A Pilot of Aptitude Testing for University Entrance

Angus S. McDonald

Paul E. Newton

Chris Whetton

with an appendix by

Samantha E. Higgs

nfer

Contents

List of Tables

List of Figures

Acknowledgements

The project team for this work consisted of:

Design and secretarial support:

Liz Gibson and Jackie Hill

Administration of project materials:

Keren Beddow, Maria Charles, and John Hanson

Statistical analysis:

Samantha Higgs, Rachel Felgate, Ian Schagen, Dougal Hutchison and Sally Green

The NFER also gratefully acknowledges the assistance of Educational Testing Services for providing and scoring the SAT I: Reasoning Test.

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.

Executive summary

This report presents the findings from a study examining the association between the Scholastic Assessment Test (SAT) and A-level attainment. This research was commissioned by The Sutton Trust and conducted by the National Foundation for Educational Research (NFER).

Recent media debate on the British university admissions system has highlighted the American SAT as a potential way of identifying able students from less privileged backgrounds. In doing so, it was offered as one way of redressing the imbalance in students from state and independent schools in some of the highest-ranked universities. However, as no recent British studies have been conducted on the SAT, there was a lack of empirical evidence for many of the claims that were made for it. The present study was undertaken to partially redress this, by looking at the association between the SAT and A-level grades, and whether this association varied according to background factors.

Three samples of students participated in this study: high- and low-achieving schools, sampled on the basis of their GCSE results, and a sample of selective independent schools provided by The Sutton Trust. Participating students completed a questionnaire which collected personal details, their GCSE and predicted A-level grades, and their intended destination when they had finished their current course. Students also completed a short version of the SAT known as the SAT I: Reasoning Test, which contained verbal and math sections. Attained A-level results were collected when available. Completed materials were received from 1,295 students.

The effects of a range of background factors on GCSE and A-level grades and the SAT I : Reasoning Test scores were studied. These included sample (high-attaining, low- attaining and independent schools), sex, parental socio-economic status, intended destination and students' intended university or college if they planned to go on to higher education. The correlations between study variables were also examined.

The main tests of the study hypotheses were conducted with multilevel modelling, with mean attained A-level grade as the dependent variable. This revealed that the SAT I: Reasoning Test was modestly associated with A-level grades, but there was no evidence that the association differed according to background factors such as ethnicity, parental socio-economic status and sample. The SAT may be of value in predicting university performance but the data provided no evidence that it was able to assess potential for study at higher education, independently of a student's social and educational experiences.

The modest correlation between the SAT I: Reasoning Test and A-level grades meant that if the SAT score were used for selection, roughly the same proportion of students would be considered from each type of school as with A-Levels. However, these might not be the same individuals. Using the SAT scores in addition to A-levels would increase the number of students selected for all types of school, but the percentage increase would be greatest for students from low-attaining schools.

Further analyses explored the functioning of the SAT I: Reasoning Test in British students. These showed that the SAT I: Reasoning Test provided a coherent assessment of verbal and math reasoning ability, and that individual items appeared to function similarly in British and American students. Item-level analyses found little evidence of bias in SAT items between males and females, or the three samples of schools.

Although this study failed to find evidence that background factors differentially affected A-level grades and SAT I: Reasoning Test scores, SAT scores were only modestly associated with A-level grades. This indicates that the SAT I: Reasoning Test assesses a distinct construct from A-levels, and that further investigation of aptitude testing as a predictor of university performance is warranted.

Background

Introduction

Despite a number of recent changes to post-compulsory education in Britain, including the emphasis on key skills and vocational qualifications, A-levels remain the most frequently taken academic qualification at the end of sixth-form or further education. Although the function of A-levels is to assess attainment in curriculum-based subjects, they are also the main method by which students are selected for higher education. In this role they effectively act as aptitude tests, being taken as indicators of how students are likely to perform on their chosen degree course.

Recent debate on the process of admissions to British universities has focused on some of the limitations of the current system. This debate was partially fuelled by statistics published by The Sutton Trust (The Sutton Trust, 2000), and the case of an able student from a state school who was rejected from Oxford but accepted by Harvard in the United States (e.g. Stein, 2000). Using league tables of British universities published in a number of national newspapers, statistics from The Sutton Trust showed that students from independent schools were over-represented in the highest-ranked universities, as were students from the higher social classes. This apparent bias remained after A-level attainment had been accounted for.

The debate on university admissions included considerable discussion of alternative or supplementary methods of selection for higher education (e.g. Carvel, 2000; Charter, 2000; Lewis, 2000; Richardson, 2000). One possibility discussed in the media was the use of admissions tests. Probably the best known example of an admissions test is the Scholastic Assessment Test (SAT) which is used in the United States. In terms of the media debate this had popular appeal as it was described as being able to identify aptitude for university education, regardless of students' social and academic background (e.g. Clare, 1999). Additional advantages of the SAT were also cited, including it giving a scaled score on a range from 400 to 1,600, so allowing greater discrimination between students than A-level grades. Further, in the United States, students can take and receive their SAT results before they apply to colleges, so reducing their reliance on predicted grades to guide their choice of institutions. (However, it should be noted that this is not an inherent feature of the SAT as attained A-level grades could also be made available earlier, if the timing of these exams in relation to university applications was changed.)

The media debate subsequently turned somewhat against the SAT, suggesting that the results obtained from it were not independent of social background and other factors, as originally claimed (e.g. Richardson, 2000; Lewis, 2000). An investigation of the SAT has also been conducted by QCA, and although this report suggested a number of changes that could be

made to the university admissions system in Britain, the adoption of the SAT was not recommended (Stobart, 2000). However, throughout this debate there was a notable lack of rigorous evidence presented for the claims made first for, and later against, the SAT (but see McDonald *et al.*, 2001, for a detailed discussion of the SAT and aptitude testing for university entrance).

In discussing the possibility of aptitude testing, one of the most fundamental questions that needs to be addressed is the association between a test such as the SAT and attainment measures such as A-levels. It is known that A-levels have only limited ability to predict degree performance (e.g. Nisbet and Welsh, 1966; Choppin *et al.*, 1973; Peers and Johnston, 1994; Mellanby *et al.,* 2000), and so clearly there is scope for other predictors of performance in higher education. However, if the underlying constructs assessed by the SAT and A-levels are very similar, there would be little value in pursuing the SAT as an entrance test for higher education. However, if the constructs these two forms of assessment measure are quite distinct, then aptitude tests may provide valuable information about a student's potential for study at higher education.

An early study of the SAT in British students was conducted by Fremer *et al.* in 1968. Just over 1,000 fifth- and sixth-form students who were expected to attend university took the SAT, and O-levels and A-levels were used as outcome criteria. Little difference was seen in the SAT verbal scores of males and females, but males showed superior performance on the math section, a finding noted by the authors to be consistent with evidence from the United States.

Fremer *et al.* (1968) reported results separately for each of the schools in their sample. Associations between verbal SAT scores and O-levels ranged from 0.27 to 0.61, with the corresponding figures for the math section being between 0.21 and 0.59. Prediction was improved when total SAT score was considered, with values ranging from 0.40 to 0.69. Attained A-level grades were available for three schools in the study. Verbal SAT scores showed correlations of between 0.28 and 0.54 with A-level grades, with the corresponding figures for the math section being 0.12 to 0.39. Again, using total SAT score improved prediction, with values ranging from 0.20 to 0.57. Overall, this study indicated that the SAT was a moderate predictor of O- and A-level attainment, although this association was seen to differ considerably between schools. Most relevant to the current study, it also suggested that the SAT assessed somewhat different constructs from O- and A-levels.

A detailed study of aptitude testing for university entrance was conducted by Choppin and colleagues during the 1960s and 1970s (Choppin *et al.*, 1972; 1973; Choppin and Orr, 1976). This work involved the construction of a Test of Academic Aptitude (TAA), which was closely modelled on the SAT, and research on the ability of this and measures of academic

attainment to predict university performance. It was found that the TAA added very little to the prediction of degree results after O-levels, A-levels and teacher assessments of students' suitability for higher education had been taken into account. However, the association between the TAA and mean A-level grade was modest, being 0.51 for students studying science-related A-levels, and 0.48 for all others.

Choppin's work indicated that the TAA was assessing a construct which appeared to be quite different from A-levels; the two shared approximately 25 per cent of variance. Although the unique variance assessed by the TAA did not substantially improve prediction of degree grades, suggesting the aptitude test was of little value, considerable changes have occurred in the British education system since this work was conducted. For example, more students are going on to attain high A-level grades and there are now more diverse routes through which higher education can be accessed (e.g. vocational and access courses). Coupled with the increasing expansion of higher education, this suggests that university students are now a less-selected group than they were when Choppin's work was conducted. As any research of this nature needs to be interpreted in terms of the characteristics of the education system of the time, recent changes suggest that an aptitude test such as the SAT may again be worth investigating.

Claims have also been made that the SAT could assess potential for higher education, regardless of students' educational and social backgrounds (e.g. Clare, 1999). This has seemed a particularly appealing idea, given the debate on the under-representation of students from state schools and the lower social classes in some of the most prestigious universities (The Sutton Trust, 2000).

Empirical evidence does not unequivocally support the claims made for the SAT. Whilst it is able to predict university performance for American students, even when coupled with high school grades the majority of the variance in university performance remains unaccounted for (e.g. Bridgeman *et al.*, 2000). In terms of its relationship with background factors, there is continuing debate as to whether the SAT is a fair assessment of potential for males and females, and different ethnic groups (e.g. Bridgeman and Wendler, 1991; Vars and Bowen, 1998; Wainer and Steinberg, 1992). Probably the most consistent evidence comes from studies predicting the performance of males and females, which show that females' university grades are consistently under-predicted by the SAT (e.g. Wainer and Steinberg, 1992). The evidence for ethnic differences is less clear. Although African Americans have been consistently seen to score lower on the SAT than other groups (e.g. Lawlor *et al.*, 1997; Vars and Bowen, 1998), this is not peculiar to the SAT, as this group scores consistently lower on virtually all tests which measure aspects of intelligence (Neisser *et al*., 1996).

It is currently unclear whether the SAT is able to assess aptitude for university study, independently of social and educational factors. Evidence from the United States is equivocal, although it does suggest that the SAT may show similar patterns of results to many other high-level intelligence tests. Further evidence on this debate can also be found in the literature review conducted by McDonald *et al.* (2001). The current study was prompted by the recent debate on the SAT in Britain, and aimed to provide empirical evidence on the link between SAT scores, A-levels and social factors.

Goals of the present study

The main goal of the present study was to investigate the association between A-levels and scores on the American SAT I: Reasoning Test. A fundamental question concerns whether aptitude tests are able to measure abilities distinct from those currently assessed by A-levels, and so provide additional information about students' potential. If the SAT I: Reasoning Test assesses virtually the same constructs as A-levels, there would be little point in studying it further. However, if A-levels and the SAT I: Reasoning Test measure different constructs, this suggests that the predictive validity of aptitude tests is worth further investigation. It was hypothesised that the SAT I: Reasoning Test and A-levels would show a moderate degree of shared variance, in accordance with previous work, but not sufficient to suggest that they were assessing the same construct.

The second goal was to determine whether the association between the SAT I: Reasoning Test and A-levels varied according to background variables. There have been media claims that the SAT I: Reasoning Test is able to assess a student's potential for higher education study, regardless of their educational experiences, ethnic background and social circumstances (e.g. Clare, 1999). Although these claims have not found consistent empirical support, the current study was intended to provide evidence from British students to address this issue. Of particular interest was the link between the SAT I: Reasoning Test and A-levels in high- and low-achieving schools. No specific hypotheses were formulated for this part of the research, it being regarded as exploratory.

A further goal of the study was to investigate the general functioning of the SAT I: Reasoning Test on a sample of British students. This focused on its reliability, item- and test-level functioning and the extent to which it showed evidence of bias between different groups of students in an English context.

Methodology

Samples

Two samples of schools were drawn from NFER's Register of Schools in England. In accordance with one of the primary goals of this study - to determine the association between the SAT and A-levels in high- and low-achieving schools - schools from the top and bottom of the GCSE grade distribution were sampled.

The sample of high-achieving schools was taken from the top 40 per cent of schools on the basis of their achieved GCSE results, with the low-achieving sample being taken from the bottom 40 per cent. Included in the pool were all schools in England with sixth forms, with the exception of independent schools and further education colleges. Each sample was stratified by school type, and included comprehensive, grammar, secondary modern and other secondary schools. Further education institutions were not included as they were less likely to have students studying A-level courses, and due to the greater difficulty in accessing students in these institutions. This was a particular issue for the present study, as it was conducted over a relatively short period in order to ensure all testing was completed before the start of A-level exams.

Seventy-five high-achieving and 120 low-achieving schools were sampled and sent letters informing them of the study and requesting their participation. A number of schools declined to participate in the study at this stage, with the primary reasons for this being lack of time and concerns over scheduling the testing. In total, 29 high-achieving schools and 53 low-achieving schools agreed to participate and were sent test materials, with completed materials being returned by 25 and 40 schools respectively.

In addition to the samples taken from NFER's Register of Schools, The Sutton Trust also supplied contact details for eight independent schools who were interested in participating in this research (referred to as 'independent schools'). All of these schools operated a selective admissions policy. These independent schools were also contacted by the NFER and sent test materials, and six of these returned completed tests and questionnaires by the cut-off date.

Materials

Participating schools were sent packs of materials which included:

- Preparation materials giving students tips for taking the SAT I: Reasoning Test (see below), and practice questions.

- A student questionnaire which collected background information from students including ethnicity, parental socio-economic status, attained GCSEs, predicted A-levels, and what students intended to do when they completed their current course.

- SAT I: Reasoning Test, which assessed verbal and numerical reasoning ability. This test is developed in the United States by Educational Testing Services (ETS) for The College Board. The SAT version used for this research is a short form of the full SAT. It contained 35 verbal questions of three types: antonyms, sentence completions and reading comprehension, and 33 math questions also of three types: multiple-choice, student-produced responses and quantitative comparisons. The test is timed, with 30 minutes being allowed for the verbal section, and 40 minutes for the math. The SAT was provided for this study by ETS with the permission of The College Board.

- SAT I: Reasoning Test answer sheets.

All materials were administered by tutors at the schools which agreed to participate in the study. To facilitate administration and to ensure this was standardised between schools, detailed administration instructions were prepared by the research team. These covered administering the student questionnaire, the test-taking tips and practice questions, and then the verbal and math sections of the SAT I: Reasoning Test. It was recommended that all materials were administered in a single session if possible, as this would minimise the amount of missing data (e.g. through students completing the questionnaire but not turning up to the SAT I: Reasoning Test session). Tutors were instructed to administer the student questionnaire first, followed by the preparation session and then the SAT I: Reasoning Test.

Study materials were dispatched to schools on 28.3.00. Schools that had not returned materials by 4.5.00 were sent a reminder letter, and subsequently contacted by telephone if necessary. The final cut-off date for the return of materials was the end of the summer term. When they were announced, attained A-level results were collected from all of the schools that had returned completed test materials.

Results

Response rates

An examination of the data set revealed that a number of students had some missing data from their records. In order to maximise the number of students retained for the analyses, minimum data requirements were set in accordance with the primary goals of the study. Students were therefore included in the data set provided they had at least an attained grade for one A-level and one SAT I: Reasoning Test score (verbal or math).

A summary of the number of schools which responded in each sample is given in Table 1, along with the number of students in each sample who met the minimum data requirements.

Table 1: Number of schools returning completed materials, and number of students meeting minimum data requirements

	Low-achieving schools	High-achieving schools	Independent schools
Number of schools returning materials	40	25	6
Number of students with minimum necessary data	630	564	101

Treatment of data

Grades for all attained and predicted exams had been coded on a common scale, ranging from A* to U. From this information, summary scores were computed for exams achieved by the end of Year 11 (primarily GCSEs), predicted A/AS-level grades at the end of year 13, and attained A/AS-level grades.

For the Year 11 grades, it was decided only to include GCSEs and GCSE short courses. Very few students reported having taken exams other than GCSEs or GCSE short courses at the end of Year 11 (less than one per cent of all exams reported), and so due to issues of statistical equivalence between different exams, these were not included.

Two methods were considered for computing the GCSE scores. The first was to sum the results from all GCSEs taken, and the second was to take the mean grade from all GCSEs. The second of these options was chosen, as higher-performing schools may have given students opportunities to take GCSEs before Year 11, and may also have had the resources to allow students to take more GCSEs than low-achieving schools. It could be argued that stretching students to take more GCSEs could have the effect of lowering mean attainment.

However, scrutiny of the GCSE grades in the high-achieving schools did not support this view.

Due to using an averaging procedure, it was also necessary to decide whether ungraded GCSEs should be included. Students had not been explicitly asked in the questionnaire to record any ungraded exams. As these made up only 0.1 per cent of all grades, this suggested that very few students had chosen to report these. As including ungraded GCSEs would therefore penalise those few students who had chosen to report them, these were omitted from the calculation of mean GCSE results. In calculating mean GCSE scores, GCSE short courses were given a weighting of 0.5, and GCSEs a weighting of one.

Similar considerations were taken when computing summary scores for predicted and attained A- and AS-levels. It was again decided to calculate a mean A-level score for each candidate, but this time N and U grades were included in this calculation. This was done because schools had been asked to report all A/AS-level results, regardless of grades. N and U grades made up 9.6 per cent of reported A-levels. This can be compared to the national results from the previous year where these grades made up 8.7 per cent of all grades (GB. DfEE, 2000), suggesting that under-reporting of these had not occurred. Although far fewer students had been predicted to achieve N and U A-level grades (only 0.2 per cent), these were still included in the calculations of mean predicted grades. The most likely explanation for the discrepancy between predicted and attained grades is over-optimistic predictions on the part of sixth-form tutors. As with GCSEs, for the calculation of mean A-level grades, AS-levels were given a weighting of 0.5 and A-levels a weighting of one.

It should be noted that as GCSEs have an A* grade whereas A-levels do not, the mean scores for GCSEs are not directly comparable to those for predicted or attained A-levels in the tables given below. The following scale should be used when interpreting GCSE and A-level grades:

Grade	Value
A*	16
A	15
B	13
C	11
D	9
E	7

Finally, students had been asked to indicate the occupation of their male and female parent/carer, as a measure of socio-economic status. Socio-economic status was taken as the highest occupational category indicated by either parent.

Descriptive statistics

Tables 2 to 7 give breakdowns of the main study variables - mean GCSE grades, predicted and attained A-level grades, total SAT I: Reasoning Test score, and verbal and math scores - by background variables. The score distributions of these variables are given in Appendix 1. When interpreting these figures, it should be noted that some categories contain the scores from a very limited number of students, and so group differences should be interpreted with caution. This applies particularly to the tables describing scores by ethnicity, socio-economic status and intentions when finishing current course of study. The exam scores can be interpreted with reference to the figures given above.

Table 2 shows exam grades and SAT I: Reasoning Test scores by sample, with the standard deviations being given in parentheses. As can be seen, highest attainment on all exams was achieved by students in the independent schools provided by The Sutton Trust. Students in the high-achieving schools showed the next highest attainment, followed by those in the low-achieving schools. For all variables, the scores of the independent schools were significantly higher (at the five per cent level) than those of the other two samples. The scores from the high-achieving schools were also significantly higher than those of the low-achieving ones.

Table 2: GCSE, A-level and SAT I: Reasoning Test scores by sample

	Low-achieving schools (N=603)*	High-achieving schools (N=538)	Independent schools (N=100)	Total (N=1241)
Mean GCSE grade	12.2 (1.6)	13.4 (1.4)	15.1 (0.8)	12.9 (1.7)
Mean predicted A-level grade	10.8 (2.2)	12.2 (1.9)	14.4 (1.1)	11.7 (2.3)
Mean attained A-level grade	8.7 (3.5)	10.9 (3.00)	13.9 (1.7)	10.0 (3.5)
Total SAT score	938.3 (152.6)	1028.4 (155.5)	1232.1 (128.0)	1001.0 (172.1)
SAT verbal	481.0 (86.0)	521.1 (83.2)	602.4 (79.7)	507.9 (90.7)
SAT math	456.5 (94.9)	507.1 (99.9)	629.7 (82.0)	492.3 (107.1)

*N indicates the minimum number of cases in each column

Table 3 shows the scores according to sex. The only statistically significant differences were for the SAT I: Reasoning Test scores, where males tended to outscore females. This was particularly noticeable on the math section and total SAT I: Reasoning Test score, but males also scored marginally significantly higher on the verbal section. These sex differences are in accordance with data from The College Board on the SAT, where despite sex differences

decreasing over recent years, males continue to outscore females, particularly in the area of math (College Entrance Examination Board, 2000).

Table 3: GCSE, A-level and SAT I: Reasoning Test scores by sex

	Male (N=576)*	Female (N=665)	Total (N=1241)
Mean GCSE grade	12.9	13.0	12.9
	(1.7)	(1.6)	(1.7)
Mean predicted A-level grade	11.7	11.6	11.7
	(2.3)	(2.2)	(2.3)
Mean attained A-level grade	10.2	9.9	10.0
	(3.6)	(3.5)	(3.5)
Total SAT score	1038.0	970.3	1001.5
	(179.9)	(158.9)	(172.2)
SAT verbal	513.8	502.9	507.9
	(93.7)	(88.3)	(90.0)
SAT math	522.8	467.0	492.8
	(112.3)	(95.1)	(107.0)

*N indicates the minimum number of cases in each column

Attainment and SAT I: Reasoning Test scores by ethnicity are shown in Table 4. Students who described their ethnicity as being Black or Black British had significantly lower mean GCSE scores than other groups. Chinese students and those from other ethnic groups showed significantly higher GCSE attainment. Predicted A-levels were significantly lower for Whites than other students, and significantly higher for Chinese and students from other ethnic groups. Attained A-levels did not differ significantly between ethnic groups.

Total SAT I: Reasoning Test scores were significantly lower for Asian and Asian British students, and Black and Black British students than for other ethnic groups. Chinese students and those from other ethnic groups scored significantly higher on the SAT I: Reasoning Test than all others. On the verbal section, Asian and Asian British students scored significantly lower than all other ethnic groups, and those who described their ethnic background as being Mixed, scored higher than all others. On the math section, Chinese and those from other ethnic groups had significantly higher scores than all other students. However, in interpreting these results the small numbers of students in some groups needs to be taken into account, as this may make the findings unreliable.

Table 4: GCSE, A-level and SAT I: Reasoning Test scores by ethnicity

	Mean GCSE grade	Mean predicted A-level grade	Mean attained A-level grade	Total SAT score	SAT verbal	SAT math
White	12.9	11.6	10.0	1008.5	513.8	493.6
(N=1033)*	(1.6)	(2.3)	(3.5)	(166.6)	(88.0)	(105.1)
Mixed (N=15)	13.2	12.5	11.4	1010.0	535.0	476.7
	(1.8)	(1.8)	(2.7)	(167.0)	(94.8)	(97.8)
Asian or Asian	12.7	11.9	9.8	938.6	463.0	476.8
British (N=150)	(1.8)	(2.2)	(3.6)	(186.9)	(92.6)	(113.8)
Black or Black	12.1	12.3	10.4	969.0	483.6	479.0
British (N=10)	(2.3)	(2.3)	(4.2)	(272.4)	(137.6)	(139.5)
Chinese or other	13.8	13.1	11.9	1084.4	513.9	568.2
ethnic group	(1.6)	(1.7)	(3.3)	(174.5)	(94.9)	(107.9)
(N=26)						
Total (N=1237)	12.9	11.7	10.0	1001.4	507.7	492.9
	(1.7)	(2.3)	(3.5)	(172.1)	(90.7)	(107.0)

*N indicates the minimum number of cases in each row

Attainment and SAT I: Reasoning Test scores by parental socio-economic status are shown in Table 5. Although this table shows the means and standard deviations for all groups, statistical analyses were only conducted using the categories which contained at least ten students. Group differences may be unreliable when very small numbers are involved, and even group sizes of ten can be considered an absolute minimum for statistical comparisons.

Mean GCSE scores were significantly lower for students with parents in the Craft or trade worker group, than all other socio-economic groups. Scores for students with parents in the Corporate manager, senior official group were significantly higher than all others, with the exception of the Professional group, who had significantly higher scores still. Very similar patterns were seen for predicted and attained A-levels, with significantly lower scores being obtained by students with parents in the Craft or trade worker group, and significantly higher scores being obtained by those with parents in the Corporate manager, senior official and Professional groups.

SAT I: Reasoning Test differences according to socio-economic status largely mirrored exam results, for both the total score and verbal and math sections. The highest scores were obtained by students with parents in the Corporate manager, senior official and Professional groups, with these scores being significantly higher than all others. Students whose parental occupation was classified as being in the Craft or trade worker group scored significantly

lower than all other students, with the exception of the General labourer group, who had comparable verbal scores, and the Plant or machine operator group, who had comparable math scores.

Table 5: GCSE, A-level and SAT I: Reasoning Test scores by parental socio-economic status

	Mean GCSE grade	Mean predicted A-level grade	Mean attained A-level grade	Total SAT score	SAT verbal	SAT math
Corporate manager, senior official (N=285)*	13.2 (1.7)	11.9 (2.1)	10.5 (3.4)	1038.5 (176.2)	525.4 (92.7)	512.3 (106.6)
Professional (N=393)	13.3 (1.6)	12.1 (2.3)	10.8 (3.3)	1039.5 (170.0)	528.4 (88.9)	511.2 (107.4)
Technician or associate professional (N=173)	12.9 (1.5)	11.4 (2.2)	10.0 (3.3)	988.5 (153.7)	496.3 (81.5)	491.1 (100.3)
Small business owner (N=92)	12.7 (1.7)	11.6 (2.2)	9.5 (3.7)	966.4 (156.7)	490.7 (91.1)	473.5 (98.0)
Clerk or secretary (N=95)	12.6 (1.5)	11.3 (2.0)	9.1 (3.3)	956.1 (149.6)	489.7 (73.6)	465.7 (94.9)
Service or sales worker (N=86)	12.2 (1.5)	10.8 (2.6)	8.8 (3.9)	929.0 (158.6)	473.3 (87.7)	454.7 (108.2)
Skilled agricultural or fishery worker (N=1)	13.6 (0.8)	12.6 (1.3)	11.0 (2.0)	860.0	413.3 (40.4)	410.0
Craft or trade worker (N=39)	11.9 (1.2)	10.7 (2.2)	8.2 (3.7)	889.5 (167.5)	453.7 (82.9)	435.5 (115.8)
Plant or machine operator (N=29)	11.9 (1.3)	10.8 (2.4)	8.7 (3.1)	903.7 (171.4)	465.2 (95.4)	437.7 (99.0)
General labourer (N=15)	12.5 (1.5)	11.7 (1.9)	9.2 (3.8)	909.3 (145.7)	450.0 (96.1)	457.3 (68.9)
Never worked outside the home for pay (N=8)	11.7 (2.6)	11.6 (2.6)	8.7 (5.3)	931.3 (217.9)	463.3 (86.2)	471.3 (139.9)
Total (N=1221)	12.9 (1.7)	11.7 (2.3)	10.1 (3.5)	1001.3 (171.9)	507.6 (90.7)	492.8 (107.0)

*N indicates the minimum number of cases in each row

Students were also asked to indicate what they intended to do when they had finished their current course of study. Attainment and SAT I: Reasoning Test scores by these intentions are

shown in Table 6. As with the analyses for parental socio-economic status, groups with less than ten students were not included in the statistical comparisons.

Students who intended to study for a degree or take a year out had significantly higher GCSE and predicted A-level grades than all other groups. Students who indicated they intended to take a job with training had significantly lower attained A-levels than all other students. Predicted A-levels were significantly higher in students who intended to take a year out than all other groups. Those who intended to study for a degree had lower predicted A-levels than students who intended to take a year out, but higher scores than the remaining groups. Predicted A-levels were lowest in students who intended to take a job with training.

Table 6: GCSE, A-level and SAT I: Reasoning Test scores by intentions at the end of current course

	Mean GCSE grade	Mean predicted A-level grade	Mean attained A-level grade	Total SAT score	SAT verbal	SAT math
Job without training	11.2	9.4	6.4	780.0	410.0	370.0
(N=5)*	(1.2)	(1.9)	(4.1)	(84.9)	(66.0)	(54.0)
Job with training	11.6	9.6	6.8	903.2	454.2	448.2
(N=65)	(1.5)	(2.5)	(3.9)	(156.8)	(82.2)	(101.8)
Take a year out	13.1	11.9	10.6	1042.5	529.2	511.6
(N=165)	(1.7)	(2.4)	(3.5)	(181.7)	(93.0)	(113.8)
Study for a degree (at	13.1	11.9	10.3	1013.7	513.8	499.3
university or college)	(1.6)	(2.1)	(3.4)	(165.7)	(88.8)	(103.9)
(N=909)						
Study at a further	11.8	10.9	8.6	861.1	450.8	410.0
education college	(1.7)	(1.9)	(3.2)	(150.6)	(79.1)	(93.0)
(N=46)						
Don't know (N=37)	12.0	10.6	8.8	935.6	467.2	468.5
	(1.5)	(2.1)	(3.7)	(172.5)	(89.8)	(106.7)
Other (N=10)	11.6	10.6	8.0	892.7	467.3	425.5
	(2.0)	(1.6)	(3.2)	(110.0)	(78.4)	(73.5)
Total (N=1237)	12.9	11.7	10.0	1001.4	507.9	492.7
	(1.7)	(2.3)	(3.5)	(172.3)	(91.0)	(107.0)

*N indicates the minimum number of cases in each row

Total SAT I: Reasoning Test scores and scores on the verbal and math sections were significantly higher in students who intended to study for a degree or take a year out, than all other groups. The remaining groups did not differ significantly from each other, with the exception that students who intended to study at a further education college, had significantly lower math scores than the remaining groups.

The majority of students indicated that they intended to study for a degree at a university or college when finishing their current course. A further set of analyses explored the relationship between students' first choice of institution and exam and SAT I: Reasoning Test scores. Universities and colleges were classified into 13 categories, details of which are given in Appendix 2. Table 7 shows mean exam grades and SAT I: Reasoning Test scores by these 13 categories of institutions.

Significance tests were conducted on the main study variables between these categories, again excluding those which had less than ten students. Students who were intending to go to Oxford or Cambridge, attained higher GCSE and A-level grades and had higher predicted A-levels than all other students. Students whose preferred higher education institution was one of the 'New new' universities (mainly polytechnics redesignated as universities) or classified as 'Other' (mainly colleges of further and higher education) had significantly lower exam grades than all other students.

Very similar patterns were seen when the SAT I: Reasoning Test scores were examined, although on the total score and the verbal section, students planning to attend Cambridge scored significantly higher then those planning to attend Oxford. Also, students who were planning to attend 'Technological universities' had verbal scores that were not significantly different from students attending 'New new' or 'Other' universities, with these being significantly lower then all other students.

Table 7: GCSE, A-level and SAT I: Reasoning Test scores by classification of universities and colleges

	Mean GCSE grade	Mean predicted A-level	Mean attained A-level	Total SAT score	SAT verbal	SAT math
Oxford (N=17)	15.3	14.8	14.1	1192.9	587.7	605.3
	(0.5)	(0.2)	(1.2)	(111.2)	(62.0)	(76.1)
Cambridge (N=37)	15.4	14.8	14.4	1287.3	640.8	646.5
	(0.7)	(0.7)	(1.1)	(135.2)	(75.6)	(91.5)
Civic universities and London (N=233)	13.8	12.9	11.7	1075.8	537.0	537.6
	(1.3)	(1.6)	(2.9)	(164.4)	(85.9)	(110.0)
Redbrick universities (N=118)	13.8	12.7	11.6	10695.9	541.9	524.9
	(1.2)	(1.7)	(2.5)	(132.8)	(80.5)	(90.4)
Durham, Keele (N=27)	14.0	13.2	12.5	1083.7	542.2	541.5
	(1.5)	(1.7)	(2.9)	(29.1)	(78.0)	(111.8)
Technological universities (N=71)	13.0	11.8	10.5	995.1	493.2	501.1
	(1.2)	(1.4)	(2.3)	(142.3)	(78.2)	(88.2)
Scottish universities (N=12)	13.6	12.5	11.6	1096.7	550.8	545.8
	(2.0)	(2.4)	(2.9)	(171.4)	(90.7)	(98.6)
Welsh universities (N=31)	13.5	12.5	11.5	1058.4	541.6	516.8
	(1.3)	(1.6)	(2.4)	(128.1)	(76.5)	(76.5)
Northern Irish universities (N=1)	15.5	11.6	14.6	1130.0	610.0	520.0
'Old new' universities (N=80)	13.8	13.1	12.0	1057.7	538.8	516.9
	(1.2)	(1.4)	(2.6)	(138.7)	(89.7)	(86.3)
'New new' universities (N=278)	12.2	10.4	8.3	939.5	481.3	458.1
	(1.3)	(2.0)	(3.2)	(141.3)	(79.8)	(92.1)
Other (N=91)	11.9	10.2	8.1	896.8	466.6	428.7
	(1.2)	(1.9)	(3.2)	(131.6)	(79.9)	(74.1)
Total (N=996)	13.2	12.0	10.5	1023.8	517.9	504.9
	(1.6)	(2.2)	(3.4)	(168.7)	(90.3)	(105.9)

*N indicates the minimum number of cases in each row

The correlations between the main study variables are shown in Table 8. It can be seen that predicted A-level grades were the best predictors of attained A-levels, closely followed by GCSE results. These values were quite consistent, with the exception that mean GCSE grades predicted attained A-levels somewhat less well in the sample of independent schools. One reason for this may be restriction of score range, as students in this sample had very high GCSE results with a low spread of scores (see Table 2). Total SAT I: Reasoning Test scores showed a modest association with attained A-levels, with values in the high- and low-achieving schools being similar to those previously seen by Choppin *et al.* (1972). When the verbal and math sections were analysed, the verbal scores were more closely associated with attained A-levels than math scores, although neither exceeded the correlations seen for total scores. Correlations were somewhat lower in the independent schools, and again this may be due to range restriction of the mean attained A-level grades.

Associations with mean GCSE grades showed a similar pattern to that seen for A-levels. These were most closely associated with total SAT I: Reasoning Test score, followed by verbal scores and then math. When comparisons within samples were made, mean GCSE grades were more closely associated with SAT I: Reasoning Test scores than A-levels, particularly in the high- and low-achieving samples. This is somewhat surprising, considering students took the SAT I: Reasoning Test far closer to the time they took their A-levels compared to their GCSEs, but corresponds to previous findings from Fremer *et al.* in 1968. One possible reason for this is that mean GCSE grades reflect a wider overall course of study than A-levels, as all students will have taken maths, English and at least one science subject at GCSE.

A further notable finding from Table 8 is the association of exam and SAT I: Reasoning Test scores with socio-economic status. These associations indicate that attainment increased as did socio-economic status. This is in accordance with many prior studies of younger children, showing that factors such as eligibility for free school meals, a surrogate for social deprivation, are associated with lower attainment. What is particularly interesting in Table 8 is that parental socio-economic status has the greatest impact on students in the independent schools. The reason why parental socio-economic status has greater impact in these very high-attaining schools is unclear, although it may be related to adjustment issues in students from relatively lower-status backgrounds.

Table 8: Correlations of GCSE, predicted and attained A-levels, SAT I: Reasoning Test scores and socio-economic status by sample

	Mean GCSE grade	Mean predicted A-level grade	Mean attained A-level grade	Total SAT score	SAT verbal	SAT math
Mean predicted A-level grade	0.59 **0.68** *0.59*					
Mean attained A-level grade	0.64 **0.65** *0.52*	0.70 **0.71** *0.70*				
Total SAT I: Reasoning Test score	0.62 **0.58** *0.38*	0.40 **0.48** *0.38*	0.45 **0.50** *0.33*			
SAT I: Reasoning Test verbal	0.54 **0.52** *0.32*	0.36 **0.46** *0.26*	0.42 **0.46** *0.24*	0.82 **0.82** *0.78*		
SAT I: Reasoning Test math	0.51 **0.47** *0.28*	0.32 **0.36** *0.35*	0.34 **0.38** *0.28*	0.86 **0.88** *0.80*	0.41 **0.45** *0.25*	
Parental socio-economic status	-0.14 **-0.13** *-0.25*	-0.05 **-0.04** *-0.27*	-0.11 **-0.06** *-0.22*	-0.21 **-0.12** *-0.35*	-0.18 **-0.16** *-0.34*	-0.16 **-0.08** *-0.21*

plain text = low-achieving schools, bold = high-achieving schools, italic = independent schools

Multilevel modelling – overview

The main statistical analysis of the data was conducted with a technique known as multilevel modelling. Multilevel modelling allows the values on a variable of interest to be predicted (the dependent variable, in this case mean attained A-level grade), given the values on one or more variables (independent variables, in this case SAT I: Reasoning Test scores, mean GCSE and predicted A-level grades, and background variables such as sample, socio-economic status, sex and ethnicity).

Multilevel modelling is based on the statistical technique of regression, but extends this by looking at data that is grouped into similar clusters at different levels. For example, individual students are grouped into year groups or cohorts, and those cohorts are grouped within schools. There may be more in common between students within the same cohort than with other cohorts, and there may be elements of similarity between different cohorts in the same school. Multilevel modelling allows us to take account of this hierarchical structure of the data and produce more accurate predictions, as well as estimates of the differences between students, between cohorts, and between schools. The model fitted to the data incorporated two levels: school and student. Sample (low-achieving, high-achieving and the independent schools provided by The Sutton Trust) was not included as a third level, as this was treated as an explanatory (independent) variable. Further details of the multilevel modelling, along with a technical report of the findings, are given in Appendix 3.

Multilevel modelling - findings

Multilevel models developed to explore two areas. The purpose of the first of these was to examine which variables were the best predictors of mean A-level grades. Unsurprisingly, predicted A-levels were identified as the best predictor, closely followed by attained GCSE results. Total SAT I: Reasoning Test score was somewhat less closely associated with A-levels, although it was still highly significant. When verbal and math scores were analysed separately, both were seen to be significant, although verbal scores showed a slightly higher association than math scores.

The second model provided a more detailed examination of school- and pupil-level variables as predictors of attained A-levels, and also studied the interaction between these variables. Mean GCSE grades and predicted A-level grades were omitted from this model, due to statistical problems caused by these variables all being highly inter-correlated, hence the need for the initial model described above.

The main findings to emerge from this multilevel model were:

- SAT I: Reasoning Test scores were identified as having the strongest association with attained A-levels (but note, GCSEs and predicted A-levels were not included here).
- The next strongest predictor of A-levels was sample, indicating that attained A-levels were lowest in the low-achieving sample of schools, and highest in the sample of independent schools (see also Table 2).
- There was a positive effect of sex, showing that females attained higher A-levels than males, after other factors, including SAT I: Reasoning Test scores, had been allowed for.

- The analysis for parental socio-economic status compared each group with the Professional category. Significant negative effects were seen for the groups Corporate manager, senior official and Clerk or secretary, indicating that students whose parental socio-economic status was in one of these categories had lower- attained A-levels.

- There was a significant interaction effect between predicted A-level grades and sample. This appeared to suggest that students from high-achieving schools who were predicted to gain lower A-level grades showed higher attainment than expected.

- No evidence was found that the association between the SAT I: Reasoning Test and attained A-levels varied according to background factors such as ethnicity and sample.

An additional model was developed, which took an alternative approach to analysing socio-economic status. Initially this had been broken down into separate categories, but in this model it was treated as a continuous variable (excluding the category Never worked outside the home for pay). Interaction terms were also created with socio-economic status for this analysis. The results of this model were very similar to the previous one, and no significant interaction effects were seen. One possible reason for the lack of significant effects of background variables may be that a large proportion of the variance in A-levels has already been explained by SAT I: Reasoning Test scores.

An alternative way of examining the link between SAT I: Reasoning Test scores and exam grades is to hold exam grades constant and then look at the scores obtained by each sample. This effectively answers the question 'Did the SAT scores of students who attained, for example, an A grade at A-level, differ between the samples?'. For these analyses, all students were placed into three grade bands on the basis of their mean A-level scores, and then, separately, on the basis of their GCSE scores. Within these three bands, the SAT I: Reasoning Test scores of each sample were then calculated.

In order to split the distributions of GCSE and A-level grades into three approximately equal groups, different bands had to be applied to each. For GCSE results, students were banded into A*/A, B, and C grades and below. For A-levels, the bands were A/B, C, and D grades and below. Due to there being low numbers of independent school students in all but the top GCSE and A-level grade bands, the findings for this sample in the lower grade bands should be treated with caution. The results of these analyses using GCSE bands can be seen in Table 9, with the ones for A-levels in Table 10.

It can be seen from Table 9 that within each of the score bands, SAT I: Reasoning Test scores tended to be lowest in the low-attaining schools and highest in the independent schools. SAT I: Reasoning Test scores therefore varied between samples, after GCSE and A-level grades

had been equalised. Although this may suggest that the association between SAT I: Reasoning Test scores and exam grades varies according to school type, this is not actually the case. Exam grades were not precisely controlled in these analyses, since they were grouped into quite broad bands. Within these broad bands, GCSE and A-level scores were not evenly distributed between the schools types, and the significant findings appear to be due to similar variations in the SAT I: Reasoning Test within each band.

Table 9: SAT I: Reasoning Test scores by GCSE grade bands and sample

	GCSE grade band	Low-attaining schools	High-attaining schools	Independent schools
SAT I: Reasoning Test total	A*/A	1083.0	1130.1	1236.3
		(122.3)	(123.3)	(124.5)
	B	980.9	990.2	1177.1
		(130.0)	(138.5)	(177.5)
	C and below	852.0	905.5	-
		(125.7)	(125.2)	
SAT I: Reasoning Test verbal	A*/A	551.7	569.4	608.0
		(76.5)	(74.8)	(77.0)
	B	502.5	501.8	547.1
		(75.0)	(74.3)	(90.5)
	C and below	439.3	461.5	-
		(75.4)	(65.0)	
SAT I: Reasoning Test math	A*/A	531.3	560.9	628.4
		(86.2)	(85.9)	(81.1)
	B	476.8	487.1	630.0
		(86.7)	(92.9)	(94.0)
	C and below	412.9	443.5	-
		(82.6)	(88.5)	

One way of understanding this apparent anomaly between the results of the multilevel model and SAT I: Reasoning Test scores when students' exam grades are banded is to look at the regressions and scatterplots of the data given in Appendix 3 and Figure 1 below. It is clear from Figure 1 that the majority of students from the independent schools are clustered towards the top right of the scatterplot, and the majority of those from the low-achieving schools towards the bottom left.

Table 10: SAT I: Reasoning Test scores by A-level grade bands and sample

	A-level grade band	Low-attaining schools	High-attaining schools	Independent schools
SAT I: Reasoning Test total	A/B	1051.9	1106.6	1236.4
		(138.4)	(147.6)	(127.6)
	C	976.4	1010.1	1226.3
		(141.9)	(124.3)	(112.0)
	D and below	886.6	933.1	1116.7
		(135.0)	(130.8)	(169.2)
SAT I: Reasoning Test verbal	A/B	540.6	559.6	604.9
		(82.7)	(78.8)	(80.3)
	C	499.2	520.7	608.8
		(77.2)	(69.0)	(51.7)
	D and below	454.7	467.2	510.0
		(78.3)	(69.3)	(90.0)
SAT I: Reasoning Test math	A/B	510.5	546.7	553.8
		(96.0)	(100.3)	(104.6)
	C	475.9	490.9	617.5
		(91.9)	(86.0)	(62.1)
	D and below	431.7	464.9	606.7
		(86.0)	(88.4)	(83.9)

The regression line for predicting A-level from SAT I: Reasoning Test (heavier line) is based on the assumption that SAT I: Reasoning Test scores are known with absolute precision while A-level grades are subject to measurement error. Because of this, its slope is lower than would be the case if we passed a single line through the centre of the cloud of points, and this is what causes the effect that more students in the sample of independent schools are above the line and more students from low-achieving schools below it.

An exactly similar argument applies to the regression of SAT I: Reasoning Test score on A-level, with the same result. In practice, both measures are subject to uncertainty and measurement error, and it is clear from the scatterplot that any apparent effect of sample is probably due to the nature of the data rather than any substantive educational effect.

Separate scatter plots for each sample of schools are shown in Figures 2 to 4. The scatter plots of A-level vs SAT I: Reasoning Test scores for low-attaining schools, high-attaining

schools and independent schools show very different patterns. For independent schools (Figure 2), the data points of the scatter plot are very concentrated in one region with 93 per cent of the sample having a SAT I: Reasoning Test score of over 1000 and A-level performance of better than 10 points (average A-level score of between C and D). For low-attaining schools (Figure 3) the data of the scatter plot is much more diverse and there are large variations in both SAT I: Reasoning Test scores and A-level performance with only 21 per cent of the sample having an A-level points score better than 10 and a SAT I: Reasoning Test score of over 1000. High-attaining state schools (Figure 4) were somewhere in between with 44 per cent of the sample having A-level performance better than 10 and a SAT I: Reasoning Test score of over 1000.

In terms of individuals the lack of a strong relationship between A levels and the SAT I: Reasoning Test means that some students score highly on one measure and not so highly on the other. This can be illustrated by taking two high cut-offs for the measures. For A levels, a points score of 14 or more was used, which is equivalent to better than an average A-level grade between A and B, the level required to gain admission to one of the top ranked universities in the UK. For the SAT I: Reasoning Test, a score of 1200 corresponds to the cut-off used in the USA at Ivy League Universities for consideration of students from non-privileged backgrounds.

Table 11 shows the percentage of students in each sample who are above these two thresholds. For each sample, the percentage meeting each threshold is about the same. However, the percentage of students above the thresholds is markedly different for the three samples ranging from approximately five per cent in low-attaining schools to around 60 per cent meeting each threshold in independent schools.

In each case, the adoption of a criterion of students being above either threshold would increase the percentage being considered compared to the proportion considered based solely on A-level performance. This would result in a much larger percentage increase in the number considered in low-attaining schools. For low-attaining schools including pupils with SAT I: Reasoning Test scores above 1200 would lead to an extra 25 pupils from the sample being considered, increasing the percentage of pupils considered from the schools from four per cent to eight per cent (a 100 per cent increase). Adopting the same policy for independent schools would also increase the percentage of students being considered to a much smaller extent (21 per cent) and for high-attaining schools an extra 53 per cent would be considered.

Table 11: Percentage of students above selection thresholds

	A-level Score 14 or above	SAT I: Reasoning Test Score Above 1200	Either or Both Thresholds Achieved	% increase in numbers considered
Low-attaining Schools	4%	5%	8%	96%
High-attaining Schools	15%	13%	23%	53%
Independent Schools	67%	63%	81%	21%

Adopting an even higher level of A-level cut off for the results gives more dramatic results. With over 80% of students who enter Oxford and Cambridge achieving three A grades at A-level the effective entry requirement for Oxbridge is 15 points. On this basis only one per cent of the low-attaining state school students would be considered, compared with four per cent of the high-attaining state school students and 30 per cent of the independent school students.

For the SAT I: Reasoning Test scores on the other hand there were 30 students (5% of total) in low-attaining state schools who scored above 1200 on the SAT I: Reasoning Test, but only one of these students scored 15 points at A level.

The utility of a procedure where SAT I: Reasoning Test scores are considered in addition to A levels remains unknown at present, and cannot be known from the present research, since the relationship of the SAT I: Reasoning Test scores to degree outcomes is not known.

Figure 1: Regression of A-level on SAT I: Reasoning Test and SAT I: Reasoning Test on A-level

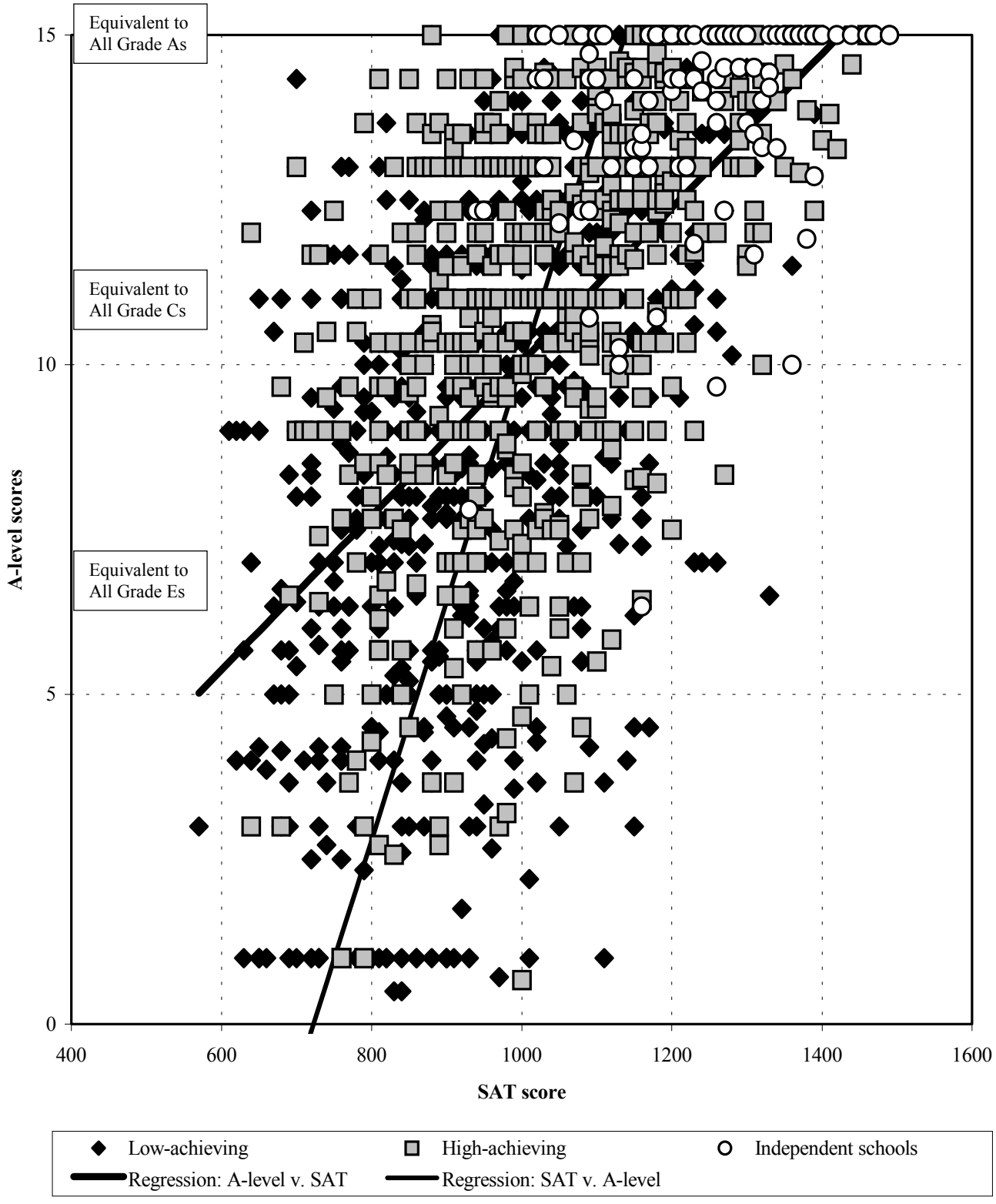


Figure 2: Regression of A-level on SAT I: Reasoning Test for independent schools

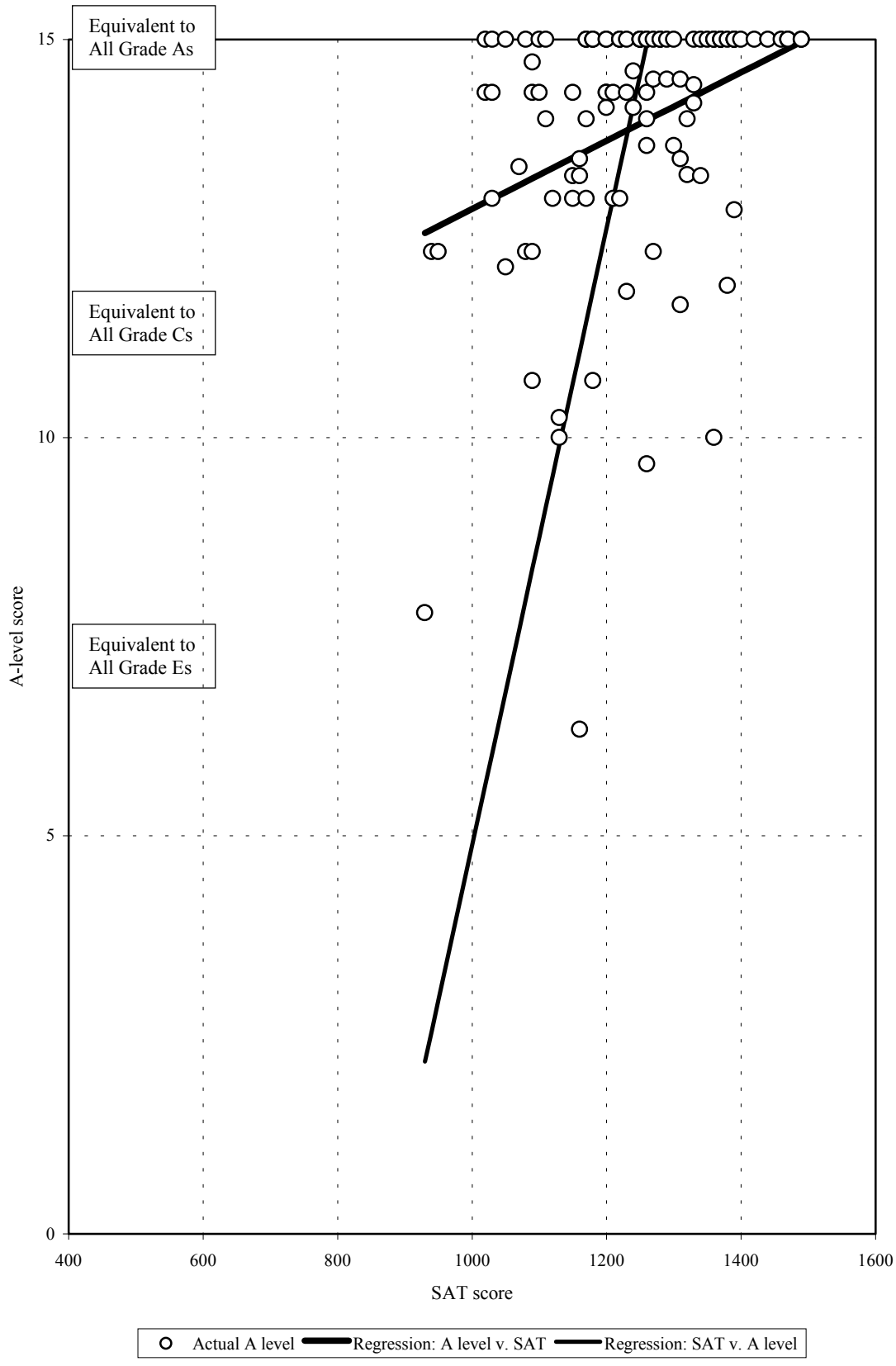


Figure 3: Regression of A-level on SAT I: Reasoning Test for low-achieving schools

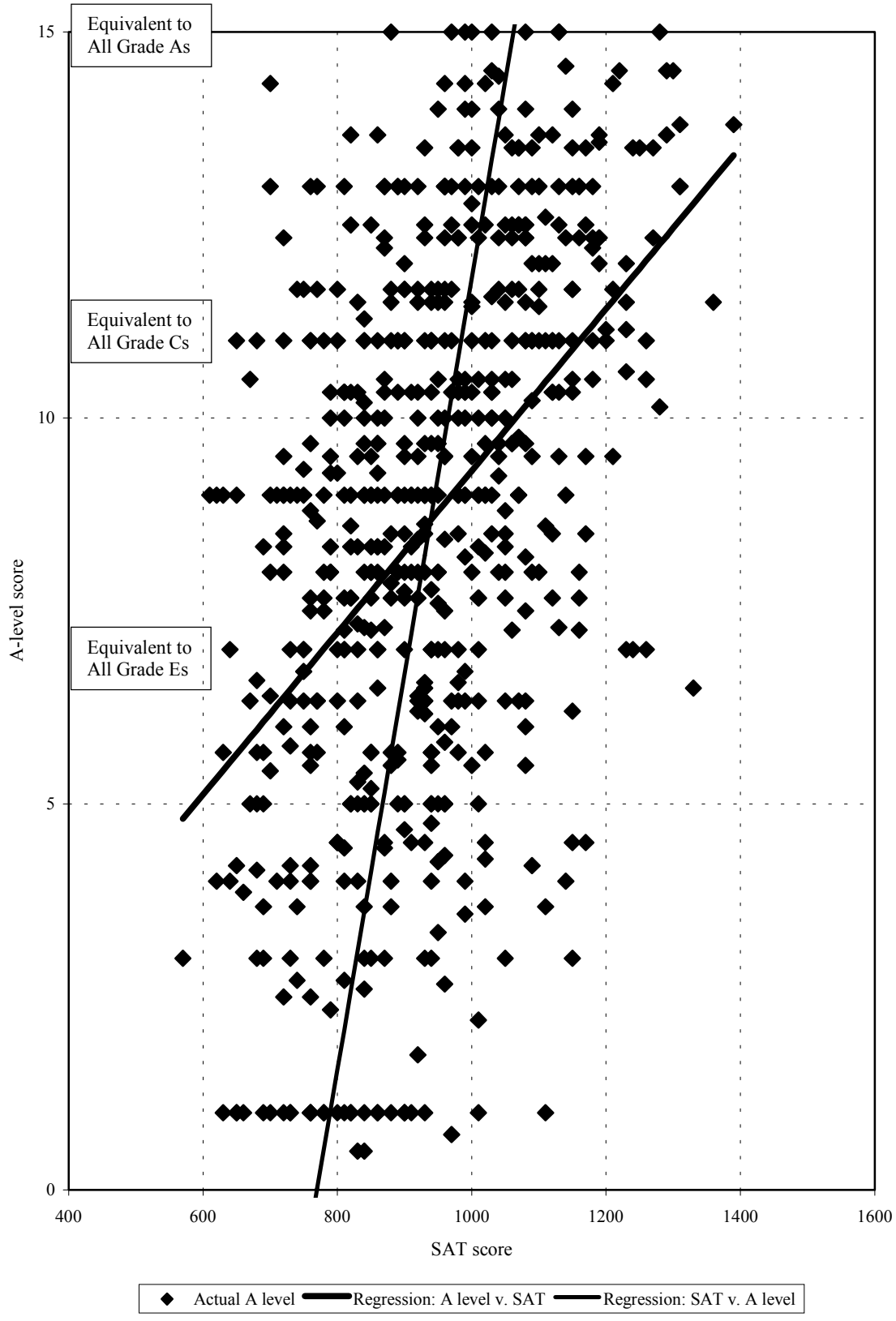
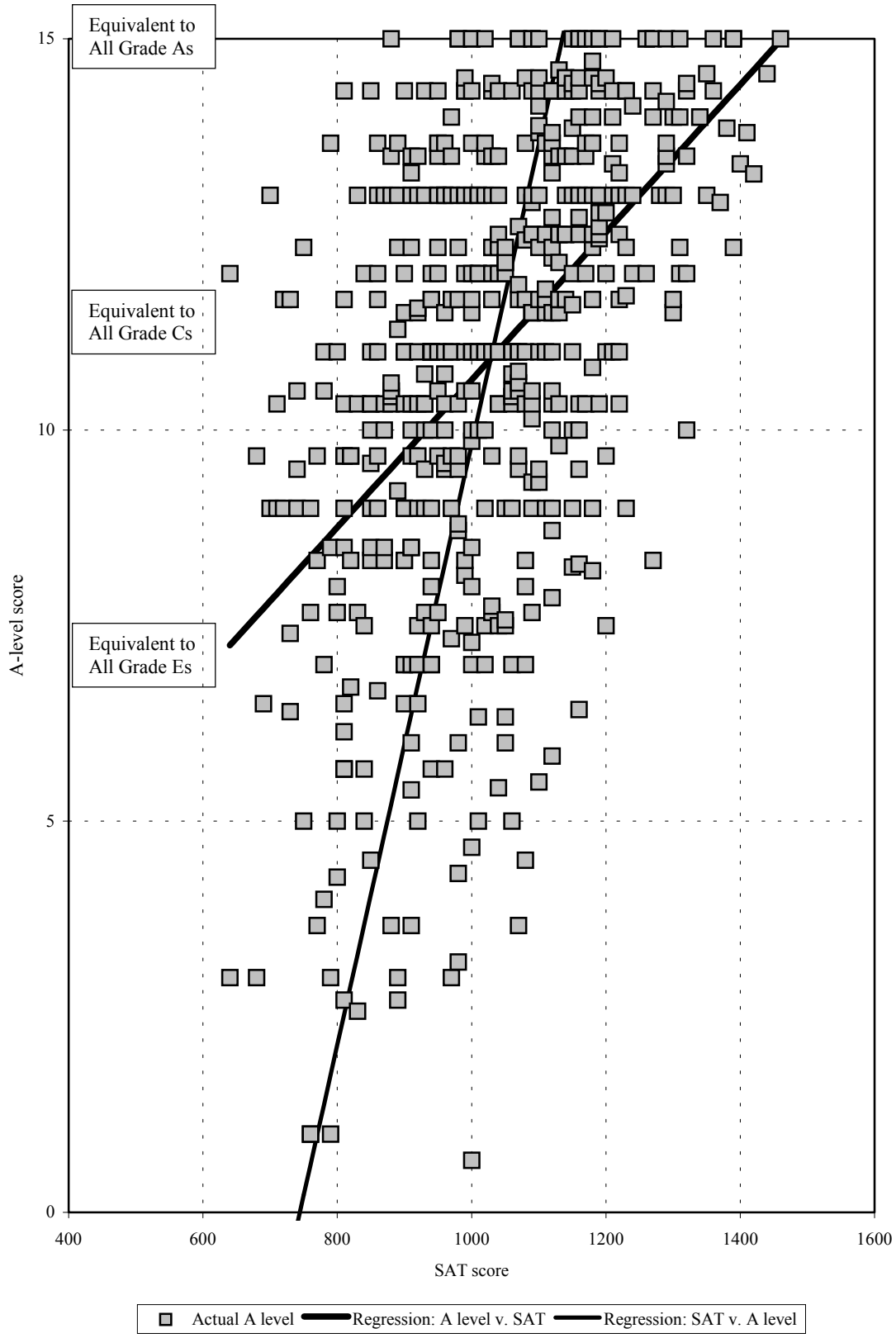


Figure 4: Regression of A-level on SAT I: Reasoning Test for high-achieving schools



Analysis of SAT functioning

The final set of analyses examined how the SAT I: Reasoning Test functioned in a sample of British students. These looked at the mean scores, reliability of the SAT I: Reasoning Test, the difficulty and discrimination of each item, the proportion of items omitted and not reached, and evidence of item-level bias.

Mean scores

Across all students, the mean verbal and math SAT scores for males were 514 and 523 respectively, with the corresponding scores for females being 503 and 467. The most recent data published by The College Board (College Entrance Examination Board, 2000) gives mean verbal and math scores of 507 and 533 for males, and 504 and 498 for females. These figures show that British students performed comparably to their counterparts in America, despite British students being less familiar with the mainly multiple-choice question format of the SAT, and having received minimal preparation before taking it. Although some differences in the scores were seen, most notably the lower scores of British females on the math section of the SAT I: Reasoning Test, this may be due to the unrepresentative nature of the samples used in this study.

Reliability

With any test such as the SAT, which is used to make decisions about individuals, an important statistic is its reliability. Reliability indicates the extent to which a test is consistent in its measurements. Consistency can be looked at in a number of ways, such as the degree to which test items are measuring a single construct or the consistency of results over time. With a test like the SAT the first of these is particularly important, as if test items are not all measuring the same underlying construct (in this case verbal or numerical reasoning), the degree of error in the test results will be large. The larger the error, the larger the score difference between two students has to be before it can be stated with a good degree of certainty that they reflect a real difference in the students' abilities - a factor particularly important if used for selection into higher education.

Table 12 shows the reliability of the SAT I: Reasoning Test across all students and for each of the samples. These figures were calculated using the Cronbach's alpha reliability formula, which gives values between zero and one. The closer the observed reliability is to one, the better the test items 'hang' together and are measuring a single underlying construct. As can be seen from Table 12, all scores obtained from the SAT I: Reasoning Test showed acceptable reliabilities considering the number of tests items, although the reliability of the math section

was slightly higher than that of the verbal section. Little variation in reliabilities was seen between the different samples, suggesting that the test-level functioning of the SAT I: Reasoning Test was not significantly affected by school type. Overall, this indicates that the SAT is a coherent measure of reasoning abilities in British students.

Table 12: Cronbach's alpha reliabilities for the SAT I: Reasoning Test

	Low-achieving schools (N=629)*	High-achieving schools (N=563)	Independent schools (N=101)	Total (N=1293)
Total SAT score	0.88	0.88	0.84	0.90
SAT verbal	0.80	0.79	0.81	0.82
SAT math	0.86	0.88	0.82	0.89

*N indicates the minimum number of cases in each column

Item analyses

ETS, who develop the SAT for The College Board, use a method known as item response theory (IRT) to analyse the functioning of SAT items. This aims to link an individual's predicted performance on an item to their ability and the characteristics of the item. The data ETS supplied on the SAT functioning gave three statistics, or parameters, for each item - slope, threshold and asymptote - which together describe the item characteristic curve or functioning of the item. This data allowed a comparison of the item functioning between British and American students.

The slope provides an indication of an item's discrimination, that is, its ability to distinguish between students who are likely to score highly on the overall test, and those likely to obtain lower scores. The steeper the slope of the line, the more discriminating an item is seen to be. The second statistic is the threshold, which indicates the difficulty of an item. The final statistic reflects the probability of obtaining a correct answer by chance. Although designed to allow for multiple-choice questions where there is a chance of guessing the answer correctly, this value can also reflect the relative ease of an item if it is answered correctly by a large proportion of test takers.

The three item statistics were calculated on the SAT I: Reasoning Test data from British students, and then compared with the figures supplied by ETS. Table 12 shows the correlations between these statistics in the British and American samples, and full details of the IRT analyses can be found in Appendix 4.

Considering the slope first, the correlations indicate that items which showed a high level of discrimination on American students were similarly able to distinguish between more and less

able British students. This association was somewhat stronger for the verbal items than the math. The correlations for threshold indicate that the difficulty of each item was highly comparable between British and American students. The final statistic, the asymptote, showed the least concordance between British and American students. This appeared to be due to American students having a very low probability of getting some questions wrong, whilst this was not the case for British students. This may have been due to the areas covered by some questions being very familiar to the majority of American students, but not so to British students. The less frequent use of multiple-choice questions in the British education system, and the very limited preparation for taking the SAT I: Reasoning Test, may also have contributed to these differences between British and American students.

Overall, these figures indicate a reasonable degree of concordance between the functioning of the SAT I: Reasoning Test items in the British and American students, with the exception of the asymptote. With the current data, it was not possible to determine why some differences were seen between the British and American students. Discrepancies may be due to relatively low numbers of British students making the IRT parameter estimates less reliable, and the non-representative nature of the British sample.

Table 13: Correlations between IRT parameters in British and American students

	IRT slope	IRT threshold	IRT asymptote
SAT verbal	0.70	0.80	0.41
SAT math	0.59	0.87	0.55

Analyses of SAT I: Reasoning Test functioning were also conducted according to the methods of classical test theory (CTT). Whereas IRT provides estimates of item functioning that are independent of the other test items, in CTT item functioning is related to the test as a whole. The full item analyses for the verbal and math sections are presented in Appendix 5. The math items showed very good discrimination, and although verbal items generally displayed acceptable levels of discrimination, three items had rather low values (below 0.20).

These analyses also provided information on the number of items students had omitted and number not reached. Omitted items were ones to which no response was made, and the proportion not reached indicates the number of students who had not made a response to an item or any subsequent items in the test section. For the verbal section, no more the 14 per cent of students omitted any single item, and just under 11 per cent did not reach the end of the test. This suggests that the verbal test was not particularly speeded, and most students were able to attempt the majority of questions.

For the math section, 57 per cent omitted or did not reach the last question. Coupled with the facility statistics, which indicate the proportion of students who answered a question correctly, this suggests that questions towards the end of the math test were found quite difficult relative to similarly placed verbal questions. This may be partially due to the math section being more speeded than the verbal, but omission rates were particularly high for the last five items. All of these items required students to generate answers and record them in grids on the answer sheet, as opposed to being multiple-choice questions. It is therefore possible that the unfamiliar format of these items had some impact on students' performance.

Bias

A final set of analyses concerned possible bias in the SAT I: Reasoning Test. Analyses of test bias can focus on overall test scores or individual items. Evidence of group differences in SAT scores have been presented above in Tables 2 to 7, although in the absence of further information, it is not possible to say whether these scores result from test bias or reflect real differences between the groups in question. An alternative way of assessing bias is to look at item-level performance for evidence of differential item functioning (*dif*). *Dif* analyses involve comparing two groups' chances of getting a test item correct, once their overall test scores have been matched. In doing this, it indicates items which are disproportionately easy or hard for a certain group.

Since 1989, SAT items have been routinely screened for *dif* before being included in live versions, with those items that exhibit extreme bias typically being removed from item pools used for SAT construction (Burton and Burton, 1993). Due to this process, it was expected that minimum levels of *dif* would be observed in the data from British students. Three sets of *dif* analyses were conducted on the data, comparing performance according to sex, ethnic status and sample. Full results of the *dif* analyses are given in Appendix 6, with the main points from these being summarised below.

Dif analyses were first conducted to compare males and females on the verbal and math sections of the SAT I: Reasoning Test. On the verbal section, only one item was identified as displaying a large degree of bias, with this item favouring females. No math items were identified as showing a large degree of bias.

Ideally *dif* analyses for ethnicity use tightly defined groups, as performance characteristics can vary considerably between specific ethnic groups. However, difficulties in obtaining sufficient numbers of test takers in all ethnic groups often result in analyses comparing 'Whites' with all 'Non-whites'. In order to overcome this limitation, the bias analyses conducted on the data compared Whites ('British', 'Irish' and 'Any other White background') with Asians ('Indian', 'Pakistani', 'Bangladeshi' and 'Any other Asian background').

Although this still resulted in the merging of some ethnic categories, this was necessary to provide a sufficiently large comparison group (153 students), but still resulted in a group likely to be more similar than all 'Non-whites'. The *dif* analyses conducted on these two groups revealed only one verbal item which showed a large degree of bias, with this item favouring Asian students. None of the math items were identified as showing significant bias towards Asians or Whites.

As *dif* analyses can only compare the performance of two groups at a time, it was necessary to compare each sample with each of the others. This resulted in a total of six analyses being conducted (three comparisons between samples by the verbal and math sections). Across all six comparisons, only two items were identified as displaying a large degree of bias. Both of these were in the comparisons between the low-achieving schools and the independent schools, and showed one verbal item to favour students from the independent schools, and one math item to favour students from the low-achieving schools.

The *dif* analyses conducted on the SAT I: Reasoning Test provide very little evidence of substantial bias at the item level according to sex, ethnicity or sample. Test-level score differences are therefore likely to result from some groups of students performing marginally better on a number of questions, rather than substantially better on a few. This indicates that the screening of items used for SAT construction appears to work well, at least for the version of the SAT considered here, although the extent to which *dif* data from American students is applicable to British students requires further study.

Conclusions

This report has presented the findings from a study which examined the association between A-level grades and the American SAT I: Reasoning Test in three samples of schools. This work was prompted by media discussion on the possibility of using a SAT-like aptitude test as part of the entrance procedure to British universities.

An important question which was not addressed in this debate was whether the SAT I: Reasoning Test provided information that was different from the information already conveyed through A-level grades. The present study found that the SAT I: Reasoning Test and A-levels were assessing relatively distinct constructs. The exact overlap between the two varied according to sample, being lowest in the independent schools, and higher in the low- and high-achieving schools. The most likely explanation for this is the restriction in range of A-levels in the independent schools, due to the high proportion of students attaining A grades. Outside of the independent schools, the strength of the association between the SAT I: Reasoning Test and attained A-levels was comparable to that seen previously (e.g. Choppin *et al.*, 1972), with the two sharing approximately 25 per cent of variance. A-level grades and the SAT I: Reasoning Test therefore assess somewhat distinct constructs, suggesting that the SAT, or a test like it, may be of value in predicting university performance.

An interesting finding from the data was that mean GCSE grades were more closely associated with SAT I: Reasoning Test scores than A-levels. This was unexpected due to the test having been taken far closer to the time students took their A-levels than GCSEs, but concurred with Fremer *et al.'s* (1968) previous study of the SAT. One reason for this may be that as students are required to study a range of subjects at GCSE, mean GCSE grades are a greater reflection of general ability than A-levels, where students may choose to specialise in an particular area.

The main analysis examined the ability of exam attainment, SAT I: Reasoning Test scores and background variables to predict A-level performance. An initial multilevel model showed predicted A-levels to have the closest association with attained A-levels, closely followed by GCSE grades. The association between total SAT I: Reasoning Test scores and A-levels was significant, but somewhat lower. When the two SAT I: Reasoning Test sections were considered separately, the verbal section showed the stronger association with A-levels.

When background variables were entered into the analysis, the strongest predictor of attained A-levels was sample, indicating that the highest mean A-level grades were obtained by the independent schools, followed respectively by the high- and then the low-achieving schools. An interaction between sample and predicted A-level grades also emerged, indicating that students from the independent schools who were predicted to attain lower grades performed

better than expected. This may have resulted from independent schools having the facilities to better prepare their students for exams, or alternatively, for these schools to under-predict grades at the lower end of the continuum. A further significant effect was seen for sex, showing that after background factors had been taken into account, females attained higher A-levels than males.

Measures were also taken of parental socio-economic status, coded as the highest occupational category of a student's male and female parents or carers. Students whose parents were in the occupational categories Corporate manager, senior official or Clerk or secretary attained significantly lower A-levels than other students. An alternative approach to studying the effects of socio-economic status was to treat this as a continuous variable instead of a categorical one. This was not done initially due to concerns over it not being a linear variable, but an exploratory analysis treating it in this way was conducted. This failed to find a significant effect for parental socio-economic status.

An important question was whether the association between the SAT I: Reasoning Test and A-levels varied according to background factors, such as sample and socio-economic status. In the media debate on the SAT, claims were made that it was able to assess students' potential independent of their social and educational experiences (e.g. Clare, 1999), although no empirical evidence was presented to support this. No evidence was found to support this possibility – both SAT I: Reasoning Test scores and mean A-level grades were seen to be highest in the independent schools and lowest in the low-achieving schools, and the associations between the two did not vary between sample. This conclusion was further supported by analysing the regression lines for each sample, which showed very similar slopes for high- and low-attaining schools. Slightly different slopes were observed for the independent schools, but this is likely to be due to the highly skewed distribution of exam results in this sample.

An alternative way of looking at the association between the SAT I: Reasoning Test and A-levels was to examine test scores when A-level attainment was held constant. This was done by placing students into three A-level grade bands: A/B, C and D or below. Within each of these bands SAT I: Reasoning Test scores were seen to vary significantly, with the highest scores being obtained by the independent schools and the lowest by the low-attaining schools. Very similar findings emerged when students were grouped on the basis of their GCSE grades. Adopting the SAT scores as a threshold for selection gives roughly the same proportion of students for consideration, but these may not be the same individuals. Using the SAT scores in addition to A-levels increases the number of students selected for all the samples. However, the percentage increase is greatest for students from low-attaining schools.

This apparent inconsistency between the multilevel model and analyses which placed students into grade bands can most readily be explained through the different statistical methods and the skewed distributions of the data.

As the SAT is designed for, and primarily taken by, American high school students, a further series of analyses was conducted to examine how the SAT I: Reasoning Test functioned in a sample of British students. Reliability analyses indicated that the SAT I: Reasoning Test provided a coherent assessment of reasoning abilities, both when total test score, and verbal and math sections were considered. Comparisons between the performance of British and American students showed considerable concordance in the difficulty and discrimination of SAT items. It also appeared that British students were able to complete the majority of each SAT I: Reasoning Test section in the time allowed, although completion rates for the verbal section were somewhat higher than the math. Finally, the SAT is screened for bias on American students as part of the development process. Bias analyses conducted on the British data showed very little evidence of differential item functioning, suggesting that bias analyses on American students may have validity for British samples.

Whilst this study has provided no evidence for the claims that the SAT is able to assess ability for university study independently of a student's background, it has shown that the SAT I: Reasoning Test and A-levels measure relatively distinct constructs. Therefore the SAT would be worth further study, particularly by looking at how it can predict performance at university. Whilst such work would require a considerable time to conduct, it is essential if the debate on aptitude testing for university entrance is to progress.

References

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.

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.

CARVEL, J. (2000). 'Plea for US-style university entrance tests to end "bias" against poor', *The Guardian*, 11 April.

CHARTER, D. (2000). 'Oxford will use test to spot creative thinkers', *The Times*, 31 July.

CHOPPIN, B. and ORR, L. (1976). *Aptitude Testing at Eighteen-Plus*. Windsor: NFER Publishing Co.

CHOPPIN, B.H.L., ORR, L., FARA, P., KURLE, S.D.M., FOGELMAN, K.R. and JAMES, G. (1972). *After A-Level? A Study of the Transition from School to Higher Education*. 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 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].

COLLEGE ENTRANCE EXAMINATION BOARD (n.d.). *SAT I: Reasoning Test*. New York, NY: College Board.

FREMER, J., COFFMAN, W.E. and TAYLOR, P.H. (1968). 'The College Board Scholastic Aptitude Test as a predictor of academic achievement in secondary schools in England', *Journal of Educational Measurement*, **5**, 3, 235–41.

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.

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.

LEWIS, S.D. (2000). 'a) frying pan or b) fire?'(Education), *The Guardian*, 6 June, (insert, 12–13).

McDONALD, A.S., NEWTON, P.E., WHETTON, C. and BENEFIELD, P. (2001). *Aptitude Testing for University Entrance: a Literature Review*. 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.

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.

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.

RICHARDSON, K. (2000). 'New tests, old results', *Times Higher Educ. Suppl.*, **1432**, 21 April, 16.

SCOTT, P. (1995). *The Meanings of Mass Higher Education*. Buckingham: Open University Press.

STEIN, J. (2000). 'A true test of talent', *Times Higher Educ. Suppl.*, **1450**, 25 August, 14.

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.

THE SUTTON TRUST (2000). *Entry to Leading Universities. Executive Summary* [online]. Available: http://www.suttontrust.com/text/Report1.doc [29 September, 2000].

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. 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.

Appendix 1: Score distributions of main study variables

Table 1.1: Mean GCSE score distributions by sample

	Low-achieving schools			High-achieving schools			Independent schools		
	Freq.	%	Cum.%	Freq.	%	Cum.%	Freq.	%	Cum.%
A*	6	1.0	1.0	28	5.1	5.1	34	34.0	34.0
A	85	13.6	14.6	181	32.9	38.0	59	59.0	93.0
B	254	40.6	55.2	255	46.4	84.4	7	7.0	100.0
C	236	37.8	93.0	83	15.1	99.5			
D	42	6.7	99.7	3	0.5	100.0			
E & below	2	0.3	100.0						

Table 1.2: Mean predicted A-level score distributions by sample

	Low achieving schools			High-achieving schools			Independent schools		
	Freq.	%	Cum.%	Freq.	%	Cum.%	Freq.	%	Cum.%
A	20	3.3	3.3	66	12.3	12.3	62	62.6	62.6
A/B	48	8.0	11.3	93	17.4	29.7	20	20.2	82.8
B	76	12.6	24.0	90	16.8	46.5	12	12.1	94.9
B/C	110	18.3	42.3	95	17.8	64.3	3	3.0	98.0
C	97	16.1	58.4	80	15.0	79.3			
C/D	88	14.6	73.0	70	13.1	92.3	1	1.0	99.0
D	72	12.0	85.0	21	3.9	96.3	1	1.0	100.0
D/E	52	8.7	93.7	12	2.2	98.5			
E & below	38	6.3	100.0	8	1.5	100.0			

Table 1.3: Mean A-level score distributions by sample

	Low-achieving schools			High-achieving schools			Independent schools		
	Freq.	%	Cum.%	Freq.	%	Cum.%	Freq.	%	Cum.%
A	14	2.3	2.3	41	7.7	7.7	48	48.5	48.5
B	105	17.5	19.9	191	35.8	43.5	40	40.4	88.9
C	128	21.4	41.2	132	24.8	68.3	8	8.1	97.0
D	136	22.7	63.9	85	15.9	84.2	1	1.0	98.0
E & below	216	36.1	100.0	84	15.8	100.0	2	2.0	100.0

Table 1.4: Total SAT score distributions by sample

	Low-achieving schools			High-achieving schools			Independent schools		
	Freq.	%	Cum.%	Freq.	%	Cum.%	Freq.	%	Cum.%
550-590	1	0.2	0.2						
600-640	8	1.3	1.5	2	0.4	0.4			
650-690	23	3.8	5.2	3	0.5	0.9			
700-740	33	5.4	10.6	13	2.4	3.3			
750-790	46	7.5	18.2	17	3.1	6.4			
800-840	65	10.6	28.8	30	5.5	11.9			
850-890	69	11.3	40.1	40	7.3	19.2			
900-940	81	13.3	53.4	67	12.3	31.5	2	2.0	2.0
950-990	73	11.9	65.3	61	11.2	42.7	1	1.0	3.0
1000-1040	59	9.7	75.0	65	11.9	54.6	5	5.0	7.9
1050-1090	57	9.3	84.3	60	11.0	65.6	9	8.9	16.8
1100-1140	34	5.6	89.9	60	11.0	76.6	7	6.9	23.8
1150-1190	32	5.2	95.1	56	10.3	86.8	13	12.9	36.6
1200-1240	13	2.1	97.2	29	5.3	92.1	15	14.9	51.5
1250-1290	10	1.6	98.9	14	2.6	94.7	16	15.8	67.3
1300-1340	5	0.8	99.7	15	2.7	97.4	12	11.9	79.2
1350-1390	2	0.3	100.0	9	1.6	99.1	13	12.9	92.1
1400-1440				4	0.7	99.8	3	3.0	95.0
1450-1490				1	0.2	100.0	5	5.0	100.0

Table 1.5: SAT verbal score distributions by sample

	Low-achieving schools			High-achieving schools			Independent schools		
	Freq.	%	Cum.%	Freq.	%	Cum.%	Freq.	%	Cum.%
250-290	3	0.5	0.5						
300-340	23	3.8	4.3	4	0.7	0.7			
350-390	80	13.1	17.3	27	4.9	5.7			
400-440	107	17.5	34.9	68	12.5	18.1	3	3.0	3.0
450-490	130	21.3	56.1	107	19.6	37.7	6	5.9	9.0
500-540	112	18.3	74.5	116	21.2	59.0	14	13.9	23.0
550-590	80	13.1	87.6	108	19.8	78.8	17	16.8	40.0
600-640	59	9.7	97.2	80	14.7	93.4	30	29.7	70.0
650-690	14	2.3	99.5	27	4.9	98.4	18	17.8	88.0
700-740	3	0.5	100.0	6	1.1	99.5	11	10.9	99.0
750-790				3	0.5	100.0	1	1.0	100.0

Table 1.6: SAT math score distributions by sample

	Low-achieving schools			High-achieving schools			Independent schools		
	Freq.	%	Cum.%	Freq.	%	Cum.%	Freq.	%	Cum.%
200-240	3	0.5	0.5						
250-290	7	1.1	1.6	3	0.5	0.6			
300-340	69	11.3	12.9	32	5.9	6.4			
350-390	93	15.2	28.2	39	7.1	13.6			
400-440	142	23.2	51.4	90	16.5	30.1	3	3.0	3.0
450-490	81	13.3	64.6	80	14.7	44.8	2	2.0	5.1
500-540	94	15.4	80.0	116	21.2	66.1	12	11.9	17.2
550-590	74	12.1	92.1	75	13.7	79.8	18	17.8	35.4
600-640	27	4.4	96.6	63	11.5	91.4	16	15.8	51.5
650-690	19	3.1	99.7	34	6.2	97.6	32	31.7	83.8
700-740				5	0.9	98.5	5	5.0	88.9
750-790	2	0.3	100.0	8	1.5	100.0	11	10.9	100.0

Appendix 2: Classification of universities and colleges

The typology was based on one developed by Peter Scott in his book *The Meanings of Mass Higher Education* (Scott, 1995). The classification is based largely on historical differences.

Type	Description	Universities included in this category
Oxford, Cambridge	ancient 12th and 13th century foundations	
Civic universities and London	established in London in the early 19th century and in other major English cities in the later part of the century	Birmingham, Bristol, Leeds, Liverpool, London, Manchester, Sheffield
Redbrick universities	founded in other cities in the early 20th century	Exeter, Hull, Leicester, Nottingham, Reading, Southampton
Durham, Keele	two anomalies: Durham founded in the early 19th century on the Oxbridge model, Keele founded after the Second World War, offering a four-year degree	Durham, Keele
Technological universities	created from former colleges of advanced technology in the 1960s	Aston, Bath, Bradford, Brunel, City, Loughborough, Salford, Surrey
Scottish universities		
Welsh universities		
Northern Irish universities		
Open University	national distance learning institution, founded in the 1960s	Open University
'Old new' universities	founded in the 1960s on campus locations	East Anglia, Essex, Kent, Lancaster, Sussex, Warwick, York
'New new' universities	polytechnics and colleges which were redesignated as universities in the early 1990s	For example: Anglia, Brighton, Central England, Coventry, De Montfort, East London, Hertfordshire, Leeds Metropolitan, Liverpool John Moores, Oxford Brookes, Sheffield Hallam, Westminster
Other	includes: colleges of higher and further education, specialised colleges and overseas institutions	

Appendix 3: Details of multilevel modelling

written by Samantha E. Higgs

Introduction

The following types of data were available for pupils:

- average achieved A-level scores;
- average achieved GCSE scores;
- average predicted A-level;
- math SAT I: Reasoning Test score;
- verbal SAT I: Reasoning Test score;
- total SAT I: Reasoning Test score (total of math and verbal scores);
- pupil background data;
- school background data.

The aim of the analysis was to investigate whether SAT I: Reasoning Test is a better predictor of university performance than A-level for certain groups of students. However, as we have no university data we could not answer this question. The next best thing is to see if SAT I: Reasoning Test predicts A-level in some way for certain groups, as A-level is the current predictor of university performance.

We investigated background factors at the school and pupil levels which might be associated with A-level scores, to see which were apparently statistically significant and whether the association between A-levels and the SAT I: Reasoning Test varies according to school type. We also looked at the association between A-levels and verbal scores and A-levels and math scores.

Setting up multilevel models

Multilevel modelling is a development of a common statistical technique known as 'regression analysis'. This is a technique for finding a straight-line relationship which allows us to predict the values of some measure of interest ('dependent variable') given the values of one or more related measures. For example, we may wish to predict schools' average test performance given some background factors, such as free school meals and school size (these are sometimes called 'independent variables').

Multilevel modelling is a recent development of regression analysis which takes account of data which is grouped into similar clusters at different levels. For example, individual pupils are grouped into year groups or cohorts, and those cohorts are grouped within schools. There may be more in common between pupils within the same cohort than with other cohorts, and there may be elements of similarity between different cohorts in the same school. Multilevel modelling allows us to take account of this hierarchical structure of the data and produce more accurate predictions, as well as estimates of the differences between pupils, between cohorts, and between schools.

When setting up the model, average achieved A-level was used as the outcome measure.

The model fitted to the data incorporated two levels:

1. school;
2. pupil.

Thus, there are assumed to be variations between schools in their average scores, and within a school there are almost bound to be variations between pupils. The sizes of these variations at each level of the model are measured in terms of 'random variances', and the relative sizes of these will be of some interest.

The fitting process was carried out in two stages:

1. the 'base case', with no background variables;
2. controlling for pupil-level and school-level background variables.

Background variables were included in the model, but to see if different groups of students, for example boys versus girls, low-achieving sample versus high-achieving sample, performed in different ways, we need to include 'interaction terms' in the model, which relate background factors to different relationships between SAT I: Reasoning Test score and outcome.

An example of an interaction term is SATSEX, which is positive. The interpretation of the model results for these variables is straightforward. If, for example, the coefficient of SATSEX is positive, this implies that the relationship between SAT I: Reasoning Test scores and A-level grades is stronger for girls than boys. A negative coefficient for GCSESEX would imply that boys with higher GCSE grades are achieving lower A-levels than equivalent girls, and so forth.

Results of multilevel analysis

Table 3.1 contains details of all the variables derived from the data collection exercise which were used in the analysis of these pupils. Categorical variables had to be broken down into dichotomous variables (0,1) in order to compare different groups of pupils. The variables of this kind were ethnicity and socio-economic status (SES). Tables 3.2 and 3.3 show some of the detailed results of the multilevel models fitted to the outcome measure. In technical language, these tables show the random variances at each level at each stage of model fitting, plus the coefficients of the background variables in the 'full model'. They also show whether or not variances or coefficients are statistically significant at the five per cent level, as well as 95 per cent confidence intervals for each parameter.

These tables, although they show the full results of all the modelling carried out at this stage, may not be easy to interpret for all readers. To help with this, therefore, the coefficients which express the estimated relationships between test scores and each of the background variables have been converted into 'normalised coefficients' which represent the 'strength' of each relationship as a percentage, and which allow the different variables to be compared in terms of their apparent influence on the test outcome, when all other variables are simultaneously taken into account.

Normalised coefficients are plotted in Figure 3.1. For each variable, the estimated normalised coefficient is plotted as a diamond, with a vertical line indicating the 95 per cent confidence interval for the estimate. Any variable whose line intersects the horizontal zero axis can be regarded as not statistically significant (at the five per cent level). Positive values imply a positive relationship with the test score outcome; negative values imply that test score tends to decrease with higher values of the given background variable.

A further element of the model fitted was the investigation of possible differential relationships of A-level scores with SAT I: Reasoning Test scores between schools. This was modelled by allowing the coefficient of SAT I: Reasoning Test score to vary from school to school. This effect was estimated as zero, implying that there were no detectable variations between schools in this relationship.

Figure 3.1: Normalised coefficients

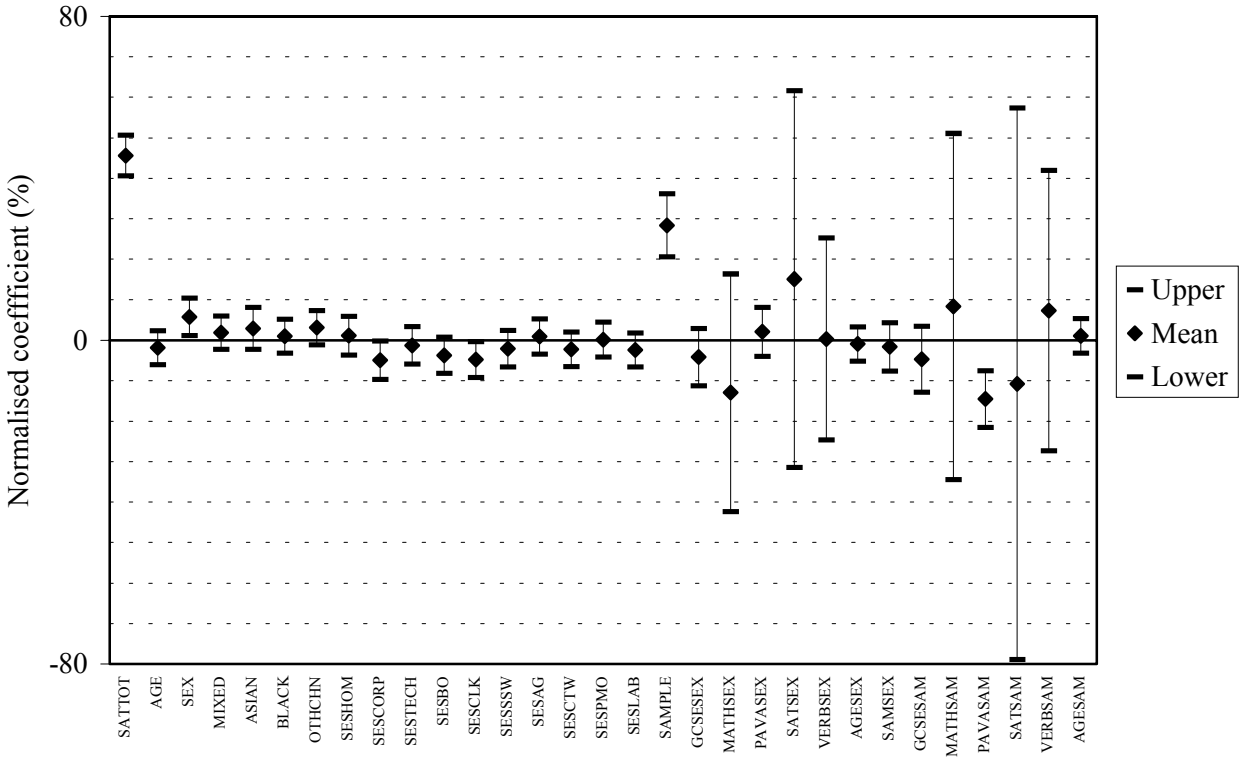
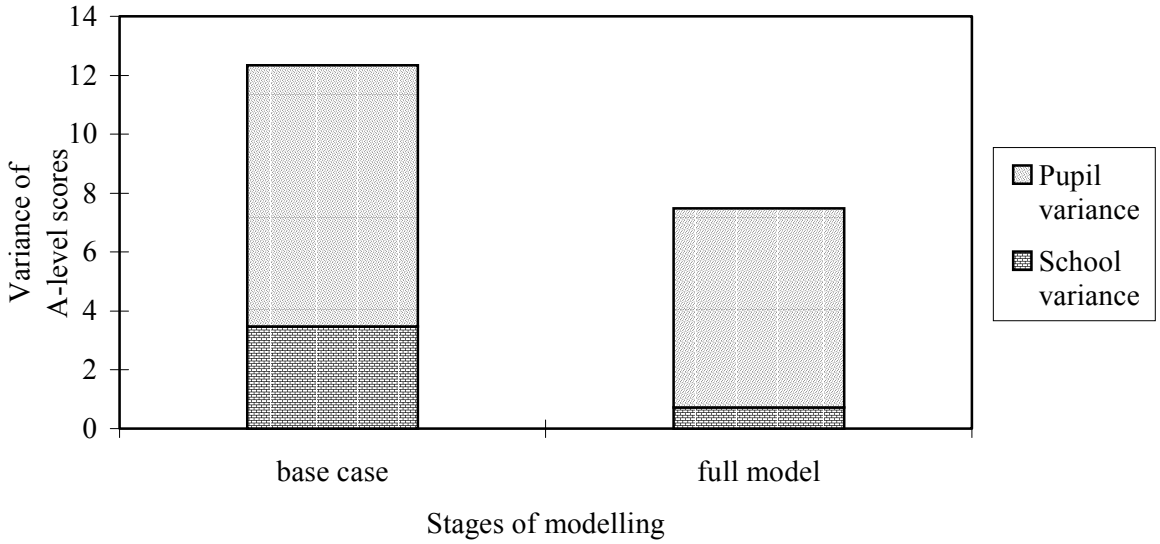


Figure 3.2: Random variances in A-levels at different levels with and without background variables



In addition to the relationships between test scores and a host of background variables described above, the multilevel model provides other information. In particular, it estimates the amount of variation in A-level grades which can be attributed to different levels in the model. The amount of variation at each level is measured by the 'variance' (basically the square of the standard deviation) at that level, and may change as extra background variables are fitted to the model.

Figure 3.2 illustrates this effect based on average achieved A-level scores. For the outcome measure, at each stage of modelling, the total variance is divided between the two levels in the model. The introduction of background variables reduces pupil-level variance by about a quarter and school-level variance by about three-quarters.

Summary of results

In this section we shall briefly summarise the findings from the multilevel analysis of the data collected.

- By far the strongest relationship for average A-level was with total SAT I: Reasoning Test score.
- The next most significant term was school type (sample) which was positive, implying pupils in the higher-achieving schools were attaining higher A-level scores (but see later discussion).
- Sex had a positive coefficient implying that girls achieved higher A-level scores than boys, allowing for other factors, including SAT I: Reasoning Test score.
- The interaction term 'PAVASAM' had a negative coefficient. This seems to imply that pupils in high-achieving schools who had been predicted low A-level scores were actually performing better when it came to actual achieved A-level.
- The other interaction terms were insignificant which meant there was no evidence that different groups of students performed in different ways when comparing the relationship of A-level with SAT I: Reasoning Test score.
- Various ethnic groups did not appear to perform differentially relative to whites.
- For social class, each group was compared to the 'Professional' category due to the fact that it had the largest mean. The variable 'SESCORP', Corporate managers, and 'SESCLK', Clerks or secretaries, both had a negative coefficient implying that students who had a parent who was either a Corporate manager or a Clerk/secretary achieved lower

A-level scores than those who had a parent in the Professional category. The other variables were not significantly related to A-level score.

- Age was not significantly related to A-level score.
- Approximately three-quarters of the variance between schools was explained by background factors, in particular the SAT I: Reasoning Test score of the pupils.

Fitting socio-economic status as a continuous variable

It was decided that instead of breaking the SES variable down into dichotomous variables, another approach would be to treat it as a continuous variable. We had a problem concerning the first category 'never worked outside the home for pay', and this would inevitably include housewives/husbands. In order to remove them from the analysis, this category was set to missing before taking the minimum value of the two parents' professions. There were only two cases where neither parent had ever worked outside the home for pay and these were also set to missing. The values of this variable were recoded so that the 'highest' Professional class was represented by the highest number.

The model was then re-fitted with this new continuous variable but it was not significant. Interaction terms were also created with this SES variable but again, none were significant implying that within each group, males for example were insignificantly different from females with regard to their A-level score. One theory could be that the A-level score has mostly been explained by the SAT I: Reasoning Test score. The normalised coefficients and the variance accounted for can be seen in Figures 3.3 and 3.4. The results are similar to those found in the first model.

Figure 3.3: Normalised coefficients when fitting SES as continuous

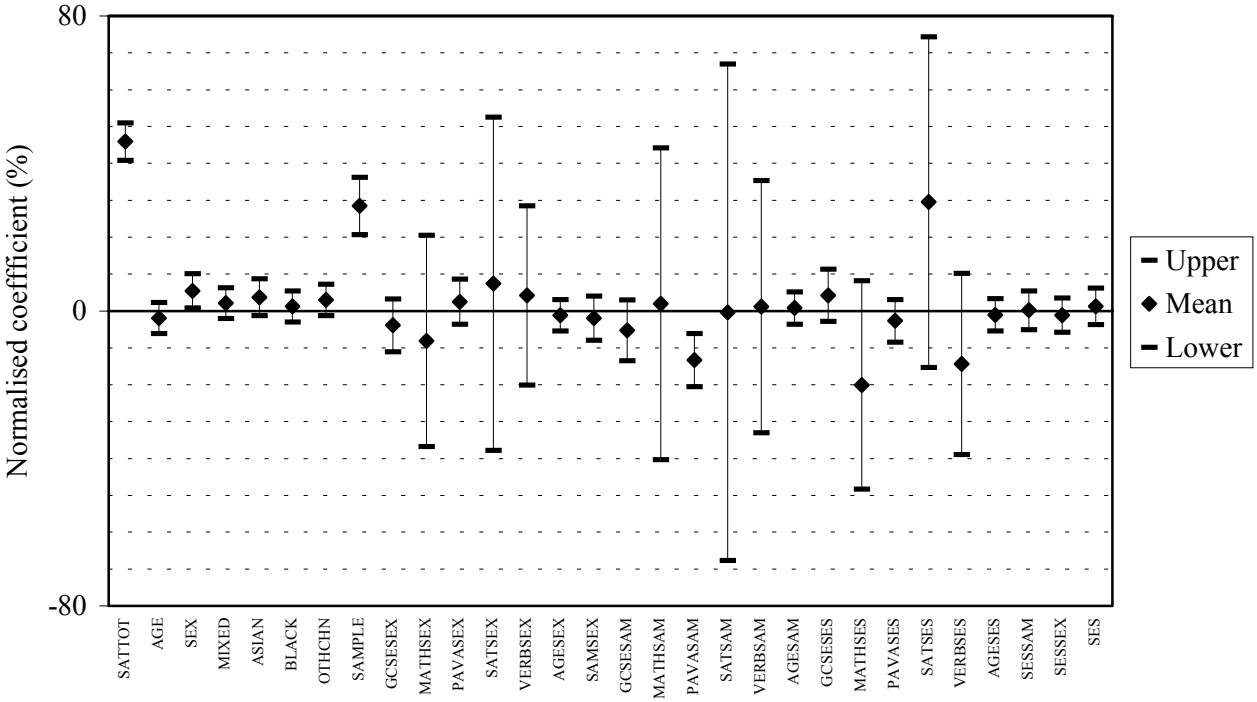
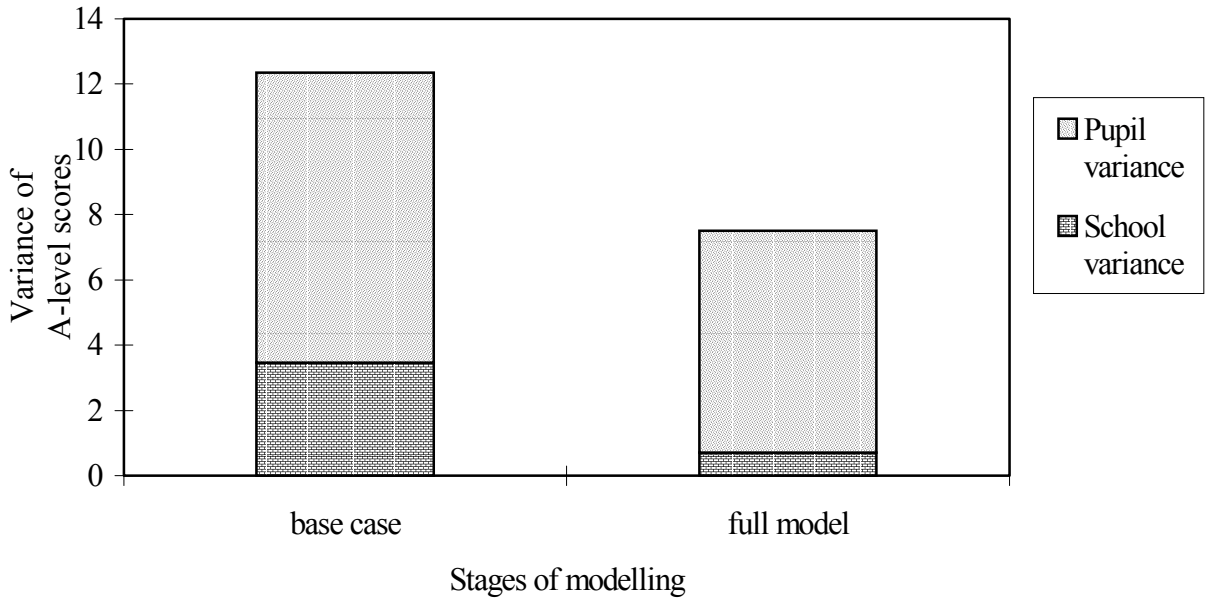


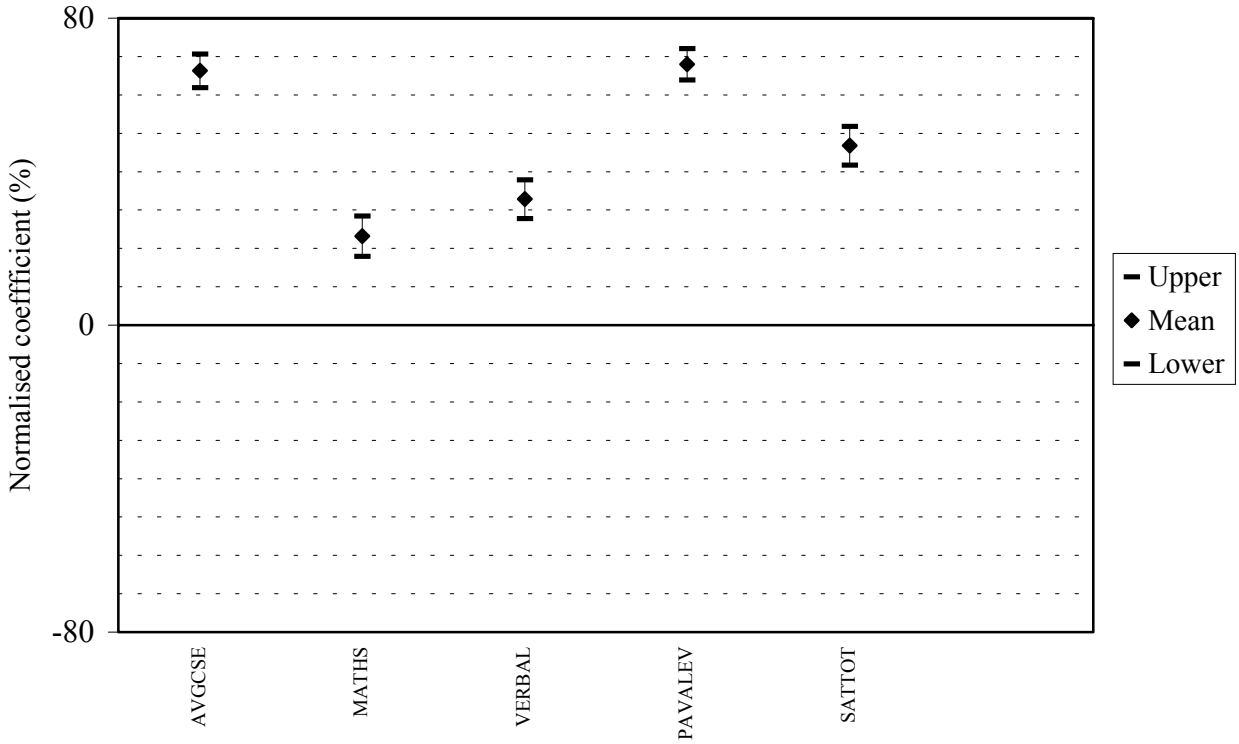
Figure 3.4: Random variances, SES as continuous



Best predictor of A-level

Another area of interest was which variable was the best predictor of A-level. In order to do this, the main variables, GCSE, MATHS and VERBAL SATS, PAVALEV and SATTOT were individually fitted and a separate model was produced for each. They could not all be fitted together due to high intercorrelations, but math and verbal SAT I: Reasoning Test scores were fitted in the same model.

Figure 3.5: Normalised coefficients, fitting four separate models



From the above chart it is apparent that, not surprisingly, predicted A-level is the best predictor of A-level. The next best predictor is GCSE score.

Fitting math and verbal SAT scores as explanatory terms

From Figure 3.5 we can see that the verbal SAT I: Reasoning Test score appears to be a better predictor of A-level than the math score. It was thought that instead of fitting SATTOT, which is just the total of each pupil's math and verbal scores, it would be of interest to fit these terms individually to the model. The normalised coefficients and the random variances can be seen in Figures 3.6 - 3.9 for each model, with the full results in Table 3.4 and 3.5. The results are similar to those we found in the first model, where total SAT I: Reasoning Test score was fitted as an explanatory variable, although there are a few differences. The variable SESCORP, pupil with a parent who was a Corporate manager, did not appear as a significant term in either of the math or verbal models but was significant with SATTOT. Sex was not significant in the verbal model but was in the math model, suggesting that girls were performing better than expected at A-level when looking at their math scores but not when looking at their verbal scores. ASIAN and OTHCHN were both significant with a positive coefficient, implying that these pupils, given their verbal score, are doing better than the white pupils at A-level.

Figure 3.6: Normalised coefficients fitting math SAT score

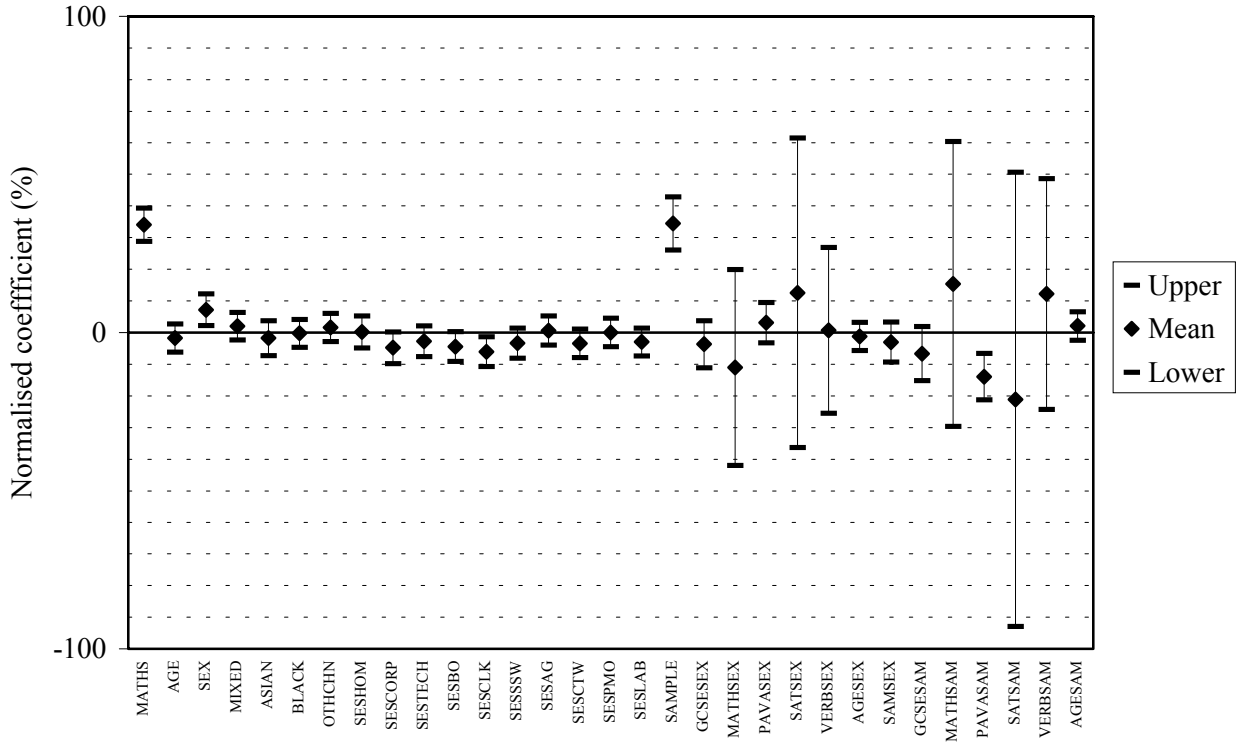


Figure 3.7: Random variances fitting math SAT score

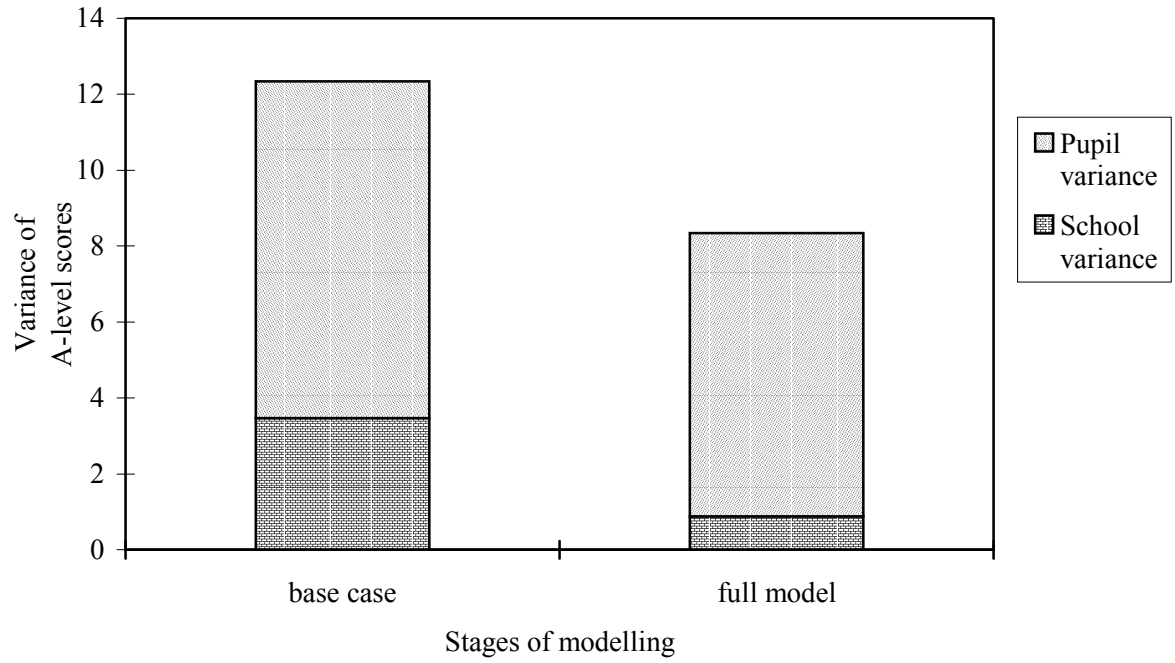


Figure 3.8: Normalised coefficients fitting verbal SAT score

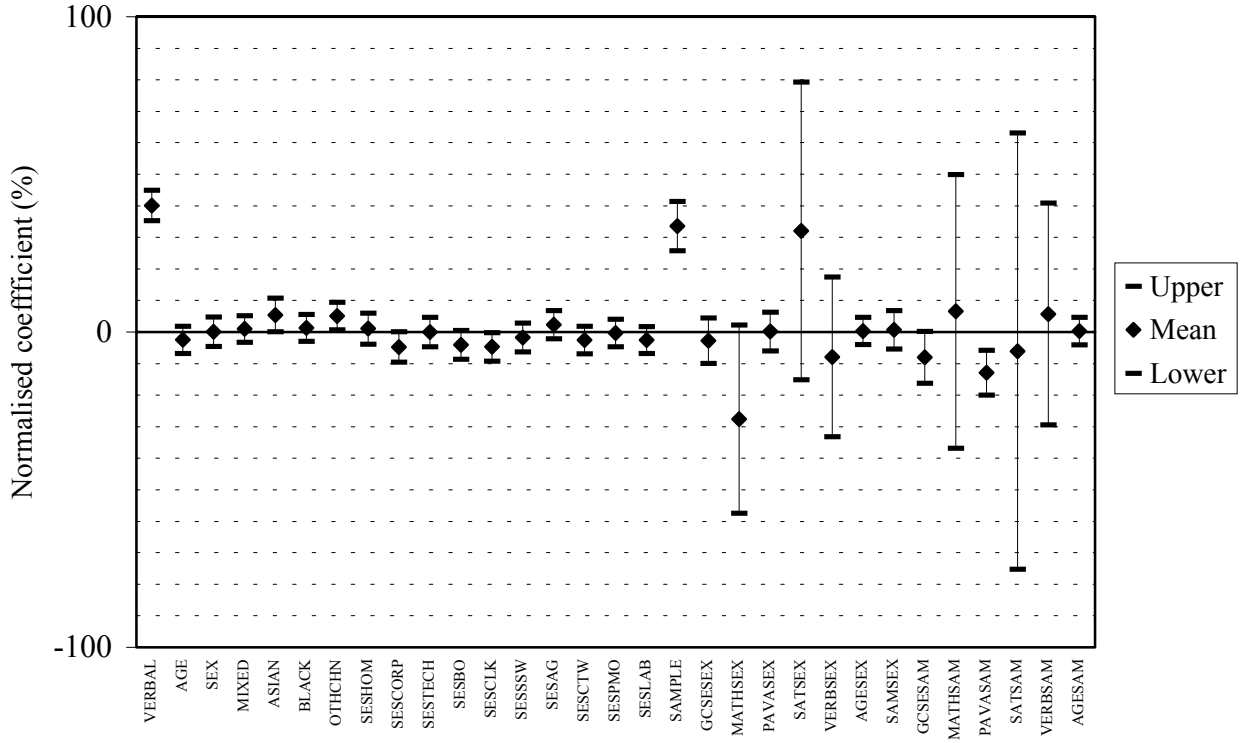
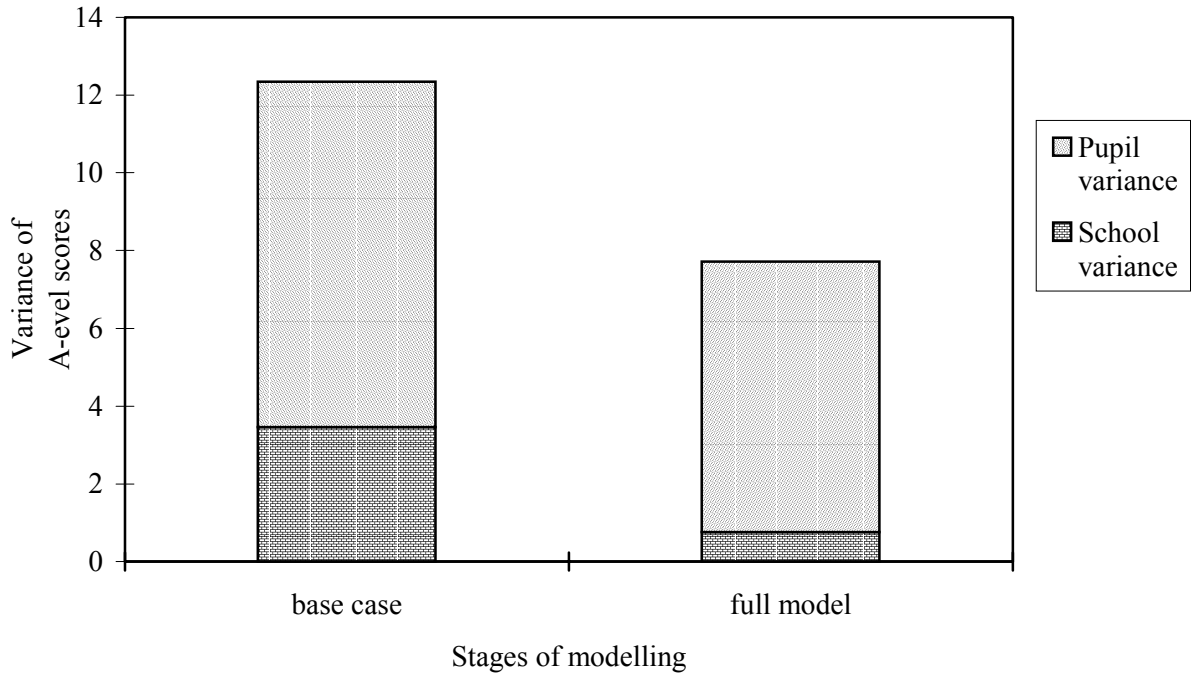


Figure 3.9: Random variances fitting verbal SAT score



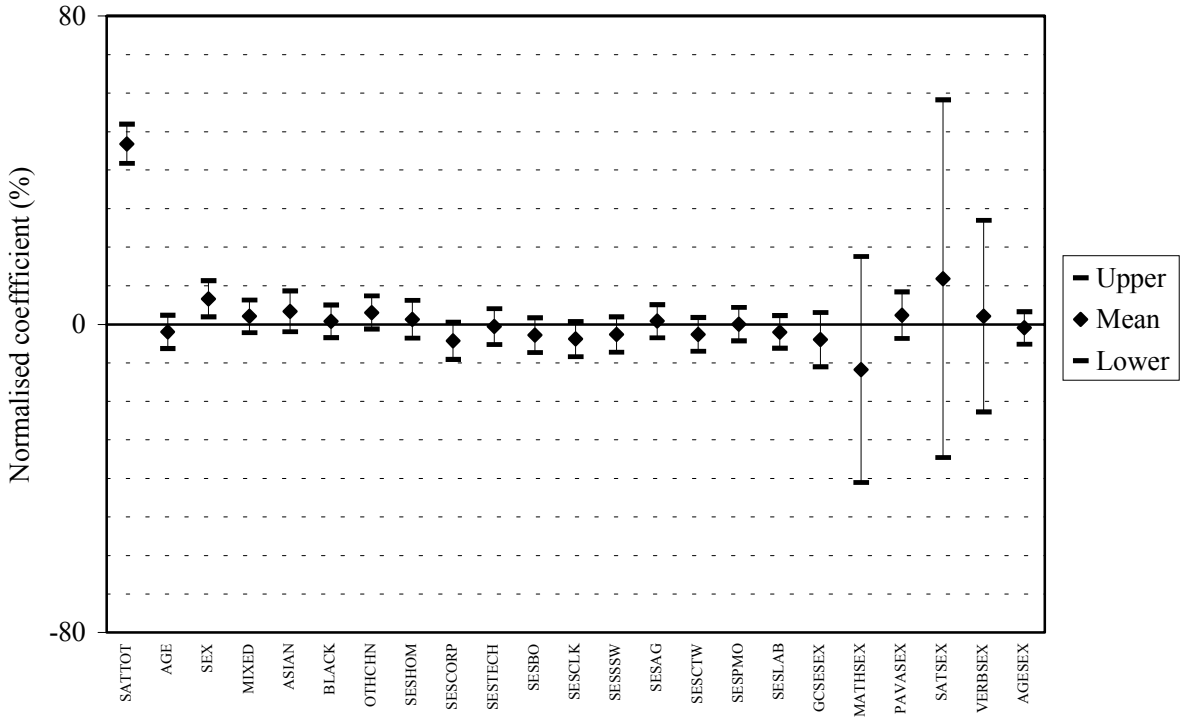
Fitting a three-level multilevel model

When looking at the previous models, it appeared that the variable SAMPLE is a significant term. An alternative approach, instead of fitting SAMPLE as an explanatory term, would be to treat this variable as a third level in the model. This would make it a little more complicated to interpret as we now have pupils within schools within sample. In this particular case, this approach is not recommended due to only having three groups at this third level, and one of these only having 100 pupils. Also, we would not actually be able to detect where the differences lie, that is between which samples, but only if the difference was there. However, it was decided to run the model anyway and look at the variances at each level.

The normalised coefficients and the random variances can be seen in Figures 3.10 and 3.11. The results are similar to those we found in the first model, where SATTOT was fitted as an explanatory variable although only SATTOT and SEX are significant now. In the base case the random variances at the third level are larger than at school level, implying that there are larger differences between samples than between schools. In the full model, the variances at each level are lower which means we have explained some of the variation by adding the explanatory variables. It should be noted that the random variances at the sample level have large standard errors. The variances are 3.933 and 0.783 for the base and full model

respectively with standard errors of 3.31 and 0.7123 clearly making them not statistically significant. This is probably due to only comparing three groups and therefore only using two degrees of freedom.

Figure 3.10: **Normalised coefficients fitting a three-level model**



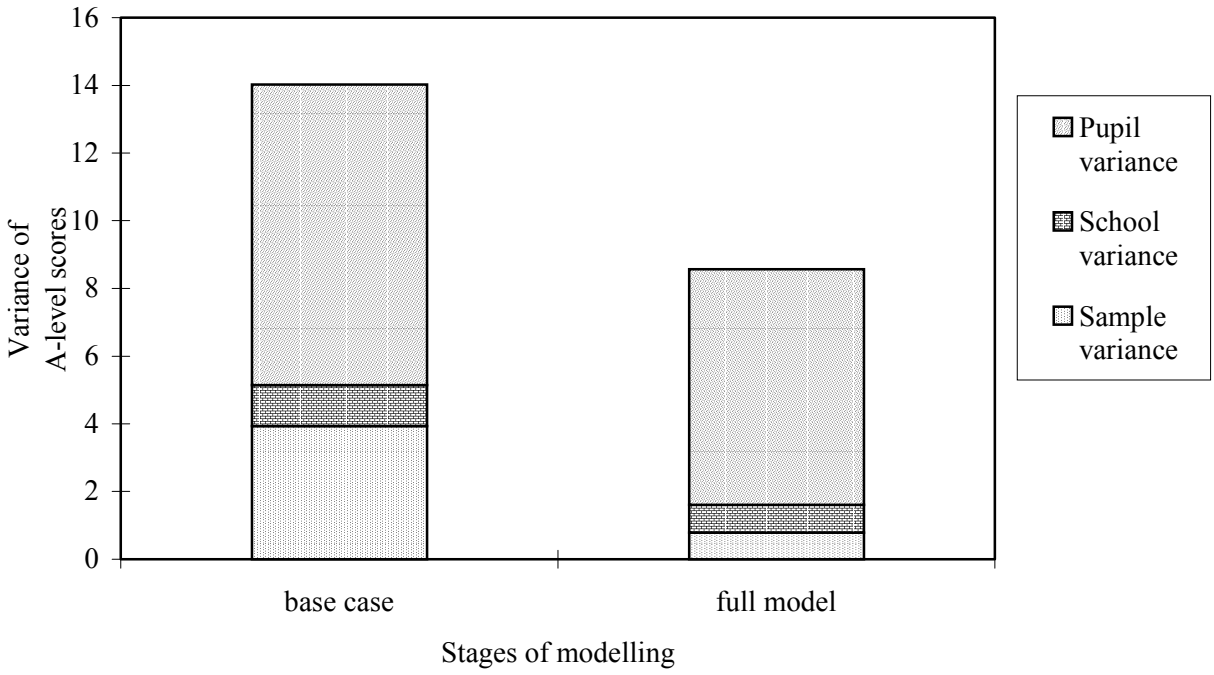


Figure 3.11: Random variances fitting a three-level model

Final conclusions

From the models that were fitted it appears that after SAT I: Reasoning Test score the most significant term is sample (school type). This consisted of three samples of schools: low-achieving, high-achieving and very high-achieving independent schools. However, during some exploratory analysis the model was switched around so that we had SAT I: Reasoning Test score as our outcome variable. This produced a similar result, implying that again sample was significant. These models should be the mirror image of one another so this result was a little strange. Scatterplots were then produced in order to look at the data and figure out what was happening (see Figures 3.12 and 3.13 and Figures 2 to 4 in the main text).

It appears that the apparent effect of school type (SAMPLE) is an artefact of the data. From Figure 3.12, pupils in independent schools (open circles) tend to lie above the darker line (regression of A-level on SAT I: Reasoning Test), and the opposite appears to be the case for the lower-achieving schools (solid diamonds). However, because of the scatter in the data the darker line has a reduced slope, which is probably sufficient to explain this effect.

Regression of SAT I: Reasoning Test score on A-level yields the lighter line. In this case pupils in independent schools appear to have higher scores than would be predicted by their A-level grades. Again, this result can be explained by the nature of the data.

Figure 3.12: Scatterplot of A-level v. SAT I: Reasoning Test for all samples

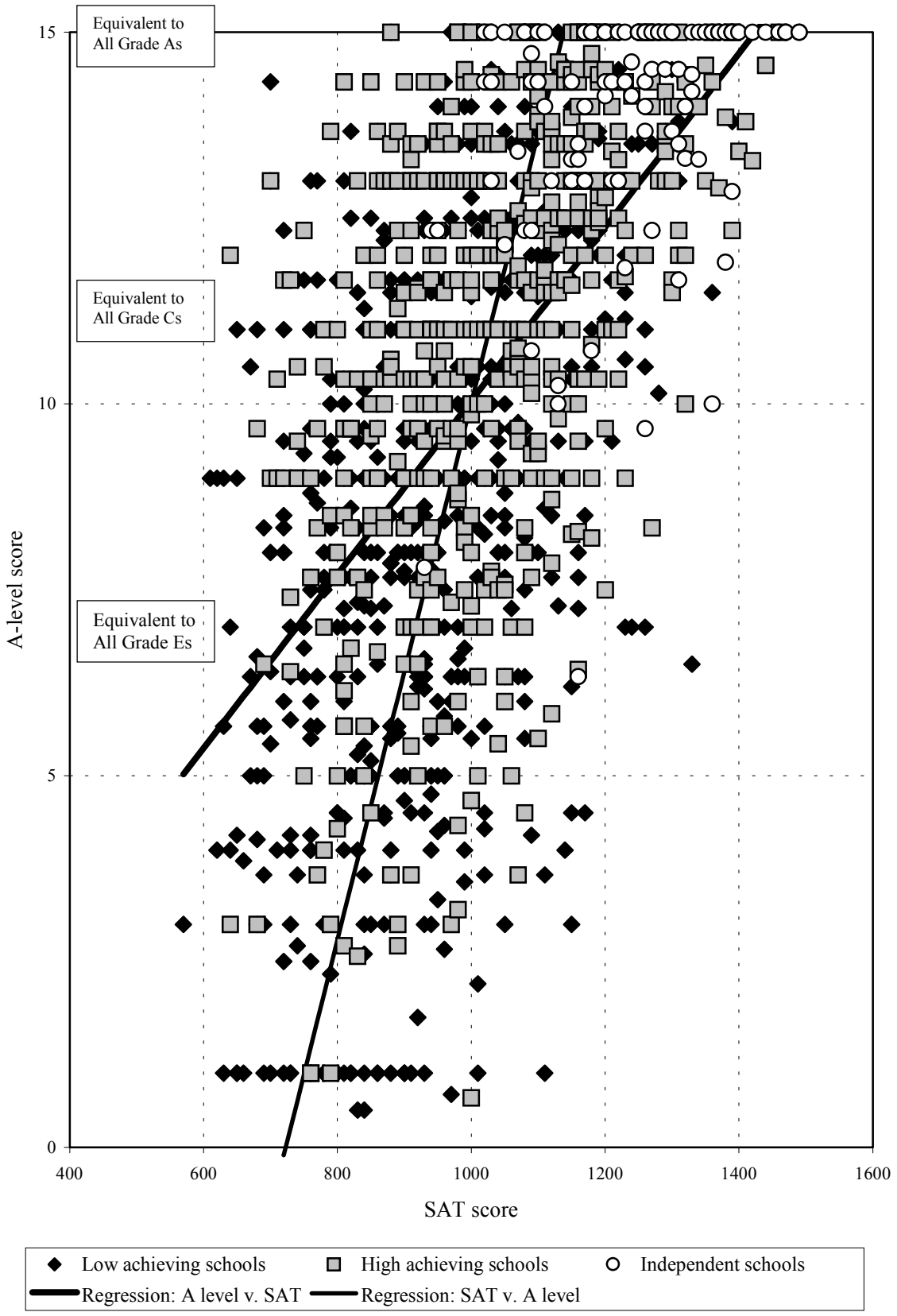


Figure 3.13: Scatterplot of GCSE v. SAT I: Reasoning Test for all samples

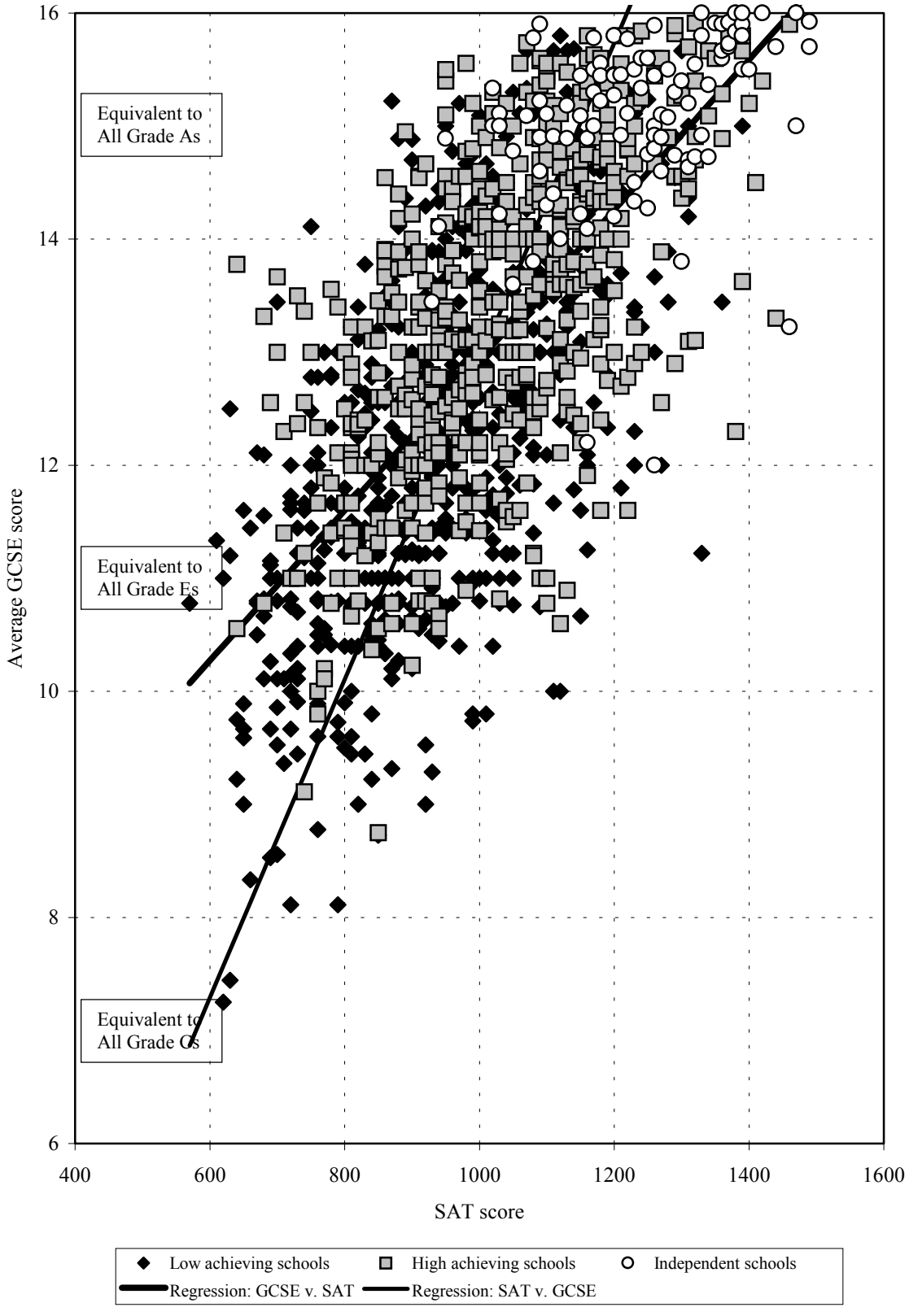


Table 3.1: Details of variables used in multilevel modelling

Var nos	Name	N	Min	Max	Mean	
1	NFER NO	1291	2	219	103.4655	School identifier
2	PUPIL	1291	1	3137	1512.625	Pupil identifier
3	AVALEV	1291	0.5	15	10.04483	Average achieved A-level
4	AVGCSE	1291	7.25	16	12.93547	Average achieved GCSE
5	MATHS	1291	220	800	492.6215	Math SAT score
6	PAVALEV	1291	1	15	11.67208	Average predicted A-level
7	SATTOT	1291	570	1490	1001.372	Total SAT score
8	VERBAL	1291	250	800	508	Verbal SAT score
9	AGE	1291	202	247	218.8729	Age in completed months
10	SEX	1291	0	2	1.078234	Sex(0=male,2=female,1=other)
11	WHITE	1291	0	1	0.826491	White
12	MIXED	1291	0	1	0.012393	Mixed
13	ASIAN	1291	0	1	0.118513	Asian or Asian British
14	BLACK	1291	0	1	0.008521	Black or Black British
15	OTHCHN	1291	0	1	0.021689	Chinese or Other Ethnic Group
16	SESHOM	1291	0	1	0.070488	Never worked outside home
17	SESCORP	1291	0	1	0.211464	Corporate manager
18	SESPROF	1291	0	1	0.298218	Professional
19	SESTECH	1291	0	1	0.134779	Technician
20	SESSBO	1291	0	1	0.064291	Small business owner
21	SESCLK	1291	0	1	0.075136	Clerk or secretary
22	SESSSW	1291	0	1	0.066615	Service or sales worker
23	SESAG	1291	0	1	0.002324	Skilled agricultural worker
24	SESCTW	1291	0	1	0.026336	Craft or trade worker
25	SESPMO	1291	0	1	0.015492	Plant or machine operator
26	SESLAB	1291	0	1	0.01007	General labourer
27	JOBNT	1291	0	1	0.004648	Job without training
28	JOBWT	1291	0	1	0.053447	Job with training
29	YROUT	1291	0	1	0.130906	Take a year out
30	DGREE	1291	0	1	0.724245	Study for a degree
31	FECOL	1291	0	1	0.03718	Study at a FE college
32	DKNOW	1291	0	1	0.02866	Don't know
33	OTHER	1291	0	1	0.008521	Other
34	SAMPLE	1291	1	3	1.59	Sample(1-low,2-high,3-ST)
35	CONS	1291	1	1	1	Constant term
36	GCSESEX	1291	-5.23	5.21	0.071597	Interaction GCSE*SEX
37	MATHSEX	1291	-332	294.4	-26.6082	Interaction MATHS*SEX
38	PAVASEX	1291	-8.44	11.53	-0.05734	Interaction PAVALEV*SEX
39	SATSEX	1291	-528	401.1	-32.1747	Interaction SATTOT*SEX
40	VERBSEX	1291	-315	241	-5.29512	Interaction VERBAL*SEX
41	AGESEX	1291	-20.7	25.88	0.14928	Interaction AGE*SEX
42	SAMSEX	1291	-1.52	1.297	-0.05869	Interaction SAMPLE*SEX
43	GCSESAM	1291	-1.72	4.322	0.518691	Interaction GCSE*SAMPLE
44	MATHSAM	1291	-164	433.4	27.62537	Interaction MATHS*SAMPLE
45	PAVASAM	1291	-4.24	6.296	0.612997	Interaction PAVALEV*SAMPLE
46	SATSAM	1291	-229	689	47.54205	Interaction SATTOT*SAMPLE
47	VERBSAM	1291	-124	411.7	20.43222	Interaction VERBAL*SAMPLE
48	AGESAM	1291	-23.8	17.1	-0.00257	Interaction AGE*SAMPLE

Table 3.2: Detailed results of multilevel analysis of A-level scores

				95% Confidence interval	
Parameter	**Estimate**	**Standard error**	**Sig.**	**Min.**	**Max.**
Base case					
School variance	3.47	0.682	*	2.130	4.804
Pupil variance	8.88	0.359	*	8.174	9.582
Final model					
School variance	0.730	0.190	*	0.357	1.103
Pupil variance	6.760	0.273	*	6.224	7.296
Fixed coefficients					
CONS	1.472	3.761		-5.900	8.844
SATTOT	0.009471	0.0005329	*	0.008	0.011
AGE	-0.01476	0.01696		-0.048	0.018
SEX	0.2045	0.08419	*	0.039	0.370
MIXED	0.5898	0.6722		-0.728	1.907
ASIAN	0.3183	0.2897		-0.250	0.886
BLACK	0.3761	0.8182		-1.228	1.980
OTHCHN	0.7469	0.5243		-0.281	1.775
SESHOM	0.1525	0.3381		-0.510	0.815
SESCORP	-0.4266	0.2089	*	-0.836	-0.017
SESTECH	-0.1324	0.2428		-0.608	0.343
SESBO	-0.5344	0.3266		-1.175	0.106
SESCLK	-0.6411	0.304	*	-1.237	-0.045
SESSSW	-0.2956	0.3249		-0.932	0.341
SESAG	0.6526	1.631		-2.544	3.849
SESCTW	-0.4981	0.4814		-1.442	0.445
SESPMO	0.04376	0.6238		-1.179	1.266
SESLAB	-0.8445	0.754		-2.322	0.633
SAMPLE	1.58	0.2222	*	1.144	2.016
GCSESEX	-0.0893	0.07745		-0.241	0.063
MATHSEX	-0.004447	0.005132		-0.015	0.006
PAVASEX	0.03317	0.04938		-0.064	0.130
SATSEX	0.003179	0.004993		-0.007	0.013
VERBSEX	0.0001221	0.004942		-0.010	0.010
AGESEX	-0.007478	0.01721		-0.041	0.026
SAMSEX	-0.09228	0.1695		-0.425	0.240
GCSESAM	-0.1529	0.1353		-0.418	0.112
MATHSAM	0.003918	0.01025		-0.016	0.024
PAVASAM	-0.3542	0.08676	*	-0.524	-0.184
SATSAM	-0.003126	0.01006		-0.023	0.017
VERBSAM	0.00418	0.01003		-0.015	0.024
AGESAM	0.01309	0.02717		-0.040	0.066

Table 3.3: Detailed results of multilevel analysis of A-level scores with SES continuous

				95% Confidence interval	
Parameter	**Estimate**	**Standard error**	**Sig.**	**Min.**	**Max.**
Base case					
School variance	3.47	0.682	*	2.130	4.804
Pupil variance	8.88	0.359	*	8.174	9.582
Final model					
School variance	0.730	0.190	*	0.357	1.103
Pupil variance	6.760	0.273	*	6.224	7.296
Fixed coefficients					
CONS	1.119	3.785		-6.300	8.538
SATTOT	0.009522	0.0005346	*	0.008	0.011
AGE	-0.01515	0.01699		-0.048	0.018
SEX	0.1927	0.08416	*	0.028	0.358
MIXED	0.6813	0.6735		-0.639	2.001
ASIAN	0.4055	0.2751		-0.134	0.945
BLACK	0.4715	0.8253		-1.146	2.089
OTHCHN	0.7151	0.5238		-0.312	1.742
SAMPLE	1.586	0.2213	*	1.152	2.020
GCSESEX	-0.08338	0.07799		-0.236	0.069
MATHSEX	-0.002765	0.004997		-0.013	0.007
PAVASEX	0.03996	0.04999		-0.058	0.138
SATSEX	0.001559	0.004852		-0.008	0.011
VERBSEX	0.001626	0.004814		-0.008	0.011
AGESEX	-0.009379	0.01728		-0.043	0.024
SAMSEX	-0.1086	0.1712		-0.444	0.227
GCSESAM	-0.1721	0.1368		-0.440	0.096
MATHSAM	0.0009203	0.01012		-0.019	0.021
PAVASAM	-0.3244	0.08903	*	-0.499	-0.150
SATSAM	-0.0001032	0.009943		-0.020	0.019
VERBSAM	0.0006749	0.009896		-0.019	0.020
AGESAM	0.01001	0.0279		-0.045	0.065
GCSESES	0.0443	0.03762		-0.029	0.118
MATHSES	-0.003161	0.00227		-0.008	0.001
PAVASES	-0.01998	0.02176		-0.063	0.023
SATSES	0.002812	0.00218		-0.001	0.007
VERBSES	-0.002493	0.002178		-0.007	0.002
AGESES	-0.00392	0.008398		-0.020	0.013
SESSAM	0.005495	0.0784		-0.148	0.159
SESSEX	-0.0193	0.03928		-0.096	0.058
SES	0.0206	0.04141		-0.061	0.102

Table 3.4: Detailed results of multilevel analysis scores with math as an explanatory variable

				95% Confidence interval	
Parameter	**Estimate**	**Standard error**	**Sig.**	**Min.**	**Max.**
Base case					
School variance	3.47	0.682	*	2.130	4.804
Pupil variance	8.88	0.359	*	8.174	9.582
Final model					
School variance	0.8867	0.2242	*	0.447	1.326
Pupil variance	7.45	0.3013	*	6.859	8.041
Fixed coefficients					
CONS	4.718	3.945		-3.014	12.450
MATHS	0.01137	0.0008955	*	0.010	0.013
AGE	-0.01385	0.01784		-0.049	0.021
SEX	0.2555	0.09001	*	0.079	0.432
MIXED	0.6503	0.7063		-0.734	2.035
ASIAN	-0.1921	0.3039		-0.788	0.404
BLACK	-0.07895	0.8586		-1.762	1.604
OTHCHN	0.3903	0.5516		-0.691	1.471
SESHOM	0.02578	0.3551		-0.670	0.722
SESCORP	-0.4119	0.2194		-0.842	0.018
SESTECH	-0.2782	0.255		-0.778	0.222
SESBO	-0.6352	0.3431		-1.308	0.037
SESCLK	-0.8041	0.3191	*	-1.430	-0.179
SESSSW	-0.4706	0.341		-1.139	0.198
SESAG	0.4499	1.714		-2.910	3.809
SESCTW	-0.7521	0.5052		-1.742	0.238
SESPMO	0.004837	0.6556		-1.280	1.290
SESLAB	-1.037	0.792		-2.589	0.515
SAMPLE	1.922	0.2383	*	1.455	2.389
GCSESEX	-0.0789	0.08138		-0.238	0.081
MATHSEX	-0.003777	0.005399		-0.014	0.007
PAVASEX	0.0502	0.05193		-0.052	0.152
SATSEX	0.002646	0.005249		-0.008	0.013
VERBSEX	0.0002742	0.005195		-0.010	0.010
AGESEX	-0.009943	0.01809		-0.045	0.026
SAMSEX	-0.1678	0.1793		-0.519	0.184
GCSESAM	-0.2159	0.1424		-0.495	0.063
MATHSAM	0.007238	0.01079		-0.014	0.028
PAVASAM	-0.3392	0.0913	*	-0.518	-0.160
SATSAM	-0.006119	0.0106		-0.027	0.015
VERBSAM	0.006926	0.01056		-0.014	0.028
AGESAM	0.02614	0.02854		-0.030	0.082

Table 3.5: Detailed results of multilevel analysis of A-level scores with verbal as an explanatory variable

A-level		Multilevel results			
				95% Confidence interval	
Parameter	**Estimate**	**Standard error**	**Sig.**	**Min.**	**Max.**
Base case					
School variance	3.47	0.682	*	2.130	4.804
Pupil variance	8.88	0.359	*	8.174	9.582
Final model					
School variance	0.7598	0.1975	*	0.373	1.147
Pupil variance	6.959	0.2814	*	6.407	7.511
Fixed coefficients					
CONS	3.793	3.804		-3.663	11.249
VERBAL	0.01555	0.0009475	*	0.014	0.017
AGE	-0.01955	0.01722		-0.053	0.014
SEX	0.002915	0.08485		-0.163	0.169
MIXED	0.312	0.682		-1.025	1.649
ASIAN	0.5877	0.2966	*	0.006	1.169
BLACK	0.5138	0.8309		-1.115	2.142
OTHCHN	1.223	0.5329	*	0.179	2.267
SESHOM	0.1485	0.3431		-0.524	0.821
SESCORP	-0.4064	0.2119		-0.822	0.009
SESTECH	0.00125	0.2468		-0.482	0.485
SESBO	-0.5796	0.3313		-1.229	0.070
SESCLK	-0.629	0.3086	*	-1.234	-0.024
SESSSW	-0.2472	0.3299		-0.894	0.399
SESAG	1.681	1.658		-1.569	4.931
SESCTW	-0.5583	0.4884		-1.516	0.399
SESPMO	-0.08465	0.6329		-1.325	1.156
SESLAB	-0.8776	0.7651		-2.377	0.622
SAMPLE	1.871	0.2228	*	1.434	2.308
GCSESEX	-0.05916	0.07855		-0.213	0.095
MATHSEX	-0.009441	0.005206		-0.020	0.001
PAVASEX	0.002705	0.05013		-0.096	0.101
SATSEX	0.006745	0.005068		-0.003	0.017
VERBSEX	-0.003058	0.005016		-0.013	0.007
AGESEX	0.002726	0.01746		-0.031	0.037
SAMSEX	0.03971	0.1722		-0.298	0.377
GCSESAM	-0.2606	0.1368		-0.529	0.008
MATHSAM	0.003087	0.0104		-0.017	0.023
PAVASAM	-0.3141	0.08798	*	-0.487	-0.142
SATSAM	-0.001759	0.01022		-0.022	0.018
VERBSAM	0.00323	0.01018		-0.017	0.023
AGESAM	0.004016	0.02759		-0.050	0.058

Appendix 4: Item functioning data (IRT analysis)

The following tables and figures present the comparison of the SAT functioning between British and American students. The data for the American students was calculated by Educational Testing Services.

Table 4.1: IRT statistics for verbal section

Slopes ('Discrimination')		Thresholds ('Difficulty')		Asymptotes ('Guessing')	
UK	USA	UK	USA	UK	USA
1.16	1.09	-2.35	-1.43	0.17	0.10
1.00	0.56	-2.71	-2.07	0.18	0.13
1.24	1.04	-0.73	-0.30	0.19	0.32
1.23	1.02	-1.02	-0.48	0.17	0.27
1.30	1.12	-0.28	-0.48	0.19	0.19
1.46	1.10	-1.11	-0.44	0.15	0.13
1.50	1.15	-0.90	0.20	0.15	0.27
1.38	1.19	0.55	1.63	0.19	0.14
2.36	1.61	1.24	1.42	0.10	0.09
2.42	1.43	1.41	1.62	0.06	0.10
1.12	0.45	-1.36	-3.59	0.16	0.13
1.18	0.87	-0.87	-0.43	0.16	0.36
1.21	0.58	-0.60	-1.58	0.15	0.13
1.06	0.54	0.55	-0.11	0.13	0.15
0.71	0.61	-0.05	0.12	0.15	0.21
1.81	1.00	0.68	0.65	0.13	0.18
0.94	0.77	0.79	0.27	0.20	0.19
1.48	0.38	1.86	0.58	0.13	0.04
0.71	0.76	2.60	1.03	0.16	0.16
1.47	0.85	1.28	0.84	0.09	0.13
1.10	0.69	1.59	1.34	0.18	0.13
0.71	0.61	2.93	1.95	0.12	0.15
1.56	0.84	1.91	1.18	0.21	0.20
0.97	0.65	-0.20	-0.56	0.12	0.08
1.58	1.20	-0.14	0.25	0.14	0.24
1.35	0.39	0.37	-0.17	0.13	0.13
1.46	0.59	-1.56	-2.48	0.17	0.13
0.50	0.34	1.22	-0.77	0.12	0.13
1.47	0.78	-1.01	-1.32	0.18	0.30
1.07	0.67	0.34	0.54	0.11	0.16
1.81	0.88	0.55	0.55	0.07	0.09
0.95	0.36	1.02	0.83	0.13	0.04
1.35	0.93	-0.37	-0.82	0.11	0.18
1.03	0.61	0.95	0.96	0.10	0.19
0.97	0.53	0.26	-0.75	0.13	0.13

Table 4.2: IRT statistics for math section

Slopes ('Discrimination')		Thresholds ('Difficulty')		Asymptotes ('Guessing')	
UK	USA	UK	USA	UK	USA
2.05	0.65	-1.32	-3.40	0.09	0.11
1.50	0.62	-1.69	-2.63	0.09	0.11
1.35	0.47	-1.05	-2.37	0.08	0.11
0.87	0.81	-0.34	0.14	0.10	0.50
1.66	0.78	-0.95	-0.65	0.08	0.27
2.33	1.08	-0.81	-1.68	0.09	0.05
1.94	0.75	-0.50	-1.13	0.06	0.07
1.38	1.09	-0.04	0.75	0.09	0.36
1.44	0.97	-0.73	-0.36	0.07	0.13
2.62	1.49	-0.04	-0.17	0.05	0.12
1.78	1.01	-0.58	-0.44	0.07	0.21
1.44	0.89	0.29	0.49	0.07	0.24
1.56	1.04	-0.71	-0.31	0.08	0.22
2.26	1.33	0.05	0.11	0.09	0.19
1.69	1.13	0.51	0.47	0.12	0.23
2.02	1.07	0.83	0.47	0.09	0.09
1.01	0.89	1.56	1.64	0.09	0.19
2.22	1.75	1.15	1.17	0.15	0.22
2.21	1.44	1.24	1.14	0.14	0.22
2.18	1.23	1.85	1.56	0.09	0.07
1.58	0.96	0.08	-0.52	0.05	0.13
1.67	1.03	-0.59	-0.83	0.08	0.28
1.33	0.44	-0.33	-1.66	0.05	0.05
0.97	0.56	0.07	-1.41	0.07	0.05
1.08	0.88	0.63	0.32	0.08	0.35
1.99	1.08	-0.22	-0.08	0.08	0.45
2.33	0.77	1.58	1.10	0.18	0.39
2.49	1.75	1.88	1.67	0.16	0.17
1.84	1.27	-0.2	-1.10	0.05	0.00
1.64	0.60	0.24	-1.31	0.04	0.00
1.84	0.65	1.10	0.70	0.03	0.00
2.28	1.18	1.25	0.69	0.01	0.00
2.95	1.09	1.65	1.37	0.01	0.00

Note: The data supplied by ETS may have set the asymptote for the last five math items (student-produced response items) to zero, whereas this was allowed to vary in the British data. This could have had the effect of slightly reducing the association between the math asymptote between British and American students.

Figure 4.1: Scatterplot of verbal IRT difficulties for British and American students

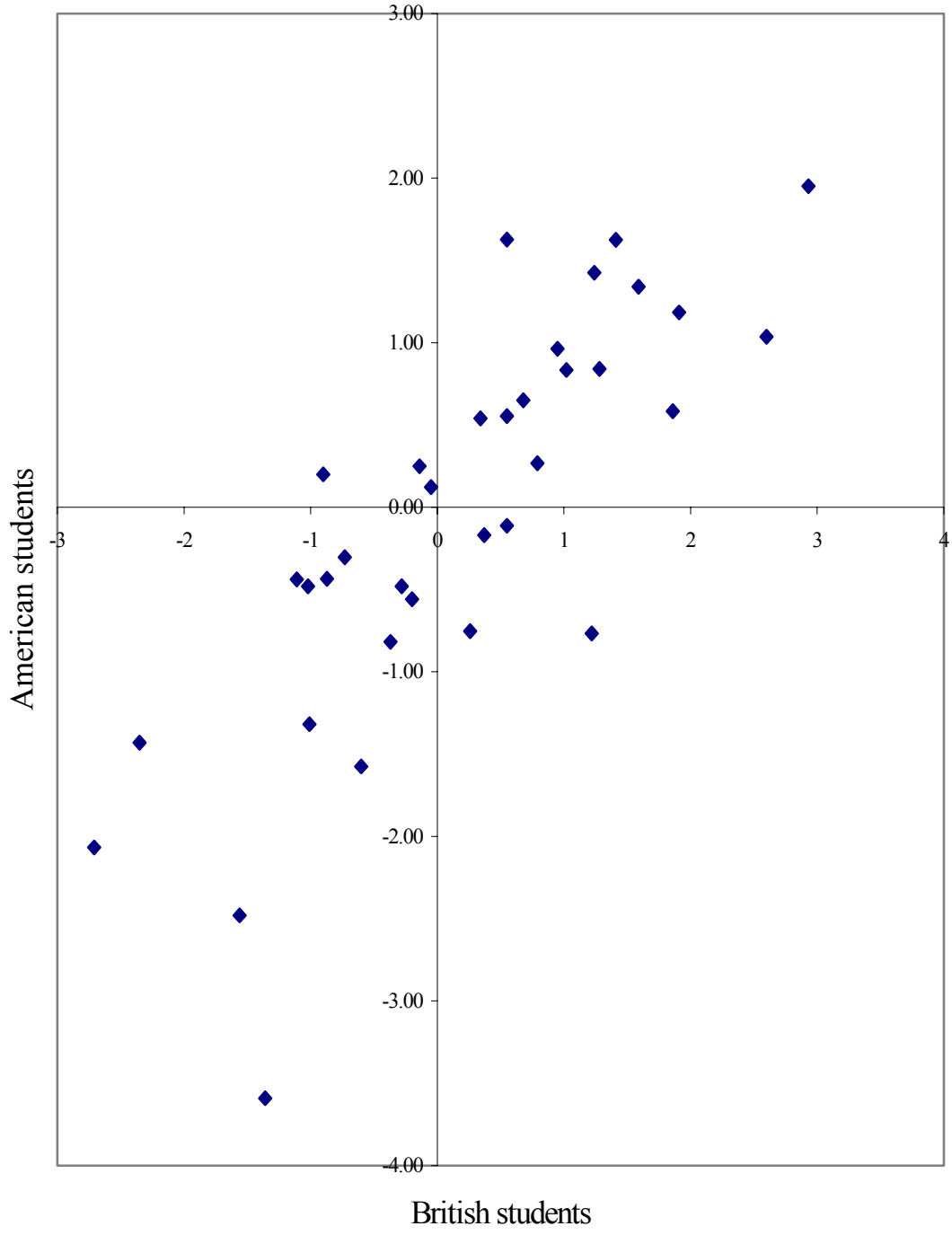
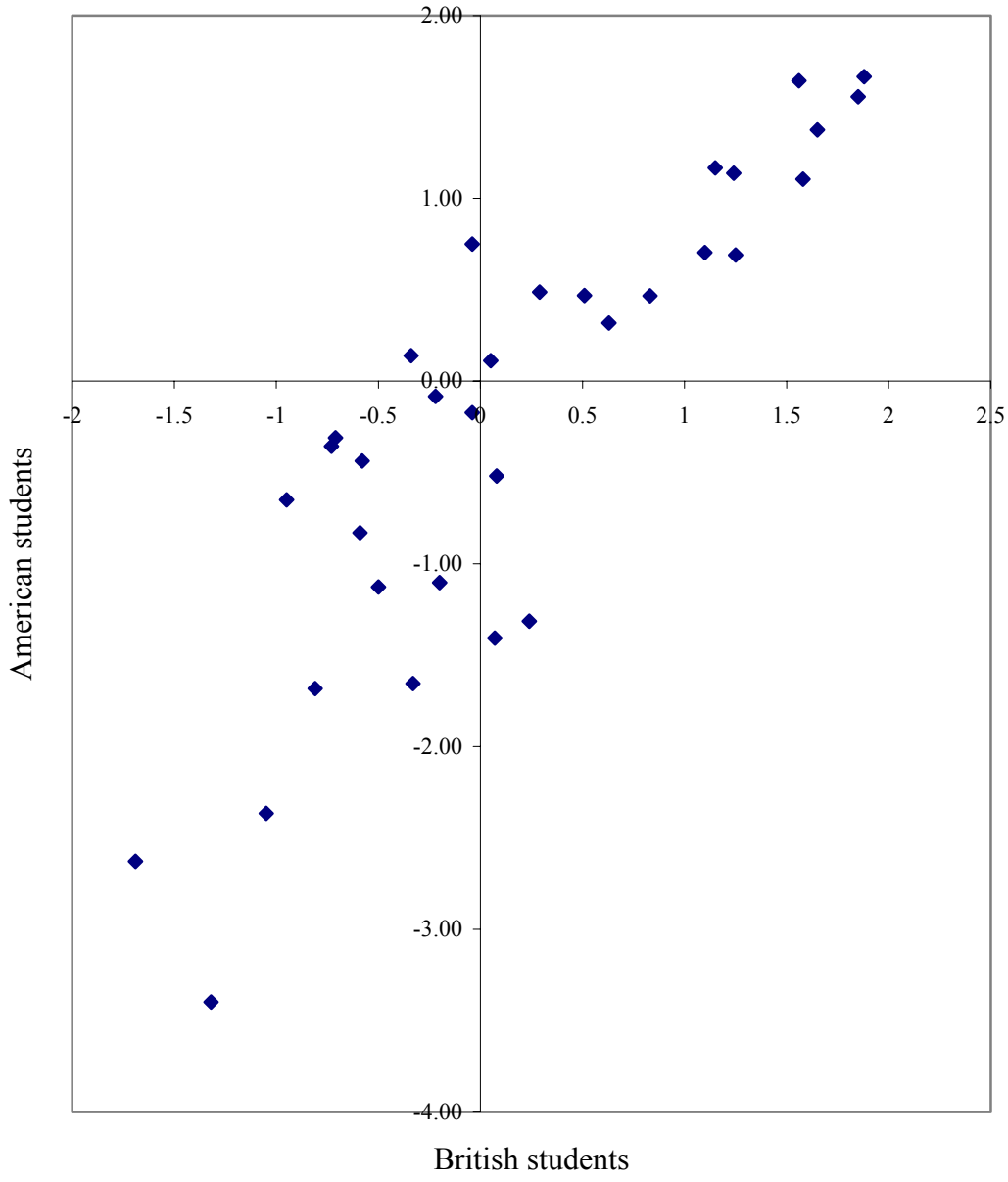


Figure 4.2: Scatterplot of IRT math difficulties for British and American students



Appendix 5: Item functioning data (classical test analysis)

Table 5.1: Item analysis of verbal section

Item	Facility (% correct)	Discrimination	% Omitted	% Not reached
1	91.9	0.27	0.9	0.7
2	92.7	0.22	1.4	0.7
3	73.2	0.36	1.6	0.7
4	77.2	0.35	1.6	0.7
5	64.8	0.38	1.8	0.7
6	80.0	0.39	2.8	0.7
7	76.5	0.41	1.3	0.7
8	48.3	0.35	2.2	0.7
9	23.8	0.38	3.3	0.7
10	17.5	0.37	2.8	0.7
11	80.9	0.33	0.9	0.7
12	74.2	0.36	2.2	0.7
13	69.3	0.38	2.4	0.7
14	46.3	0.34	1.7	0.7
15	58.3	0.26	2.6	0.8
16	39.9	0.42	6.3	0.9
17	47.9	0.27	8.7	0.9
18	22.9	0.23	1.9	0.9
19	29.6	0.15	8.5	0.9
20	27.67	0.35	6.2	1.0
21	33.4	0.24	6.9	1.0
22	23.2	0.16	6.8	1.2
23	28.9	0.20	12.0	1.5
24	59.7	0.33	5.1	2.1
25	60.1	0.44	3.9	2.4
26	48.6	0.39	5.1	2.8
27	87.1	0.34	3.1	3.0
28	43.4	0.19	5.3	3.2
29	79.2	0.38	3.9	3.6
30	48.8	0.35	5.7	4.7
31	39.1	0.47	8.4	5.8
32	39.9	0.29	8.8	6.8
33	63.7	0.42	8.6	7.3
34	38.0	0.32	14.0	9.4
35	52.1	0.33	10.6	10.6

Table 5.2: Item analysis of math section

Item	Facility (% correct)	Discrimination	% Omitted	% Not reached
1	85.9	0.46	4.0	3.7
2	87.6	0.39	4.2	3.8
3	76.6	0.42	6.4	3.9
4	60.7	0.33	4.5	3.9
5	76.8	0.46	7.6	3.9
6	77.4	0.52	6.7	4.0
7	67.3	0.52	6.5	4.2
8	55.8	0.45	10.0	4.2
9	70.7	0.45	4.9	4.3
10	54.1	0.60	7.1	4.3
11	69.0	0.51	7.3	4.3
12	46.5	0.45	11.3	4.8
13	71.4	0.47	8.2	5.0
14	53.2	0.56	12.0	5.3
15	43.4	0.45	18.8	5.4
16	32.4	0.46	8.7	5.4
17	28.1	0.28	6.5	5.5
18	30.7	0.37	16.7	6.2
19	28.0	0.36	22.6	6.4
20	15.9	0.26	15.0	6.5
21	50.9	0.50	8.8	7.6
22	69.3	0.49	9.1	7.9
23	60.4	0.46	10.5	8.3
24	52.0	0.37	10.1	8.8
25	41.8	0.38	13.7	9.6
26	60.5	0.54	11.7	9.8
27	26.3	0.26	22.5	11.1
28	20.7	0.20	23.7	11.8
29	58.4	0.53	27.5	20.9
30	46.0	0.50	27.4	23.7
31	23.3	0.46	40.2	30.8
32	16.5	0.47	42.0	39.1
33	8.3	0.40	57.2	57.2