brings together the main findings from the previous sections, and discusses aptitude testing specifically in relation to the current system of admissions to British universities.

An appropriate starting point for this discussion is probably to consider why the debate on aptitude testing for university entrance in Britain arose. Media attention focused on statistics published by The Sutton Trust (2000) showing that students from independent schools were over-represented in the universities ranked highest according to newspaper league tables, as were students from the higher social classes. This degree of bias could not be justified purely on the basis of A-level attainment, suggesting that certain universities were biased against admitting students from state schools.

As part of this debate, the difficulties admissions tutors face when selecting between applicants were also highlighted. In universities which attract high numbers of well-qualified candidates, it is often not possible to discriminate between them on the basis of their predicted A-level grades. The increasing number of students predicted to attain top grades at A-level and actually achieving these has at least partially caused this problem. For example, in the year 1998/9, 40.7 per cent of 17-year-olds who took A-levels achieved grades A or B (GB. DfEE, 2000). In the absence of adequate evidence to make selection decisions, admissions tutors may favour students from schools known to consistently produce high-achieving students, and whose students have previously been successful at the university. If these schools are more likely to be independent schools, this at least partially accounts for the over-representation of students from independent schools in some of the most competitive universities.

Rather than seeing the SAT as the ideal solution to this problem, it has been suggested that admissions tutors are in favour of any information which would provide better discrimination between applicants than A-levels currently do (e.g. Clare, 1999; Tate, 2000). As the SAT gives scaled scores, this would at least allow greater discrimination than is presently possible through A-level grades. This promise of greater discrimination was also accompanied by the view that the SAT was able to identify potential for university study, independently of school experiences or social background (e.g. Clare, 1999). Whilst appealing to those who make decisions, there was a dearth of evidence presented in this debate on whether the SAT discriminated between students on constructs that were important to university success, or whether its scores were actually independent of factors such as social background.

A further appeal of the SAT was that results from it could be available when students applied to university. This would reduce the current reliance on predicted A-level grades with their considerable unreliability. However, it is important to recognise that the availability of attained rather than predicted grades is not an inherent feature of the SAT, merely a result of when it can be taken in relation to university applications. The same situation could be achieved by bringing A-levels forward to earlier in the academic year (e.g. Charter and Baldwin, 2000).

6.2 Aptitude testing in the British education system

The appeal of tests such as the SAT appears to lie in them helping admissions tutors to make selection decisions, and in the promise of 'fairer' assessment. The College Board initially introduced the SAT with the goal of standardising the admissions procedure to colleges in the United States, as many colleges had their own processes and tests. Although relatively successful in achieving this goal, there is now considerable variation in how SAT and ACT scores are used in admission to colleges, with them not being considered at all in some cases (see Section 2.2.3). This suggests that if the goal is to obtain a diverse, representative student population in systems which take account of individual students' needs, it may not be possible to have a fully standardised admissions procedure. At least, the vision that the College Board had for the United States has been seen to be not fully capable of achieving these goals.

In Britain, the university admissions process for undergraduates is largely standardised through The Universities and Colleges Admissions Service (UCAS), although some variation in the system still exists (e.g. some universities and/or courses require students to attend interviews, whereas others do not). However, as was seen in Section 5.3, a number of universities are looking at additional methods of assessing students. At present, this work is being conducted on a local level, either within universities or within departments. If a number of institutions start to use additional assessments in their selection, this would result in the system being less standardised as universities require applicants to take their own admissions tests in addition to the UCAS procedure.

It is questionable whether a situation where many universities have their own admissions tests is desirable, but this may depend on the purpose of these tests. If they are used simply as a way of discriminating between students who have been predicted similar A-level grades, then this may not be the best way forward. The Advanced Extension Awards offer an alternative way to discriminate between students, as they are designed for the most able A-level students. These will be available from 2002, and it is likely to take a number of years before their usefulness can be properly evaluated. An alternative suggestion is to use the marks from A-level papers instead of grades (Tate, 2000), but this raises difficult issues of measurement reliability and comparability between exam boards.

It is further worth considering that British students are generally more selected than their counterparts in the United States, by the time they come to apply for university. This selection takes place through a national examination system, again in contrast to the United States where students do not take exams on a national level. However, it is worth noting that some critics of the SAT have argued that this should be replaced by attainment tests (e.g. Crouse and Trusheim, 1988). Crouse and Trusheim cited the College Board's Advanced Placement Program as a possible model for this, which with established course descriptions, shares a number of similarities with A-levels.

This touches on a further goal of the SAT, to act as a national standard in the absence of this in the high school system. There would be no need for a British aptitude test to act as a national standard, as this role is already fulfilled by qualifications such as A-levels. Indeed, claims that tests like the SAT are no more than high-level intelligence tests that may measure fixed attributes (see Jencks, 1998) suggest that any such benchmarking of IQ may have negative consequences, particularly if used in a negative, discriminatory sense. Exams such as A-levels have the advantage, at least from an ethical perspective, of measuring 'attainment', and so do not imply fixed ability.

Admissions tests will be more desirable if they aim to assess characteristics important for the successful completion of degrees, which are not measured by A-levels. This appears to be closer to the goal of current developments in a number of British universities (see Section 5.3). Both Oxford and King's College London are exploring tests to assess mental abilities considered necessary for study at degree level. Whilst the tests being developed for UCL also include a test of ability, their main focus is on providing a more objective assessment of students' personal qualities to supplement the information obtained through interviews. In all cases, the hope is that the tests will provide information in addition to A-levels, to indicate which students are likely to benefit most from higher education.

Research on the ability of A-levels to predict university performance is limited, with far fewer studies having been conducted than in the United States on the predictive validity of the SAT and high school record. Generally, there has been much more open debate in the United States on the college system than there has in Britain, particularly on the ability of high school record and admissions tests to predict success. What evidence is available suggests that on average, A-levels appear to be able to account for around eight per cent of the variation in degree class (e.g. Peers and Johnston, 1994). Although this is a very modest prediction, it does not provide a fair reflection of the utility of using A-levels to select students for university. This is because it may not be the A-level grades themselves which are important, but rather that completing A-levels shows the ability, commitment and motivation necessary for a degree course. There will be much more to completing any course of study than ability, although attainment may be the only factor seen to be directly reflected in an exam grade. Whilst other factors such as motivation may be important predictors of aptitude for university education, they are yet to be adequately researched.

6.3 The consequences of aptitude testing

The need to carefully consider the potential consequences of any new testing system has been increasingly recognised since Messick's (e.g. 1989) influential work on consequential validity. Although the exact consequences of any revision to the education system are hard to evaluate, it is possible to speculate on what the effects of introducing a university admissions test in Britain may be.

One of the most obvious effects is likely to be a reduced focus on A-levels, as an admissions test would need to be taken during the course of A-level study. This makes the assumption that students would divide their time between preparation for the admissions test and A-level work, rather than maintaining the time spent on A-levels and finding additional time for test preparation. Although the truth may lie somewhere in the middle, some detraction from A-levels appears inevitable.

If students spend less time on their A-levels, this may also have consequences when students start university. Some university courses require students to have taken specific A-levels as part of their entrance requirements, and so rely on these courses to furnish students with the fundamental knowledge necessary for the degree course. If students have been able to acquire less of this knowledge through their A-level studies, this may have an effect on their degree performance. In courses which do not require students to have studied specific subjects, the effects of preparing for an aptitude test may be less noticeable.

Much of the burden in preparing students for an admissions test is likely to lie with sixth-form tutors, again because it is during this time that students are likely to sit the test. It is important to consider the potential effects of this on tutors, who already have high workloads. Placing

additional demands on sixth-form tutors may have a negative effect on morale and recruitment in an area where it is already difficult to find adequate staff.

As has been seen from countries such as the United States and Israel, a considerable coaching industry has grown up alongside the SAT and the PET. It is almost inevitable that a parallel industry would develop in Britain if a university admissions test was introduced. Whilst the effects from coaching have been seen to be modest, they none the less exist, and may be most noticeable in highly selective institutions which require high test scores. This may again bias access to prestigious universities in favour of those who can afford to pay for additional preparation, although adequately disseminating the true effects of coaching and its quickly diminishing returns could do much to pre-empt the development of coaching courses.

A further issue concerns the cost of the admissions test development and administration, and who would meet this. In the United States the costs are kept down through the large volume of students who take the SAT and ACT at each sitting. In Britain, the numbers of students would be far less, meaning that it would probably cost considerably more or that the costs would need to be at least partially met through other means. As scores have been seen to improve on simply retaking the SAT, this suggests that poorer students may be again disadvantaged, if they have to pay to take the test. Alternatively, costs could be met by colleges, as the costs of A-levels often are, but such an expense would place further strains on budgets that are often already stretched.

The financial considerations outlined above touch on the issue of fairness in testing for university admissions. As has been seen in Section 3, the issue of bias in testing is very complex and it is difficult to come to any definite conclusions. Emerging themes from the literature are that Blacks and other ethnic groups often score considerably lower on tests such as the SAT and ACT than White students, with this difference being up to one standard deviation in some cases. This observation is not unique to these tests, as despite attempts to remove this difference, Blacks consistently score lower on a range of tests that assess aspects of intelligence. The evidence on whether these differences are a fair reflection of the subsequent college performance of these ethnic groups is inconclusive.

A second finding concerns sex differences on admissions tests in the United States and also Israel. Females consistently score lower than males, but this difference is often not reflected in their college attainment, as they go on to attain comparable or higher GPAs. The patterns of male and female performance on the SAT are also not consistent with their performance on tests measuring comparable constructs, although this may be due to male SAT takers being a more selected group than their female counterparts. The issue of bias has proved very controversial in the United States, and any admissions test introduced in Britain is unlikely to escape this.

If an admissions test gave a scaled score for each student, this may go some way to reducing the problem of discriminating between students outlined above. However, the most important issue is whether it would provide an accurate prediction of subsequent university performance. All available evidence shows that the ability of admission tests to predict performance is limited. Although there are limitations in the statistical methods used to estimate prediction, there are likely to be many other reasons for this. Factors not measured by aptitude tests such as the ability of students to adjust to college life have been proposed as important, but little focused research has been conducted in this area.

In light of these findings, the utility of an admissions test, particularly when set in the context of the British education system, which already involves national assessments, is clearly open

to debate. Evidence from the United States shows that overall prediction is modest, and can vary considerably between different institutions. This may well result from the extent to which colleges provide support for students who have difficulties adjusting to university. Some evidence of similar variations in the predictive power of A-levels has been seen, and such variations may well occur if an admissions test was introduced.

It is almost inevitable that an admissions test used as part of access into British universities would come under considerable scrutiny, both from education professionals and the general public. Equally inevitable would be the debate over fairness to different social and ethnic groups. Whilst open, public debate over this and other education issues should be welcomed, clear evidence for the value of the test would be needed if it were to be adequately defended. Possibly the strongest evidence against the utility of introducing an admissions test is Bruce Choppin's work in the 1960s and 1970s. However, the British education system has changed in many ways since this work, including offering far more diverse ways of entering higher education, so it is unclear whether these findings remain valid today. It is fundamental that further research is conducted into predictive validity, before any introduction of a national university admissions test.

References

ALLALOUF, A. and BEN-SHAKHAR, G. (1998). 'The effect of coaching on the predictive validity of scholastic aptitude tests', *Journal of Educational Measurement*, **35**, 1, 31 - 47.
AMERICAN COLLEGE TESTING PROGRAM (2000a). *Selections from the 2000 National Score Report: Academic Abilities and Nonacademic Characteristics of ACT-Tested Graduates* (2000 ACT National and State Scores) [online].
Available: http://www.act.org/news/data/00/00data.html [31 October, 2000].
AMERICAN COLLEGE TESTING PROGRAM (2000b). *Test Preparation: Test Taking Strategies* (ACT Assessment) [online].
Available: http://www.act.org/aap/testprep/index.html [10 October, 2000].
BAIRD, L.L. (1987). 'Do students think admissions tests are fair? Do tests affect their decisions?' *Research in Higher Education*, **26**, 4, 373 - 88.
BARON, J. and NORMAN, M.F. (1992). 'SATs, achievement tests, and high-school class rank as predictors of college performance', *Educational and Psychological Measurement*, **52**, 4, 1047 - 55.
BAYDAR, N. (1990). *Effects of Coaching on the Validity of the SAT: a Simulation Study* (ETS Research Report 90 - 4). Princeton, NJ: Educational Testing Service.
BELLER, M. (1994). 'Psychometric and social issues in admissions to Israeli universities', *Educational Measurement: Issues and Practice*, **13**, 2, 12 - 20.
BEN-SHAKHAR, G., KIDERMAN, I. and BELLER, M. (1996). 'Comparing the utility of two procedures for admitting students to liberal arts: an application of decision-theoretic models', *Educational and Psychological Measurement*, **56**, 1, 90 - 107.
BLACKBURN, J.C. (1990). 'No one, including admissions officers, fails the SAT', *College & University*, **66**, 1, 17, 20.
BRAUN, H., RAGOSTA, M. and KAPLAN, B. (1986). *The Predictive Validity of the Scholastic Aptitude Test for Disabled Students* (ETS Research Report 86 - 38). Princeton, NJ: Educational Testing Service.
BRIDGEMAN, B., McCAMLEY-JENKINS, L. and ERVIN, N. (2000). *Predictions of Freshman Grade-Point Average from the Revised and Recentered SAT I: Reasoning Test* (College Board Research Report No.2000 - 1/ETS Research Report 00 - 1). New York, NY: College Board.
BRIDGEMAN, B. and WENDLER, C. (1991). 'Gender differences in predictors of college mathematics performance and in college mathematics course grades', *Journal of Educational Psychology*, **83**, 2, 275 - 84.
BRODNICK, R.J. and REE, M.J. (1995). 'A structural model of academic performance, socio-economic status, and Spearman's *g', Educational and Psychological Measurement*, **55**, 4, 583 - 94.
BURTON, E. and BURTON, N.W. (1993). 'The effect of item screening on test scores and test characteristics.' In: HOLLAND, P.W. and WAINER, H. (Eds) *Differential Item Functioning*. Hillside, NJ: Lawrence Erlbaum Associates.
BURTON, N. (1996). 'Have changes in the SAT affected women's mathematics performance?' *Educational Measurement: Issues and Practice*, **15**, 4, 5 - 9.
CARNOY, M. (1995). 'Why aren't more African Americans going to college?' *Journal of Blacks in Higher Education*, **6**, 66 - 9.
CHAPMAN, K. (1996). 'Entry qualifications, degree results and value-added in UK universities', *Oxford Review of Education*, **22**, 3, 251 - 64.
CHARTER, D. and BALDWIN, T. (2000). 'A levels in April may end scramble', *The Times*, 22 June.

CHOPPIN, B. and ORR, L. (1976). *Aptitude Testing at Eighteen-Plus*. Windsor: NFER Publishing Co.
CHOPPIN, B.H.L., ORR, L., KURLE, S.D.M., FARA, P. and JAMES, G. (1973). *The Predication of Academic Success*. Windsor: NFER Publishing Co.
CLARE, J. (1999). 'SAT – a simple way to grade students' (Education), *The Daily Telegraph*, 10 November, 25.
COLLEGE BOARD.COM (2000). *PSAT/NMSQT* [online].
Available: http://www.collegeboard.org/psat/student/html/indx001.html [31 October, 2000].
COLLEGE ENTRANCE EXAMINATION BOARD (1978). *Taking the SAT: a Guide to the Scholastic Aptitude Test and the Test of Standard Written English*. Princeton, NJ: Educational Testing Service.
COLLEGE ENTRANCE EXAMINATION BOARD (2000). *SAT Math Scores for 2000 Hit 30-Year High; Reflect Gains for American Education. Verbal Scores Hold Steady Amidst Increasing Diversity* (College Board News 2000 - 2001) [online].
Available: http://www.collegeboard.org/press/senior00/html/000829.html [31 October, 2000].
CORLEY, E.R., GOODJOIN, R. and YORK, S. (1991). 'Differences in grades and SAT scores among minority college students from urban and rural environments', *The High School Journal*, **74**, 3, 173 - 7.
CROUSE, J. and TRUSHEIM, D. (1988). *The Case Against the SAT*. Chicago, IL: The University of Chicago Press.
CROUSE, J. and TRUSHEIM, D. (1991). 'How colleges can correctly determine selection benefits from the SAT', *Harvard Educational Review*, **61**, 2, 125 - 47.
DAVIES, S. and GUPPY, N. (1997). 'Fields of study, college selectivity, and student inequalities in higher education', *Social Forces*, **75**, 4, 1417 - 38.
DELAP, M.R. (1994). 'An investigation into the accuracy of A-level predicted grades', *Educational Research*, **36**, 2, 135 - 48.
DELAP, M.R. (1995). 'Teachers' estimates of candidates' performances in public examinations', *Assessment in Education*, **2**, 1, 75 - 92.
FINCHER, C. (1990). *Trends in the Predictive Validity of the Scholastic Aptitude Test* (ETS Research Report 90 - 13). Princeton, NJ: Educational Testing Service.
FLEMING, J. and GARCIA, N. (1998). 'Are standardized tests fair to African Americans? Predictive validity of the SAT in black and white institutions', *The Journal of Higher Education*, **69**, 5, 471 - 95.
FREEDLE, R. and KOSTIN, I. (1990). 'Item difficulty of four verbal item types and an index of differential item functioning for black and white examinees', *Journal of Educational Measurement*, **27**, 4, 329 - 43.
FREEDLE, R. and KOSTIN, I. (1997). 'Predicting black and white differential item functioning in verbal analogy performance', *Intelligence*, **24**, 3, 417 - 44.
FUERTES, J.N., SEDLACEK, W.E. and LIU, W.M. (1994). 'Using the SAT and noncognitive variables to predict the grades and retention of Asian American university students', *Measurement and Evaluation in Counseling and Development*, **27**, 2, 74 - 84.
GALLAGHER, A.M. and DE LISI, R. (1994). 'Gender differences in Scholastic Aptitude Test – mathematics problem solving among high-ability students', *Journal of Educational Psychology*, **86**, 2, 204 - 11.
GANDARA, P. and LOPEZ, E. (1998). 'Latino students and college entrance exams: how much do they *really* matter?' *Hispanic Journal of Behavioral Sciences*, **20**, 1, 17 - 38.
GARDNER, H. (1993). *Frames of Mind: the Theory of Multiple Intelligences*. London: Fontana Press.

GOULD, E.M. and M'COMISKY, J.G. (1958). 'Attainment level on leaving certificate and academic performance at university', *British Journal of Educational Psychology*, **28**, 2, 129 - 34.
GRAFF, A.S. (1993). 'The new SAT: the future of transition assessment', *Education Libraries*, **17**, 3, 7 - 13.
GREAT BRITAIN. DEPARTMENT FOR EDUCATION AND EMPLOYMENT (2000). *Statistics of Education: Public Examinations GCSE/GNVQ and GCE/AGNVQ in England 1999.* London: The Stationery Office.
HAMP-LYONS, L. (1997). 'Washback, impact and validity: ethical concerns', *Language Testing*, **14**, 3, 295 - 303.
HAND, C.A. and PRATHER, J.E. (1985). 'The predictive validity of scholastic aptitude test scores for minority college students.' Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, 31 March - 4 April.
HARRIS, A.M. and CARLTON, S.T. (1993). 'Patterns of gender differences on mathematics items on the Scholastic Aptitude Test', *Applied Measurement in Education*, **6**, 2, 137 - 51.
HENRIKSSON, W. and WOLMING, S. (1998). 'Academic performance in four study programmes: a comparison of students admitted on the basis of GPA and SweSAT scores, with and without credits for work experience', *Scandinavian Journal of Educational Research*, **42**, 2, 135 - 50.
HERRNSTEIN, R.J. and MURRAY, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. New York, NY: Bell Press.
HYDE, J.S., FENNEMA, E. and LAMON, S.J. (1990). 'Gender differences in mathematics performance: a meta-analysis', *Psychological Bulletin*, **107**, 2, 139 - 55.
HYDE, J.S. and LINN, M.C. (1988). 'Gender differences in verbal ability: a meta-anlysis', *Psychological Bulletin*, **104**, 1, 53 - 69.
JENCKS, C. (1979). *Who Gets Ahead? The Determinants of Economic Success in America*. New York, NY: Basic Books.
JENCKS, C. (1998). 'Racial bias in testing.' In: JENCKS, C. and PHILLIPS, M. (Eds) *The Black-White Test Score Gap.* Washington, DC: The Brooking Institution.
JONES, L.V. (1994). 'Perspectives on educational testing: discussion', *Educational Measurement: Issues and Practice*, **13**, 2, 28 - 30.
KANOY, K.W., WESTER, J. and LATTA, M. (1989). 'Predicting college success of freshmen using traditional, cognitive, and psychological measures', *Journal of Research and Development in Education*, **22**, 3, 65 - 70.
KELSALL, R.K. (1963). 'University student selection in relation to subsequent academic performance: a critical appraisal of the British evidence', *Sociological Review*, **7**, 99 - 115.
LAWLOR, S., RICHMAN, S. and RICHMAN, C.L. (1997). 'The validity of using the SAT as a criterion for black and white students' admission to college', *College Student Journal*, **31**, 4, 507 - 15.
LAWRENCE, I.M., LYU, C.F. and FEIGENBAUM, M.D. (1995). *DIF Data on Free-Response SAT I Mathematical Items* (ED 389 742). Princeton, NJ: Educational Testing Service.
LEWIS, S.D. (2000). 'a) frying pan or b) fire?'(Education), *The Guardian*, 6 June, (insert, 12 - 13).
LINN, M.C. and HYDE, J.S. (1989). 'Gender, mathematics, and science', *Educational Researcher*, **18**, 8, 17 - 19, 22 - 27.
LINN, R.L. (1983). 'Testing and instruction: links and distinctions', *Journal of Educational Measurement*, **20**, 2, 179 - 89.
McCRUM, N.G. (1996). 'Gender and social inequality at Oxford and Cambridge universities', *Oxford Review of Education*, **22**, 4, 369 - 97.

McDONALD, A.S., NEWTON, P.E. and WHETTON, C. (2001). *A Pilot of Aptitude Testing for University Entrance*. Slough : NFER.
MELLANBY, J., MARTIN, M. and O'DOHERTY, J. (2000). 'The "gender gap" in final examination results at Oxford University', *British Journal of Psychology*, **91**, 3, 377 - 90.
MELLANBY, J.H. and RAWLINS, J.N.P. (1997). 'The gender gap – the case of PPP', *Oxford Magazine*, **44**, 2, 2.
MESSICK, S. (1989). 'Meaning and values in test validation: the science and ethics of assessment', *Educational Researcher*, **18**, 2, 5 - 11.
MESSICK, S. (1995). 'Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning', *American Psychologist*, **50**, 9, 741 - 9.
MESSICK, S. and JUNGEBLUT, A. (1981). 'Time and method in coaching for the SAT', *Psychological Bulletin*, **89**, 2, 191 - 216.
MINISTRY OF EDUCATION (2000). *New University Admission System* (EDUN N25-02-004) [online]. Available: http://www1.moe.edu.sg/press/pr23102000adm.htm [31 October, 2000].
MINISTRY OF EDUCATION. COMMITTEE ON UNIVERSITY ADMISSION SYSTEM (1999). *Preparing Graduates for a Knowledge Economy: a New University Admission System for Singapore.* Singapore: Ministry of Education.
MINKE, A. (1996). 'A review of the recent changes in the Scholastic Aptitude Test I: Reasoning Test.' Paper presented at the Annual Meeting of the Southwest Educational Research Association, New Orleans, Louisiana, 26 January.
MOFFATT, G.K. (1993). 'The validity of the SAT as a predictor of grade point average for nontraditional college students.' Paper presented at the Annual Meeting of the Eastern Educational Research Association, Clearwater Beach, Florida, 17 - 22 February.
MORGAN, R. (1990). *Predictive Validity within Categorizations of College Students: 1978, 1981, and 1985* (ETS Research Report 90 - 14). Princeton, NJ: Educational Testing Service.
NATHAN, J.S. and CAMARA, W.J. (1998). *Score Change When Retaking the SAT I: Reasoning Test* (Research Notes RN-05). New York, NY: College Board.
NEISSER, U., BOODOO, G., BOUCHARD, T.J., BOYKIN, A.W., BRODY, N., CECI, S.J., HALPERN, D.F., LOEHLIN, J.C., PERLOFF, R., STERNBERG, R.J. and URBINA, S. (1996). 'Intelligence: knowns and unknowns', *American Psychologist*, **51**, 2, 77 - 101.
NISBET, J. and WELSH, J. (1966). 'Predicting student performance', *University Quarterly*, September, 468 - 80.
PEARSON, B.Z. (1993). 'Predictive validity of the Scholastic Aptitude Test (SAT) for Hispanic bilingual students', *Hispanic Journal of Behavioral Sciences*, **15**, 3, 342 - 56.
PEERS, I.S. (1994). 'Gender and age bias in the predictor-criterion relationship of A levels and degree performance: a logistic regression analysis', *Research in Education*, **52**, 23 - 41.
PEERS, I.S. and JOHNSTON, M. (1994). 'Influence of learning context on the relationship between A-level attainment and final degree performance: a meta-analytic review', *British Journal of Educational Psychology*, **64**, 1, 1 - 18.
PERFETTO, G., ESCANDON, M., GRAFF, S., RIGOL, G. and SCHMIDT, A. (1999). *Toward a Taxonomy of the Admissions Decision-Making Process: a Public Document Based on the First and Second College Board Conferences on Admissions Models*. New York, NY: College Board.
PILKINGTON, G.W. and HARRISON, G.J. (1967). 'The relative value of two high level intelligence tests, advanced level, and first year university examination marks for predicting degree classification', *British Journal of Educational Psychology*, **37**, 3, 382 - 9.
POWELL, B. and STEELMAN, L.C. (1996). 'Bewitched, bothered, and bewildering: the use and misuse of state SAT and ACT scores', *Harvard Educational Review*, **66**, 1, 27 - 59.

POWERS, D.E. (1993). 'Coaching for the SAT: a summary of the summaries and an update', *Educational Measurement: Issues and Practice*, **12**, 2, 24 - 30, 39.
POWERS, D.E. and ALDERMAN, D.L. (1983). 'Effects of test familiarization on SAT performance', *Journal of Educational Measurement*, **20**, 1, 71 - 9.
POWERS, D.E. and ROCK D.A. (1999). 'Effects of coaching on SAT I: Reasoning Test scores', *Journal of Educational Measurement*, **36**, 2, 93 - 118.
RAGOSTA, M., BRAUN, H. and KAPLAN, B. (1991). *Performance and Persistence: a Validity Study of the SAT for Students with Disabilities* (College Board Report No.91 - 3). New York, NY: College Board.
RECKASE, M.D. (1998). 'Consequential validity from the test developer's perspective', *Educational Management: Issues and Practice*, **17**, 2, 13 - 16.
REUTERBERG, S-E. (1998). 'On differential selection in the Swedish Scholastic Aptitude Test', *Scandinavian Journal of Educational Research*, **42**, 1, 81 - 97.
RICHARDSON, K. (2000). 'New tests, old results', *Times Higher Educ. Suppl.*, **1432**, 21 April, 16.
RICHARDSON, P.H., WINDER, B., BRIGGS, K. and TYDEMAN, C. (1998). 'Grade predictions for school-leaving examinations: do they predict anything?' *Medical Education*, **32**, 3, 294 - 7.
ROBBINS REPORT. GREAT BRITAIN. DEPARTMENT OF EDUCATION AND SCIENCE. COMMITTEE ON HIGHER EDUCATION (1963). *Higher Education* (Cmnd. 2154). London: HMSO.
ROONEY, C. with SCHAEFFER, B. (1998). *Test Scores do not Equal Merit: Enhancing Equity & Excellence in College Admissions by Deemphasizing SAT and ACT Results*. Cambridge, MA: National Center for Fair and Open Testing (Fair Test).
ROSSER, P. (1989). *The SAT Gender Gap: Identifying the Causes*. Washington, DC: Center for Women Policy Studies.
ROZNOWSKI, M. and REITH, J. (1999). 'Examining the measurement quality of tests containing differentially functioning items: do biased items result in poor measurement?' *Educational and Psychological Measurement*, **59**, 2, 248 - 69.
RUDISILL, E.M. and MORRISON, L.J. (1989). 'Sex differences in mathematics achievement: an emerging case for physiological factors', *School Science and Mathematics*, **89**, 7, 571 - 7.
SCHAFFNER, P.E. (1985). 'Competitive admission practices: when the SAT is optional', *Journal of Higher Education*, **56**, 1, 55 - 72.
SCHMIDT, F.L. and 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 - 74.
SCHMITT, A.P. and DORANS, N.J. (1990). 'Differential item functioning for minority examinees on the SAT', *Journal of Educational Measurement*, **27**, 1, 67 - 81.
SCHNEIDER, D. and DORANS, N. (1999). *Concordance Between SAT I and ACT Scores for Individual Students* (Research Notes 07). New York, NY: College Board.
SCHURR, K.T., HENRIKSEN, L.W. and RUBLE, V.E. (1990). 'The use of the College Board classification of high schools in predicting college freshman grades', *Educational and Psychological Measurement*, **50**, 1, 219 - 23.
SEASHORE, H.G. (1962). 'Women are more predictable than men', *Journal of Counseling Psychology*, **9**, 3, 261 - 70.
SHAYER, M. and ADEY, P.S. (1981). *Towards a Science of Science Teaching: Cognitive Development and Curriculum Demand.* London: Heinemann.
SHEEHAN, K.R. and GRAY, M.W. (1992). 'Sex bias in the SAT and the DTMS', *The Journal of General Psychology*, **119**, 1, 5 - 14.

SMITHERS, A. and ROBINSON, P. (1991). *Beyond Compulsory Schooling: a Numerical Picture*. London: Council for Industry and Higher Education.
SMYTH, F.L. (1995). 'How the ACT and the SAT are used and compared', *The Journal of College Admission*, **148**, 24 - 31.
STAGE, C. (1992). *Betyg och högskoleprov* [Grades and SweSAT scores] (Prov-memoria, Nr 53). Umeå universitet, Pedagogiska institutionen, Avdelningen för pedagogiska mätningar. Cited in: WOLMING, S. (1999). 'Validity issues in higher education selection: a Swedish example', *Studies in Educational Evaluation*, **25**, 4, 335 - 51.
STEIN, J. (2000). 'A true test of talent', *Times Higher Educ. Suppl.*, **1450**, 25 August, 14.
STERNBERG, R.J. (1999). 'The theory of successful intelligence', *Review of General Psychology*, **3**, 4, 292 - 316.
STOBART, G. (2000). 'The Scholastic Aptitude Test (SAT) as a model for an academic aptitude test in England.' Paper presented to the Advisory Group on Research into Assessment and Qualifications, QCA, London, 22 June.
STRICKER, L.J., ROCK, D.A. and BURTON, N.W. (1996). 'Using the SAT and high school record in academic guidance', *Educational and Psychological Measurement*, **56**, 4, 626 - 41.
THE SUTTON TRUST (2000). *Entry to Leading Universities. Executive Summary* [online]. Available: http://www.suttontrust.com/text/Report1.doc [29 September, 2000].
TATE, N. (2000). 'Extending the grade', *Times Higher Educ. Suppl.*, **1418**, 14 January, 18.
TEKIAN, A. (2000). 'Minority students, affirmative action, and the admission process: a literature review, 1987 - 1998', *Teaching and Learning in Medicine*, **12**, 1, 33 - 42.
VARS, F.E. and BOWEN, W.G. (1998). 'Scholastic aptitude test scores, race, and academic performance in selective colleges and universities.' In: JENCKS, C. and PHILLIPS, M. (Eds) *The Black-White Test Score Gap.* Washington, DC: Brookings Institution Press.
WAINER, H. (1999). 'Comparing the incomparable: an essay on the importance of big assumptions and scant evidence', *Educational Measurement: Issues and Practice*, **18**, 4, 10 - 16.
WAINER, H. and STEINBERG, L.S. (1992). 'Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: a bidirectional validity study', *Harvard Educational Review*, **62**, 3, 323 - 36.
WALLER, J.H. (1971). 'Achievement and social mobility: relationships among IQ score, education, and occupation in two generations', *Social Biology*, **18**, 252 - 9.
WEDMAN, I. (1994). 'The Swedish Scholastic Aptitude Test: development, use, and research', *Educational Measurement: Issues and Practice*, **13**, 2, 5 - 11.
WILDER, G., CASSERLY, P.L. and BURTON, N.W. (1988). *Young SAT-takers: Two Surveys* (College Board Report No. 88 - 1). New York, NY: College Board. Cited in: POWERS, D.E. and ROCK D.A. (1999). 'Effects of coaching on SAT I: Reasoning Test scores', *Journal of Educational Measurement*, **36**, 2, 93 - 118.
WILLIAMS, E.M. (1950). 'An investigation of the value of Higher School Certificate results in predicting performance in first-year university examinations', *British Journal of Educational Psychology*, **20**, 83 - 98.
WILMOUTH, D. (1991). *Should the SAT be a Factor in College Admissions?* (ED 345 592). Washington, DC: U.S. Department of Education, Educational Resources Information Centre.
WOLCHOVER, J. (2000). 'New tests aim to dig out hidden university talent', *Evening Standard*, 12 April, 11.
WOLFE, R.N. and JOHNSON, S.D. (1995). 'Personality as a predictor of college performance', *Educational and Psychological Measurement*, **55**, 2, 177 - 85.
WOLMING, S. (1999). 'Validity issues in higher education selection: a Swedish example', *Studies in Educational Evaluation*, **25**, 4, 335 - 51.

ZEIDNER, M. (1987). 'Age bias in the predictive validity of scholastic aptitude tests: some Israeli data', *Educational and Psychological Measurement*, **47**, 4, 1037 - 47.

Appendix 1: Methodology for the Review

Inclusion/Exclusion Criteria

Study Design

Literature from refereed journals and conference papers was included, and articles and statistics were also obtained from searches of websites.

Target Population

Literature pertains to students who were attending university or college, or who were about to make the transition from high school to university or college.

Time and Place

Searches in this review date from 1989 and cover all countries that use aptitude tests for university or college selection. Secondary references were also obtained where they were considered to make a substantial contribution to the discussion, and so may pre-date 1989.

Search Strategies

As the primary method of identifying published literature for this review, staff at the NFER Library searched a range of different sociological, educational and psychological databases: ASSIA, British Education Index, ERIC, PsycLIT and TES/THES, as well as the Library's own internal databases. Searches were limited to articles published in the English language. Due to limited resources, other recommended means of searching, such as handsearching of journals, were not undertaken. A record of the searches undertaken for the various databases has been documented and is outlined below.

Internet searches

The websites of Educational Testing Services, the College Board and the American College Testing Program were searched for statistics and articles on tests for the admissions process. Internet searches were also conducted to search for other countries which use aptitude testing for university entrance.

Databases

PSYCLIT (1989 - 2000)

#1 College Entrance Exam Board Scholastic Aptitude Test

#2 Entrance Examinations **NOT** College Entrance Exam Board Scholastic Aptitude Test

#3 College Academic Achievement

#4 Academic Achievement Prediction **NOT** College Academic Achievement

#5 Academic Aptitude

#6 Attitude Measures **NOT** Academic Aptitude

#7 Consequential Validity

#8 Scholastic Aptitude Test **AND** Differential Item functioning (free-text)

ERIC

#1 Scholastic Aptitude Test

#2 College Entrance Examinations Scholastic Aptitude Test

#3 Academic Aptitude **NOT** (Scholastic Aptitude Test **OR** College Entrance Examinations)

#4 Predictive Validity **NOT** (Scholastic Aptitude Test **OR** Academic Aptitude **OR** College Entrance Examinations)

#5 Scholastic Aptitude Test **AND** Differential Item functioning (free-text)

BEI

#1 Admission Criteria **AND** (Universities **OR** Higher Education)

#2 University Admission **NOT** Admission Criteria

#3 Scholastic Aptitude Test (free-text)

#4 Predictive Validity

#5 Scholastic Aptitude Test **AND** Differential Item functioning (free-text)