Fall 5-1-2013

Does the SAT predict academic achievement and academic choices at Macalester College?

Jing Wen
Macalester College, jwen@macalester.edu

Does the SAT predict academic achievement and academic choices at Macalester College?

Honors Project

Mathematics, Statistics and Computer Science

Macalester College

Student's Name: Jing Wen

Advisor: Professor Vittorio Addona

May, 2013

Abstract

This paper examines the predictive power of the Scholastic Aptitude Test (SAT) for Macalester students' college success and academic choices. We use linear regression to study whether the SAT can predict students' first year or four-year grades. Using Kullback-Leibler divergence and classification trees, we also examine the SAT's predictive ability for other aspects of students' academic experience, for example, major selection, or academic division of study. After controlling for major and course level, we find that the SAT does not explain a large proportion of the variability in Macalester students' college success. However, the SAT does provide some useful information in predicting students' major choice or academic division of study.

CONTENTS

ACKNOWLEDGEMENTS

Thank you to my readers Daniel Kaplan, and David Chioni Moore, and to my advisor Vittorio Addona.

1. Introduction

The Scholastic Assessment Test (SAT) is a widely used standardized test for undergraduate admissions (Kobrin et al., 2008). The test's primary purpose is to successfully measure a student's reasoning ability and educational achievement related to academic performance in college (Kobrin and Michel, 2006). The College Board is responsible for designing the SAT test. In the Standards for Educational and Psychological Testing (AERA/APA/NCME, 1999), the test maker is "responsible for furnishing relevant evidence and a rationale in support of the intended test use" (p.11). Prior to 2005, the SAT was consisted of verbal and mathematical reasoning sections. The verbal section included both long and short passages and the mathematics section covered three years of high school mathematics. The students were allowed three hours to complete the test. In March 2005, the College Board revised the SAT format and incorporated many changes. The verbal section was renamed as the critical reading section, and more questions on long reading passages replaced the analogies questions (Kobrin et al., 2008). The inclusion of a writing section was the most notable change of the new SAT. The current SAT is 3 hours and 45 minutes in length, which gives takers more time to complete the full test.

According to the Standards for Educational and Psychological Testing (AERA/APA/NCME, 1999), "a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (p.17). One of the most common criticism for admissions tests such as the SAT is their predictive validity. Since many schools use the SAT as a required test for admission, for decades many researchers have discussed whether the SAT is a good predictor of college performance.

This paper examines the validity of the SAT for predicting student academic performance at Macalester College. We also investigate what ability, if any, the SAT possesses for predicting the academic choices of Macalester students. As is done in most previous literature, we use linear regression as part of our statistical analysis to study whether the SAT can predict students' first year, or four-year, grades. We also examine the SAT's predictive power for other aspects of students' academic experience, for example, major selection, or academic division of study.

In this paper, we find that the SAT does not explain a large proportion of the variability in Macalester students' college success. However, the SAT does contribute to predicting students' major choice or academic division. These findings are subject to limitations due to the restricted dataset available for analysis.

The purpose of the present study is to examine the SAT's predictive power for students' academic performance and academic choices at Macalester. This paper will proceed as follows: Section II reviews previous literature that has illustrated the SATs' predictive power. Section III introduces the Macalester College data used and describes why it consists of an adequate, even if not ideal, sample for our study. Section IV describes models estimated on this sample concerning the predictive power of the SATs and presents a variety of different results, along with interpretations of these results. Section V concludes and provides directions for future research.

2. Literature Review

In 1926, the College Board began the process of designing the Scholastic Aptitude Test (SAT) in order to predict the academic performance of secondary school students as they entered college (Donlon, 1984). In the 90 years since then, the College Board has claimed that the SATs have remained an effective predictor of students' college performance (The College Board, 2011). The format of the Scholastic Assessment Test I has changed several times in its history, but the SATs' predictive power of applicants' grades has been justified by numerous studies (e.g. Leonard & Jiang, 1999; Striker, 1991; Willingham et al., 1990). More recently, however, several prestigious colleges have de-emphasized the role of the SATs in their admission procedure. For instance, since 2001, University of California (UC) have implemented the "Four Percent Plan" — students ranking at, or near, the top of their high school classes are admitted without consideration of SAT scores (Rothstein, 2003). In addition, a few more well-known small colleges, such as Middlebury and Bennington also enacted an admissions rule whereby completing the SATs was optional (Rothstein, 2003).

Studies on the predictive validity of SAT scores for college performance have been conducted over several decades. Rothstein (2002) measured the SAT's predictive power considering demographic variables omitted by previous literature: The fraction of students who are Black, Hispanic, and Asian; the fraction of students receiving subsidized lunches; and the average education of students' parents. He concluded that the SATs have less predictive power than previously believed. Most previous studies used students' first year GPA, which were only available for students enrolled at a single college, and based on highly selected samples. As a result, these students probably did not serve as a representative group and resulted in models with biased SAT contribution estimates (Rothstein, 2002). However, Rothstein used data which contained academic information on all California residents from the 1993 high school graduating class who enrolled at any of the eight University of California campuses. This approach differed from the restrictive samples employed in previous literature. Accepted students tend to have higher SAT scores and may be more likely to achieve a higher GPA. However, the students with relatively lower SAT scores are also less likely to get accepted since they are deemed less likely to succeed in college. This "restriction of range" may lead to a lower correlation between the SAT scores and grades. Rothstein (2002), on the other hand, implemented restriction of range corrections by using an algorithm derived from regression-omitted variables results to resolve this problem (Camara and Echternacht, 2000; Willingham et al., 1990).

Rothstein (2002) also studied the effect of school-level demographic variables which might serve as confounders, and the potential endogenous selection into campuses related to geographic instruments. According to a linear regression model with students' high school GPAs, SAT scores, official majors, and demographic characteristics of high schools as the independent variables, and students' first year GPA as the dependent variable, the results showed that the SAT Verbal and Math scores have reliabilities approximately 75% less what College Board indicated, which was about 0.9 (College Board, 2001). The aforementioned "Four Percent Plan" he proposed was adopted by the University of California, Berkeley.

Baron and Norman (1992) also examined whether the SAT is a significant factor in predicting students' first year college performance. Their data came from 4170 students from four departments enrolled at the University of Pennsylvania. By implementing linear combinations of high-school class rank (CLR), total shcolastic-apitude-test (SAT) and average achievement-test score (ACH) as the independent variables, Baron and Norman (1992) stated that when CLR and ACH were available, the R-squared value was 0.136, but the R-squared value remained essentially constant when the SAT was added to the model. Thus they concluded that while CLR and ACH contributed significantly to prediction of students' first year college performance, SAT merely added a relatively small contribution.

However, the majority of SAT predictive validity studies confirmed the SAT's significant contribution to predicting students' college success defined by first year grades. Authors typically used first year, instead of four-year, GPA because of proximity to the achieved SAT score. Wilson (1983) implemented linear regressions and found that either the SAT scores alone or high school GPAs was a good predictor of students' first year grades. However, the combination of SAT scores and high school grades could give better predictions than either one alone. Morgan (1989) analyzed the predictive power of the SAT and high school grades using data from the College Board's Validity Study Service by studying the dynamic pattern of the correlation between students' first year grades and SAT scores. The data contained more than one million students from 1976 to 1985. After correcting for restriction of range, he concluded that the correlation between first year grades and SAT scores decreased from 0.51 to 0.47 over the decade, but the correlation was still statistically significant.

Among the recent studies examining the validity of SAT scores for predicting college performance, most scholars, including Geiser & Studley (2002), and Agronow & Studley (2007), used SAT scores with high school grades and demographic characteristics of the students to predict first year GPA. They all concluded that the SAT was a significant coefficient in the linear regression models and a good predictor of students' first year GPA. Using data on 34,000 students who entered 30 colleges in 1995, a multiyear validity study by Kobrin and Michel (2006) studied whether the SAT had more predictive power for students with high first year GPAs compared to lower first year GPAs. The study implemented a logistic regression model to approach the probability that a student would be correctly classified as successful or unsuccessful in their first year of college based on the student's high school grade point average and SAT scores. The results showed that the SAT was able to predict high levels of college success, even better than high school grades (Kobrin and Michel, 2006).

Kobrin et al. (2008) used data on students from 726 four-year institutions that got more than 200 SAT scores in 2005. To approach the increment in prediction of students' first year grades by SAT scores over high school grades, they calculated the difference between the correlation of high school grades alone with first year grades and the multiple correlation based on the SAT scores and high school grades (Kobrin et al., 2008). They found that the increment in the SAT's predictive validity over high school grades was 0.10, which was statistically significant.

Although the main focus of this section is the SAT's predictive power for students' college performance, the predictive validity of a similar standardized test for high school achievement – American College Testing (ACT), is also worth studying, since the ACT is used predominantly in the Midwest. Moreover, comparing the two tests' predictive validity is useful to provide recommendations to Macalester concerning the tests' respective value at the time of admission.

With regards to the ACT's predictive ability, Noble (1991) examined the prediction of students' overall freshman GPA based on ACT test scores along with high school grades. The results showed that although neither ACT scores nor high school grades predicted students' first year grades well, the joint ACT scores and grades produced a good prediction (Noble, 1991). Noble and Sawyer (2002) did a similar study by pooling samples across ACT's Prediction Research history files. (ACT, 1997b; 1998c) They used logistic regression models and incorporated high school course grades and ACT test scores to predict the first-year college GPAs. Their results revealed that ACT Composite scores were effective at predicting students' all first-year GPA levels and the accuracy rate ranged from 0.78 to 0.93 for all first-year GPA levels.

3. Methods

We obtained the dataset used in the subsequent analysis from Macalester College's Registrar's Office. The data contains information on 478 students who enrolled at Macalester in Fall 2008 and were expected to graduate in May, 2012. There are 16 variables in total, including student ID number, gender, grade for each course, course number and corresponding department, course credits and year taken, race, range of financial aid, high school class rank, SAT score for each section of the exam, ACT score, a faculty ID for each course, and graduation date for each student. In this dataset, each section of the SAT (Math, Verbal, and Writing) has a score ranging from 200 to 800; ACT scores range from 25 to 30; race is divided into 7 categories: NR=non-resident alien, HI = Hispanic of any race AS = Asian, not Hispanic, BL = Black, not Hispanic, IS = Pacific Islander, not Hispanic, WH = White, not Hispanic, MR = Two or more races, not

Hispanic. We converted students' letter grades to point values according to Macalester grading policy[1]. We do not consider the grades of courses taken Pass/Fail by students, since such grades are not counted in GPA calculations. There are 33 departments and 4 course levels (100, 200, 300, and 400) in the dataset, and we use these two variables to adjust students' grades; the adjustment process is explained in section V.

The target population of our study includes all applicants to Macalester College, so our ideal data should be representative of this target population. The ideal dataset is similar to our actual dataset but with several exceptions. First, the goal of our study is to measure the SAT's validity for predicting student academic performance for the entire applicant pool of Macalester College. Unfortunately, we were only able to obtain detailed information for students who were admitted, and enrolled at, Macalester College. This leads to an underestimation of the true predictive power of the SAT, and this phenomenon is called "restriction of range" (Rothstein, 2006). The range is restricted in that the range of scores is smaller for admitted students than the applicant pool, so analyzing only enrolled students restricts the SAT's variation and its predictive validity (Kobrin et al., 2008). This problem exists in our dataset since we do not have the full data for the entire applicant pool. In addition, an ideal dataset would include students from different graduating classes at Macalester College to make the sample more representative of the target population. However, we only had access to information for a single graduating class. Therefore, the conclusions stemming from this analysis must be interpreted with some care.

The summary statistics for students' cumulative first year GPA (FYGPA), four

[1]Please refer to the website: http://www.macalester.edu/academic/catalog/ap4.html for more information on the grading policy.

year GPA (FOURYGPA), SAT scores and ACT scores are given from Table 1 to Table 4 below.

Table 1: Summary statistics

	25% quantile	75% quantile	Mean	SD	Observations
FYGPA	3.29	3.73	3.46	0.38	478
FOURYGPA	3.21	3.68	3.42	0.50	478
SAT Reading	640	740	677	78.22	337
SAT Math	630	710	671	67.16	337
SAT Writing	670	730	670	78.00	337
ACT	28	32	30	2.85	267

Table 2: GPA by Category

Category	Observations	FYGPA	GPA SD

Total	478	3.455	0.379
Citizen Country			
Non-resident Alien	48	3.465	0.350
Hispanic	29	3.274	0.297
Native American	4	3.425	0.269
Asian	35	3.364	0.386
Black	8	2.863	0.539
White	343	3.491	0.366
Two or more races	11	3.5	0.221
Gender			
Female	279	3.523	0.327
Male	199	3.36	0.424
Financial Aid			
No Financial Aid	134	3.468	0.357
Less than 15000	67	3.554	0.332
15000 – 24999	59	3.479	0.437
25000 – 34999	78	3.421	0.320
35000 – 44999	140	3.403	0.416
First Year Transfer			
Transfer	21	2.994	0.734
Not Transfer	457	3.475	0.343

Table 3: SAT scores by Category

Category	Observations	Mean Math	Mean Verbal	MRSAT[2] SD
Total	337	671	677	185
Citizen Country				
Non-resident Alien	47	668	585	231
Hispanic	23	595	615	193
Native American	1	800	660	N/A
Asian	21	682	685	176
Black	5	614	600	153
White	231	681	702	135
Two or more races	9	618	688	124
Gender				
Female	192	663	678	179
Male	145	681	675	194
Financial Aid				
No Financial Aid	95	676	685	160
Less than 15000	51	674	699	161
15000 – 24999	44	681	708	159
25000 – 34999	52	691	687	140
35000 – 44999	95	648	636	228
First Year Transfer				
Transfer	16	659	691	162
Not Transfer	321	671	676	187

Table 4: ACT scores by category

Category	Observations	Mean	ACT SD

[2]MRSAT is the sum of students' Math and Verbal scores.

Total	267	30	2.85
Citizen Country			
Non-resident Alien	1	29	N/A
Hispanic	15	27	4.15
Native American	4	31	3.11
Asian	18	27	3.16
Black	3	24	3.46
White	219	30	2.34
Two or more races	7	28	2.06
Gender			
Female	163	30	2.82
Male	104	30	2.89
Financial Aid			
No Financial Aid	70	30	2.18
Less than 15000	42	30	2.63
15000 – 24999	31	31	2.60
25000 – 34999	43	30	2.76
35000 – 44999	81	29	3.41
First Year Transfer			
Transfer	11	31	3.75
Not Transfer	256	30	2.81

4. Analyses (Models and Results)

This section contains two subsections. Section 4.1 investigates whether the SAT or ACT can be a good predictor of Macalester students' college success defined by cumulative first year, and four-year, grades. Section 4.2 examines the SAT's capability to predict other aspects of Macalester students' academic experience – major selection and academic division of study.

4.1. The SAT's and ACT's predictive power for students' first year, and four-year, GPA

In order to investigate the SAT's predictive power for students' college success, we use the students' college GPA as one reasonable surrogate for their academic success. There are, of course, many other facets to college success, such as leadership, volunteer experience, publications and honors, which are not considered in the scope of this paper.

The basic model used in much of the literature for looking at the predictive power of the SAT is:

$$Y_i = \beta_0 + \beta_1 SAT_i + \beta_2 HSGPA_i + \beta_3 X_i + error_i \qquad (1)$$

In equation (1), the response variable Y_i is the college GPA for student i. SAT_i stands for the SAT score of a student, and $HSGPA_i$ represents a student's high school GPA, a proxy for high school academic success. X_i is a vector of a student's non-academic characteristics, including gender, ethnicity, and graduation class. The previous literature indicates that the strong predictive ability of the SAT is observed when mistakenly not considering other factors which contribute to college success and are also correlated with high SAT scores. This means that although the SAT is a strong explanatory variable of college success itself, the predictive power is not necessarily maintained when controlling for other variables such as high school performance and other indicators.

This section begins with the most basic model, using only SAT score as an explanatory variable for a student's college GPA at Macalester College:

$$Y_i = \beta_0 + \beta_1 SAT_i \qquad (2)$$

If the SAT serves as a significant predictor of students' college grades, in equation (2), then we could build a model like the one shown in equation (1) to consider additional independent variables. However, if the SAT is not a significant predictor in this simplified model, it means that even without considering other factors, the SAT hardly contributes to our ability to predict students' academic success. In such a case, we could conclude that the SAT is an insignificant predictor of Macalester students' GPA.

The basic model to approach the ACT's predictive validity for students' grades is the same as equation (2), except replacing the SAT variable with ACT Composite scores.

4.1.1 Adjustment of students' GPA

Students choose different courses in college based on personal interests and college requirements. The grades in different departments, and at different levels, are not comparable. It is normal that some departments' grading criteria are stricter than others, and the grades in lower level courses may be lower than in upper level courses. These observations lead to the possibility that an A in one department may not be considered as equivalent to an A in another department. Similarly, an A in a 100-level course does necessarily mean the same thing as an A in a 400-level course. We thus find it necessary to adjust the students' GPA by taking into account the department and the course level of each class which a student has taken.

To verify that adjusting students' GPA by departments and courses levels is necessary, we present two histograms and one table shown below. Figure 4.1.1a shows the histogram of students' GPA mean by department; Figure 4.1.1b shows the histogram of students' GPA standard deviation by department. Table 4.1.1 presents the students' GPA mean and standard deviation by course level. The plots and table show that generally, the standard deviation and mean of grades do vary across departments and/or course levels.

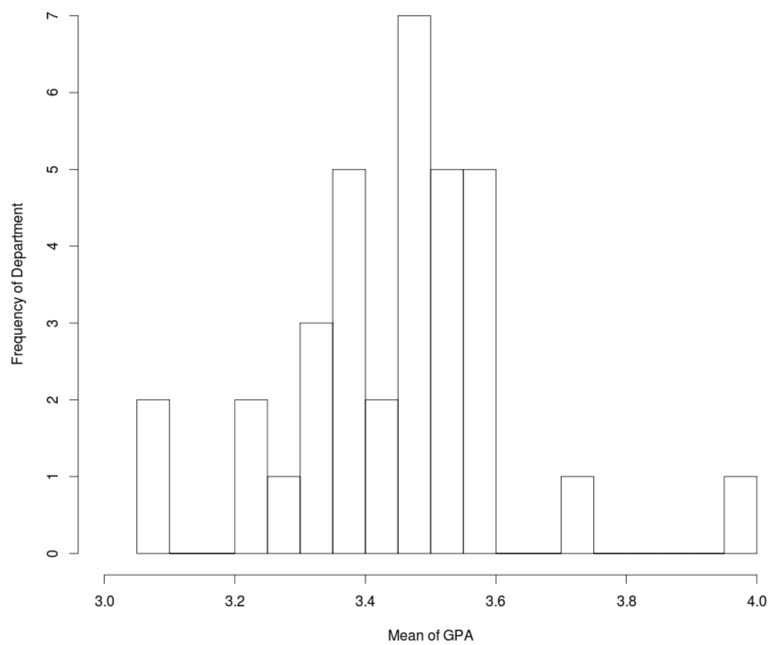


Figure 4.1.1a: Histogram of students' GPA mean by department

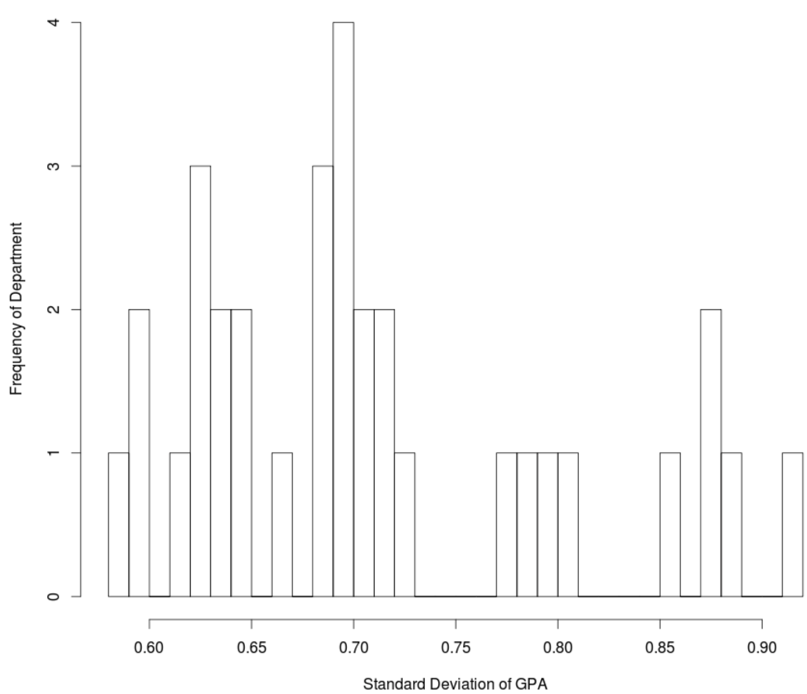


Figure 4.1.1b: Histogram of students' GPA mean by department

	Mean	SD
below 100	3.64	0.097
100	3.36	0.75
200	3.37	0.72
300	3.44	0.73
400	3.55	0.68
600	3.82	0.61

Table 4.1.1: Students' GPA mean and standard deviation by course level

The technique which we employ for adjusting student grades is based on the equation:

$$Points_i = \beta_1\ StudentID_i + \beta_2\ Department_i + \beta_3\ CourseLevel_i + error_i \quad (3)$$

$Points_i$ represents student i's grade for each course taken. $StudentID_i$ is the identification number for a student. $Department_i$ and $CourseLevel_i$ stand for the corresponding academic department and level for each course, respectively.

By controlling for both a course's level and for the department in which the course was taken, each student (through their StudentID) is given a model coefficient, which is that student's isolated impact on grades.

4.1.2 The SAT's predictive power for students' four-year GPA

In this section, by implementing the adjustment strategy described in Section 4.1.1, we use students' four-year GPA, adjusted or unadjusted, as the independent variable and the sum of SAT Verbal scores and SAT Mathematical scores as the only independent variable (MRSAT)[3]. The models are both built based on equation (2).

The regression output is presented in Table 5 on page 26. For adjusted and unadjusted GPA, although the p-values of the sum of SAT Verbal scores and SAT Mathematical scores are significant in both cases at the 5% significance level, the values of R-squared are extremely small: 0.003 for adjusted GPA and 0.0013 for unadjusted GPA. We conclude that the sum of SAT Verbal scores and SAT Mathematical scores does not predict students' four-year adjusted or unadjusted GPA very effectively.

Limitations in our data and adjustment technique may partially explain the low predictive ability of SAT scores for students' four-year grades:

1. The dataset only contains the information of the students who applied, were admitted, and enrolled in Macalester College in 2008. This may contain bias since the restriction of data makes it not representative of all applicants. We are not able to observe the students who applied but were not accepted into Macalester College, nor the students who were admitted but did not enrolled. This "restriction of range" may lead to a lower correlation between the SAT scores and grades. Although we do not correct for this selection bias, we compare the range and standard deviation of the enrolled students' SAT and ACT scores with those of the entire 2008 applicant pool.
2. Among our sample, 141 out of 478 students did not take the SAT exam but rather only took the ACT exam. These missing values may not be missing at random. To deal with this issue, we converted the ACT scores to the SAT scores for the students with no SAT scores and rerun the models.

[3]The writing scores are very subjective with no standard grading rule so it is not taken into consideration.

We obtained group level data from Macalester's Institutional Research Department for the 2008 applicant pool. The summary statistics of our dataset and the applicant pool is shown respectively in Table 4.1.2d and Table 4.1.2e. From these two tables, we see that, although the test scores' standard deviation and range in our dataset are smaller than those in the whole applicant pool, the difference seems fairly small. This indicates that the "restriction of range" problem may not be playing a major role in our dataset.

Table 4.1.2d: Summary statistics of test scores in our dataset

	Mean	Minimum	Maximum	SD
SAT Reading	677	380	800	78.22
SAT Math	671	410	800	67.16
SAT Writing	670	400	800	78.00
ACT	29.9	20	36	2.85

Table 4.1.2e: Summary statistics of test scores in applicant pool

	Mean	Minimum	Maximum	SD
SAT Reading	665	240	800	89.79
SAT Math	662	310	800	81.49
SAT Writing	662	260	800	84.53
ACT	29	12	36	3.71

To attempt to address concern 2., we impute the ACT scores using the sum of Verbal SAT scores and Mathematical SAT scores to obtain a complete set of SAT scores for all students in our data. We accomplish the imputation by running a linear regression between the SAT and the ACT, which is shown in equation (4). Although this is not a perfect procedure, the R-squared value of model (4) is 0.64, indicating strong correlation between ACT and SAT scores.

$$\text{ImputedSAT}_i = \beta_0 + \beta_1 \text{ACT}_i \tag{4}$$

We run equation (5) by using the new, complete, set of SAT scores. By implementing the adjustment strategy described in Section 4.1.1, we use students' unadjusted or adjusted four-year GPA as the independent variable and the imputed sum of SAT Verbal scores and SAT Mathematical scores, ImputedSAT_i, as the only independent variable.

$$Y_i = \beta_0 + \beta_1 \text{ ImputedSAT}_i \tag{5}$$

The regression output is shown in Table 5 on page 26. The R-squared value is 0.016 for adjusted GPA and 0.015 for unadjusted GPA. Although the R-squared values increase compared to those obtained by using students' original SAT scores, they are still very low, indicating that the sum of Mathematical SAT scores and Verbal SAT scores is not a useful variable for modeling students' GPA. The poorness of SAT scores as a predictor persisted even once a complete set of scores was used.

4.1.3 The ACT's predictive power for students' four-year GPA

Section 4.1.2 showed that the SAT is not a good indicator of students' four-year GPA. In this section we investigate whether the ACT is a better predictor of students' four-year GPA.

We begin by using only the ACT score as an independent variable, and the students' four-year adjusted GPA as the dependent variable. We then repeat this model after adjusting GPA as described in Section 4.1.1. The fundamental strategy is the same as equation (2):

$$Y_i = \beta_0 + \beta_1 \, ACT_i \tag{6}$$

The R-squared of the model shown in (6) is approximately 0.12 when using either adjusted, or unadjusted, GPA. Although this R-squared value demonstrates that there is still much variability left unexplained, it is substantially larger than what we obtained for all the models in Section 4.1.2, which used SAT scores.

However, this does not provide conclusive evidence that the ACT is a better predictor than the SAT, since many students did not take the ACT exam. Only 267 out of 478 students enrolled in 2008 took the ACT exam. We augment our ACTscores by imputing values based on SAT scores as in equation (4). The new ACT variable is called Imputed ACT.

$$Y_i = \beta_0 + \beta_1 \, ImputedACT_i \tag{7}$$

Running the models based on equation (7) by using both adjusted, and unadjusted, GPA, we get R-squared values 0.031, and 0.036, respectively. These R-squared values are lower than the ones from the model in (6), however, they are still larger than those of the models using SAT scores. It thus seems reasonable to conclude that the ACT has better, although still not good, predictive ability compared to that of the SAT scores for students' four-year GPA.

4.1.4 The SAT's predictive power for students' first-year GPA

In this section, we investigate the relationship between students' first-year GPA and the SAT scores. The rationale for using first-year GPA instead of four-year GPA stems from two sources: 1) Proximity to when the SAT was written. 2) The tendency for grades in sophomore, junior, and senior years to show less variability, perhaps because students in sophomore, junior and senior years become more accustomed to college life. Thus, our hypothesis is that the SAT is more likely to correlate well with a student's first-year grades.

In this model, the dependent variable is a student's first-year cumulative GPA, and the independent variable is the student's sum of SAT Verbal scores and SAT Mathematical scores, without imputation:

$$FY_i = \beta_0 + \beta_1 \, SAT_i \qquad (8)$$

FY_i represents student i's first year cumulative GPA, SAT_i is the sum of student i's SAT Verbal scores and SAT Mathematical scores.

The R-squared of the model in (8) is 0.022. Although it is still very small, compared to the result of model (2), the R-squared value has improved a great deal.

Similarly, we build another model by using the Imputed SAT scores. The model is:

$$FY_i = \beta_0 + \beta_1 \, ImputedSAT_i \qquad (9)$$

The R-squared value of the model in (9) is 0.030, larger than the results of equation (5) by using unadjusted, or adjusted, four-year GPA. These results are consistent with our hypothesis that first year GPA can be better predicted than four-year GPA.

4.1.5 The ACT's predictive power for students' first-year GPA

In this section we test whether the ACT scores has more predictive power for students' first-year GPA than the SAT scores. We will also compare the results to those in section 4.1.3 to see whether the relationship between the ACT and first-year overall GPA is stronger than that between the ACT and four-year GPA. The model is:

$$FY_i = \beta_0 + \beta_1 ACT_i \qquad (10)$$

The R-squared value of the model in (10) is 0.113, which is larger than that of model (8) and (9). However, this R-squared value is a little bit smaller than that of the model in equation (6), which used the four-year grades. Thus, the ACT can better predict students' first year grades than the SAT. Also, the relationship between the ACT and first-year overall GPA is not stronger than that between the ACT and four-year GPA.

The model in (11) uses Imputed ACT as the independent variable:

$$FY_i = \beta_0 + \beta_1 ImputedACT_i \qquad (11)$$

Running model (11), we obtain an R-squared value of 0.048. This is substantially lower than what was found using only the observed ACT scores.

The R-squared value in this model is larger than that from the model using either the regular SAT (equation 8) or the Imputed SAT scores (equation 9), suggesting that the ACT is indeed a better measurement for students' first-year college academic performance than the SAT, although the predictive power is not very strong.

In addition, compared to the R-squared value of the model investigating the relationship between the Imputed ACT scores and students' four-year GPA (equation 8), the R-squared of model (11) is higher, which confirms that the Imputed ACT scores have more predictive power for students' first-year GPA than the four-year GPA

Table 5. SAT's and ACT's consistency in predicting students' four-year GPA

Dependent Variable	Independent Variable	P-value	Observations	R-squared
Overall GPA	MRSAT	0.0080	337	0.003
Adjusted Overall GPA	MRSAT	0.0370	337	0.001
Overall GPA	Imputed MRSAT	0.0018	478	0.015
Adjusted Overall GPA	Imputed MRSAT	0.0061	478	0.016
Overall GPA	ACT	$1.3*10^{-9}$	267	0.130
Adjusted Overall GPA	ACT	$3.6*10^{-9}$	267	0.123
Overall GPA	Imputed ACT	$3.2*10^{-5}$	478	0.036
Adjusted Overall GPA	Imputed ACT	0.0001	478	0.031

Table 6. SAT's and ACT's consistency in predicting students' first-year GPA

Dependent Variable	Independent Variable	P-value	Observations	R-squared
Overall GPA	MRSAT	0.5350	337	0.022
Overall GPA	Imputed MRSAT	0.0061	478	0.030
Overall GPA	ACT	$1.9*10^{-8}$	267	0.113
Overall GPA	Imputed ACT	$1.4*10^{-6}$	478	0.048

4.1.6 Summary

In general, the proceeding analysis shows that SAT scores have very little ability to predict academic performance at Macalester. We draw the following conclusions:

1. SAT scores are not a good predictor of students' grades, either for the first-year or for all four-year study in college.
2. Compared to the SAT scores, the ACT is a better predictor of students' college GPA, but it is still not a reliable predictor.
3. Generally, the SAT scores and ACT scores have more predictive power for students' first-year grades than the four-year grades. This is intuitive since a student's freshman year is closer to when the SATs were written, and upper division grades generally show less variability.

4.2 The SAT's Prediction of Major & Division of Study

Section 4.1 indicates that the SAT is not a valid predictor of Macalester students' college performance. This conclusion is in contrast with what College Board suggests in its literature. In this section, we study whether the SAT has the capability to predict other aspects of Macalester students' academic experience, for example, their major selection, or their academic division of study.

Since each student's major choice was not provided in our original dataset, we created a new *major* variable by selecting the department in which he/she took the most courses. In total, there are 33 majors represented by the graduating class of 2012. One drawback of this method is that, we are not able to ascertain whether a student was a double major. Also, we acknowledge that it is possible for a student to enroll in more courses in one department, but major in another department; however, this seems rather unlikely.

In addition, we categorize the different departments into the 4 unofficial academic divisions at Macalester: Social Sciences, Natural Sciences and Mathematics, Humanities, and Fine Arts. We are interested in the SAT's predictive power for division since there are many majors, and generally, majors in the same division require similar skills, and/or indicate similar interests. That is, major misclassification may be high even if the SAT is doing a reasonable job of "sorting" students. For example, our method may misclassify some Biology major students into Chemistry; however, since Biology and Chemistry require common courses, the spirit of the prediction is not far off. Hence we collapse similar majors according to their division. Section 4.2.1 to 4.2.5 use different strategies to investigate whether the SAT can predict students' major and division, and the predictive power of the SAT for such aspects. With the aid of Macalester's website, we classified each department into these 4 divisions as follows; with the number of students in each division shown in parentheses:

- **Social Sciences (170):** Anthropology, Economics, Geography, Linguistics, Political Science, Psychology, Sociology
- **Natural Sciences and Mathematics (133):** Biology, Chemistry, Geology, Mathematics, Statistices and Computer Science, Physics and Astronomy
- **Humanities (129):** Classics, English, French and Francophone Studies, German and Russian Studies, Hispanic Studies, History, Japanese, Media and Cultural Studies, Philosophy, Religious Studies
- **Fine Arts (46):** Art, Theater and Dance, Music.

This classification is not perfect, since many departments have courses that fall into more than one division. Although some misclassification will occur with both our major and division designations, we think that general trends can still be observed using our procedures. More precise classification for each student is not possible with the information provided to us.

4.2.1 Kullback–Leibler Divergence Approach – Majors

Kullback-Leibler (KL) divergence is a measure in statistics that compares the entropy of two different distributions over the same random variable (Shlens, 2007). Specifically, it quantifies the distance between a probability distribution p={P(i)} and a model distribution q = {Q(i)} in bits, that is, the required extra bits to code samples from P by using a code based on Q (Shlens, 2007). For two probability distributions, the following equation gives the value of the Kullback-Leibler divergence for discrete probability distributions P and Q.

$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i).$$

An intuitive understanding of the KL divergence is related to likelihood theory – the average probability of observing a set of data with the distribution P if the model Q indeed generated the dataset. It measures how much information is lost when Q is used to estimate P (Cover and Thomas, 1991).

In this study, we use KL divergence in order to investigate whether Macalester students' SAT scores could predict their choices of majors by comparing the distributions of major choice in different SAT categories.

As Figure 4.2.1 shows, we divide students' SAT scores into 9 categories according to their Verbal and Math scores. The horizontal lines represent students' Verbal scores ranging from 380 to 800; 660 and 720 correspond to 33.3 and 66.7

quantiles, respectively, of the Verbal scores. Similarly, the vertical lines represent the Math scores ranging from 410 to 800; 650 and 700 correspond to 33.3 and 66.7 quantiles, respectively, of the Math scores. We label each category as "Low Verbal" and "Low Math", "Median Verbal" and "Low Math", "High Verbal" and "Low Math", etc.

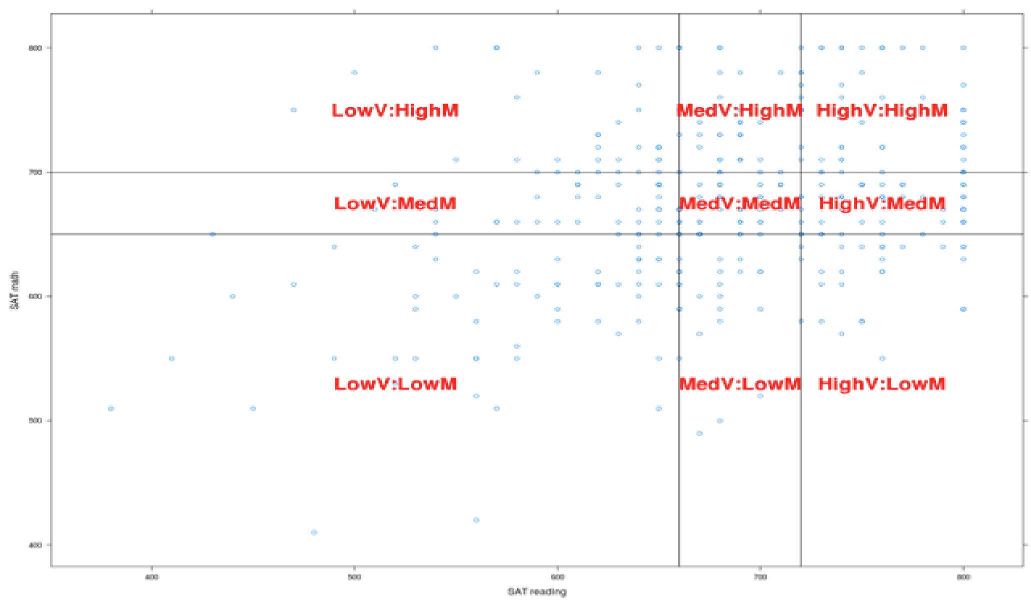


Figure 4.2.1: Nine categories of students' SAT scores based on Verbal and Math scores

We quantify the distance between two probability distributions of major corresponding to two of the nine SAT score categories. A larger KL divergence value indicates more distance between the probability distribution of major, thereby suggesting more significantly different major choice for students in these two categories. We compute 81 pairwise KL divergence values.

We would like to compare these observed differences to what would be expected if two categories truly had no difference in their major probability distributions. Implementing such a null hypothesis would provide us with a threshold value to decide whether the observed distance is big enough to be considered "significant". A straightforward approach implementing the null hypothesis is to obtain the sampling distribution by shuffling the majors randomly and calculating the KL divergence many times. We could then use the 95th percentile of these estimated sampling distributions as threshold values for deciding on significance, that is, if the observed KL divergence values are larger than the 95% cutoff values, then we would conclude that the major probability distributions are different for the two groups.

The 81 KL divergence values as well as the 95% cutoff values are shown in Table 1 in the appendix. We observe that for most comparisons, the null hypothesis cannot be rejected. As for those pairs which are deemed significantly different, we observe that the Verbal scores can better explain such difference than the Math scores in most cases. The pairs HighV:LowM and LowV:LowM, HighV:MedM and MedV:MedM, LowV:LowM and MedV:LowM show significant KL divergence values with the same Math scores, indicating Verbal scores' contribution to the significant distance between categories. HighV:MedM and LowV:HighM, HighV:MedM and LowV:LowM, LowV:HighM and MedV:LowM and LowV:LowM and MedV:MedM also show significant distance. Notice that comparing the same pairs leads to similar but not exactly the same KL divergence values. For instance, the KL divergence value for LowV:LowM and HighV:LowM is close but not equal to that for HighV:LowM and LowV:LowM, because the KL values are based on different Q distributions. Section A.2 in the appendix presents these results more clearly as graphs. Section A.3 in the appendix presents the 9 categories' major distributions, and we can see that these distributions do vary among the 9 categories, indicating that students with different SAT scores have different major choices.

4.2.2 Kullback–Leibler Divergence Approach – Divisions

We repeat the process from Section 4.2.1 but with division in place of major.

Since majors belonging to the same division may require many common courses, we are interested in how different SAT scores are related to this broader notion of division categories including similar majors.

The result is shown in the appendix Table 2. We see that although for most cases the KL divergence values are not large enough to reject the null, the rejection is more likely compared to the result in Section 4.2.1. Also, Section B.3 in the appendix presents the nine categories' division distributions and we observe that the distributions do not look similar across the categories. Therefore, we conclude that the SAT Verbal and Math scores can provide useful information on students' course divisions, and the prediction of the SAT scores for students' course divisions is expected to be better than that for students' major choice. The more strict measure of the SAT scores' prediction of students' course divisions is shown in Section 4.2.4.

4.2.3 The SAT's prediction of Majors – Classification Trees Approach

Section 4.2.1 and 4.2.2 indicate that although the SAT is not a perfect predictor of students' majors and divisions, it does have some predictive power worth probing further. KL divergence provides one way to see whether there was a difference in the major or division distribution, however, we still cannot tell much about which majors or divisions are being selected by which type of students. In this section, we use classification trees for constructing prediction models in order to measure the prediction power of the SAT and specifically see which majors are more likely to be selected by which type of students.

Classification trees are machine-learning methods for building prediction models from data (Lol, 2011). The models are constructed by partitioning the data space and then in each partition fitting a prediction model. Then the partitioning process can be graphed as a decision tree.

The basic idea is straightforward. We want to predict a categorical response variable from inputs X_1, X_2, X_3... X_p. One approach is called a **partition**, which requires subdividing the data space into small regions. We then do the partitioning again, a process called recursive partitioning, until eventually we can fit simple models to the small regions of the data space. Figure 4.2.3a helps to explain this. The other way to present the prediction result is called a **decision tree** (Lol, 2011). We first apply a test to an input X_i, at each node in the tree. We then go to the right or the left branch of the tree depending on the result of the test. A case goes to the left node when the given condition is satisfied, goes to the right if not. This process continues until a leaf node is reached where we can make a prediction. Figure 4.2.3a gives an example of this decision tree structure wherein there are 2 explanatory variables 3 classes (Lol, 2011).

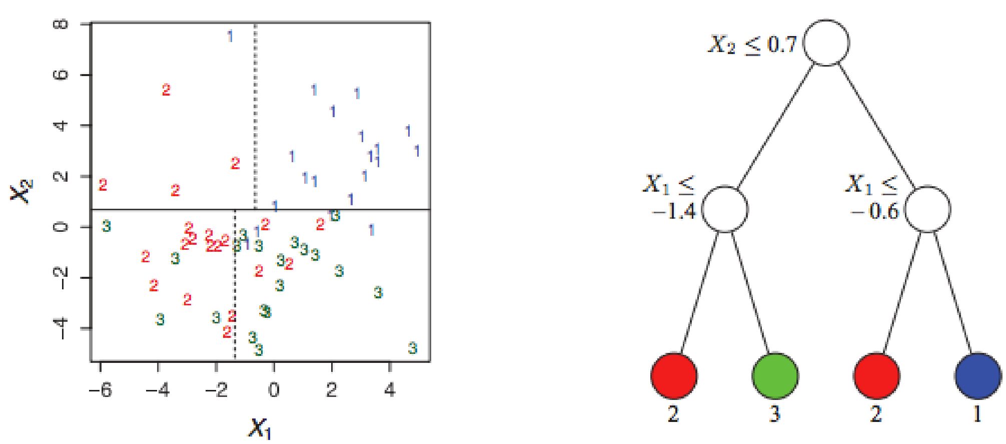


Figure 4.2.3a: Left: example of partition Right: example of decision tree structure.

We apply the method of classification trees to our dataset in order to investigate the SAT Verbal and Math scores' prediction of students' majors. Some majors are so small with few students and we only consider the majors with over 15 students. These 11 majors include Sociology (15 students), Anthropology (16 students), Biology (55 students), Economics (42 students), English (41 students), Geography (28 students), History (17 students), Mathematics/Computer Science (29 students), Music (28 students), Political Science (23 students), and Psychology (33 students). By doing this, we omit 22 majors and 105 students' major information.

We build a tree model in which major is the response variable and SAT Verbal and Math scores are the predictors. Figure 4.2.3b shows the plot of students' SAT Verbal and Math scores along with the partition of the tree. Figure 4.2.3c shows the decision tree structure. The partition omits three majors: Sociology, History and Geography, and this is related to the notion of overfitting. Overfitting is one of the classification trees' limitations, that the algorithm creates over-complex trees that does not classify the data well, so the decisions are poorly made towards the categories with little data (Mooney, 2007). To deal with the issue of overfitting, when there is not enough data to make reliable decisions, software (such as R) stops growing tree at some point during the construction. In our model, the number of students in Sociology, History or Geography majors is relatively small among the 11 selected majors.

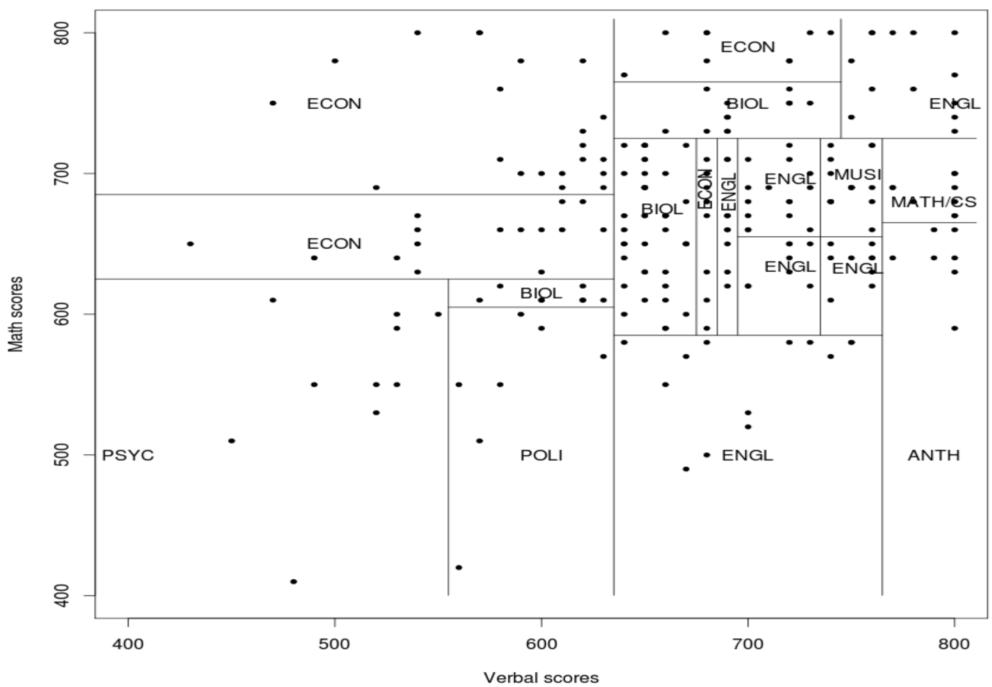


Figure 4.2.3b: Plot of students' SAT Verbal and Math scores, and the partition.

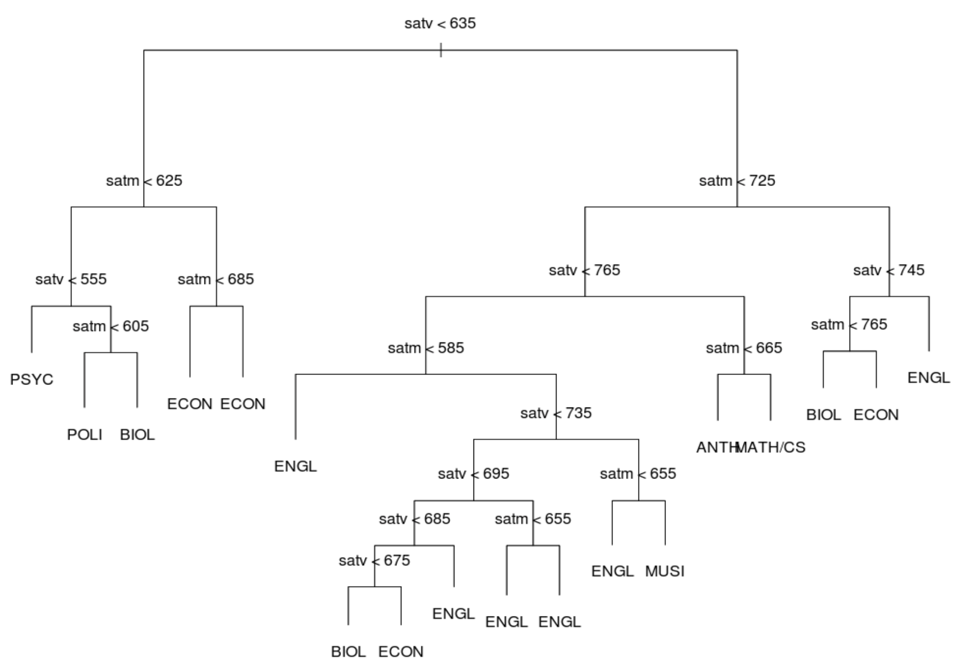


Figure 4.2.3c: Classification tree for predicting majors from the SAT scores.

According to Figure 4.2.3b and Figure 4.2.3c, students in English tend to have the highest Verbal and Math scores, while students in Psychology are more likely to have lower Verbal and Math scores. Notice that although some English major students' SAT Math scores are not the highest, their Verbal scores are generally in a top range. Also, students with relatively lower Verbal scores but higher Math scores most likely majored in Economics. Students with median Verbal and Math scores tend to major in Biology. The Political Science major attracted students mostly with median Verbal scores but low Math scores, and Anthropology major students generally gained top Verbal scores but low Math scores. Math major students seem to get both high Verbal and Math scores. The tree structure in Figure 4.2.3c conveys the same information with 18 nodes.

One common way to evaluate a model of classification tree is to calculate the model's misclassification rate, which is the fraction of cases assigned to the wrong class. For our model, the misclassification rate is 0.6897, which means that there are approximately 69% of students assigned to the wrong majors after implementing our tree model. Misclassification is illustrated in the plots in appendix C. For example, the Mathematics/Computer Science (MSCS) plot shows that students who majored in MSCS obtained high Math scores. However, our model predicts an MSCS major in the region with highest Verbal scores where only two MSCS students are included. Nevertheless, our classification tree method successfully categorizes a substantial number of students into their actual major.

One way to evaluate our tree model's misclassification rate is by comparing it to what we would obtain as a misclassification rate from a "naïve" model. In our sample, there are 11 majors and 327 students in total. Suppose we only have information on how many students are in each major and have to guess each of our sample student's major. Perhaps the simplest way to do this would be by assigning each student to the largest major, in our case, Biology, with 55 students. This naïve model would result in a misclassification rate of 0.8318. Thus, our model outperforms this naïve model.

Alternatively, we can use the observed major distribution to predict a student's major by simply randomly generating a prediction. We can do this for every student, and repeat the procedure many times in order to estimate the misclassification rate distribution. Figure 4.2.3d shows the sampling distribution of 1000 misclassification rates, and the 2.5th percentile is 86%. The misclassification rate of our tree model is 68.97%, which indicates that SAT scores provide useful information about a student's major selection.

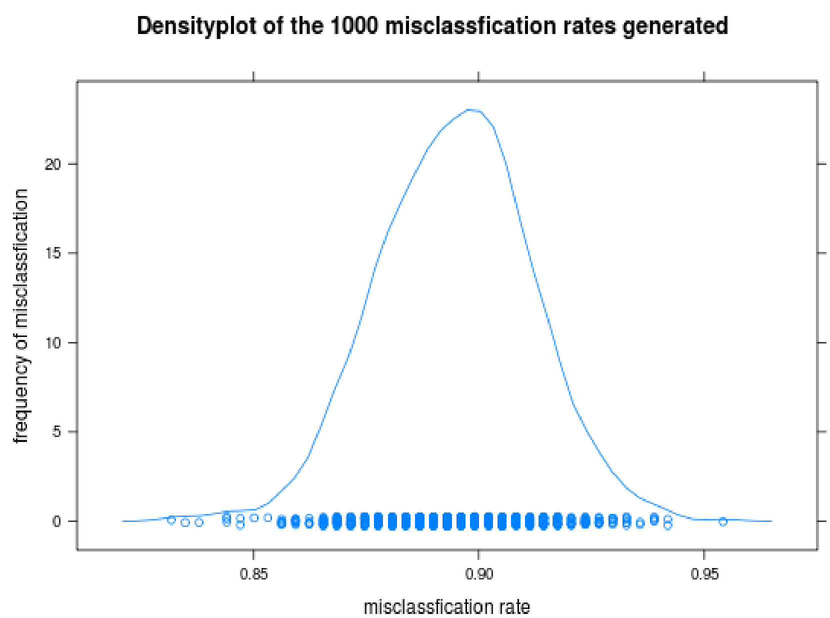


Figure 4.2.3d: The sampling distribution of 1000 misclassification rates

4.2.4 The SAT's prediction of Divisions – Classification Trees Approach

In order to evaluate the SAT's potential prediction of courses' divisions, we replicate the same procedure above except replacing the major variable with division.

We build a tree model that uses division as the response variable and SAT Verbal and Math scores as the predictors. Figure 4.2.4a shows the plot of students' SAT Verbal and Math scores where each color corresponds to a division, along with the partition of the tree. Figure 4.2.3b shows the decision tree structure.

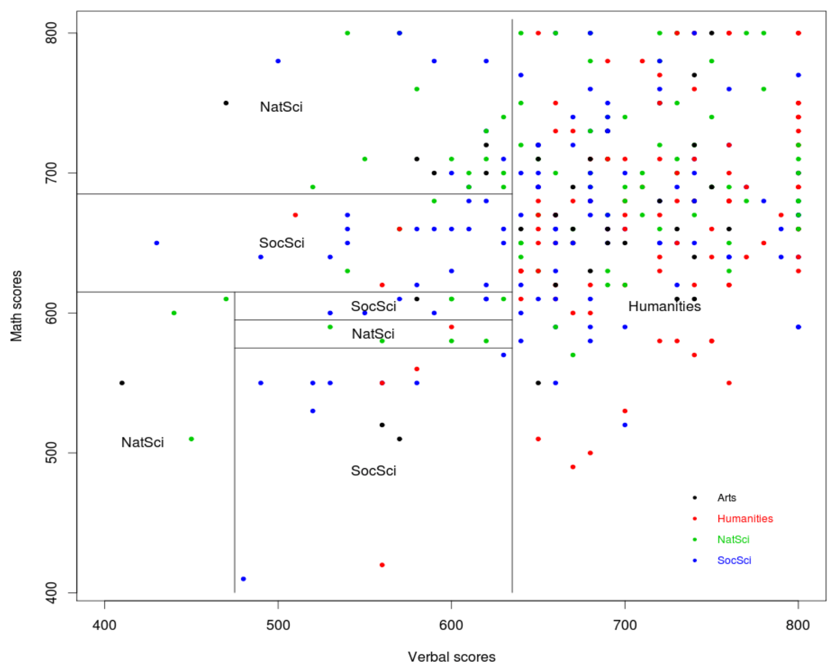


Figure 4.2.4a: Plot of students' SAT Verbal and Math scores, and the partition.

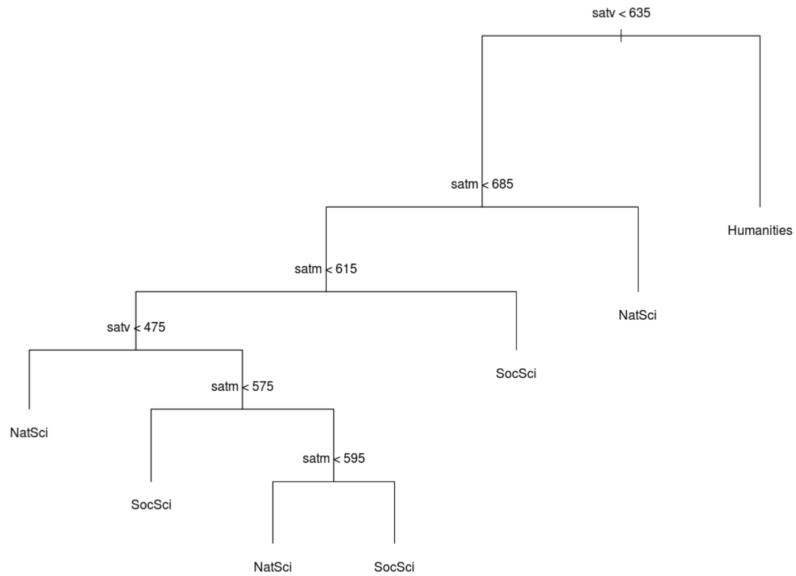


Figure 4.2.4b: Classification tree for predicting divisions from the SAT scores.

In the first partition plot, black refers to Arts division; red refers to Humanities; green refers to Natural Sciences and Mathematics; blue refers to Social Sciences. Notice the Arts division (black dots) is omitted in both plots because the small number of students in this division makes our tree model hard to produce reliable decisions. According to the two graphs, students with the highest Verbal scores are more likely to major in Humanities. As for Natural Sciences and Mathematics and Social Sciences, the distributions are not clustered. Some students who majored in Natural Sciences obtained high Math scores but low Verbal scores. Other students in the Natural Sciences got both low Math and Verbal scores. Students in Social Sciences obatained either median or low Math scores with generally median Verbal scores.

The misclassification rate is 0.5786, indicating that there are approximately 58%

of students assigned to the wrong divisions by implementing our tree model. Notice this rate is lower than the one in Section 4.2.3 by using the SAT scores to predict majors. In addition, according to the figures in appendix D, we observe that although our classification tree method successfully categorize most of students in Humanities division, for students in either Natural Sciences division or Social Sciences, our tree method misclassifies a large portion of them into Humanities. Also, by implementing the naïve method described in Section 4.2.3, we assign each student the biggest division, Social Sciences. We then compare the naïve method's misclassification rate with the tree model's. The naïve method's misclassification rate is 64.44%, which is larger than that of the tree model, 57.86%, so we conclude that our tree model is better than the method merely based on the base rate.

Similar to the method we used in Section 4.2.3, we also randomly attribute a division to each student and constructe 1000 misclassification rates based on different tree models. The densityplot of the 1000 misclassification rates is shown in Figure 4.2.4c. The 2.5% significance level is 67%, which is larger than the true misclassification rate, so we conclude that that our classification tree model that predicts students' divisions can provide useful information based on the criteria of misclassification rate.

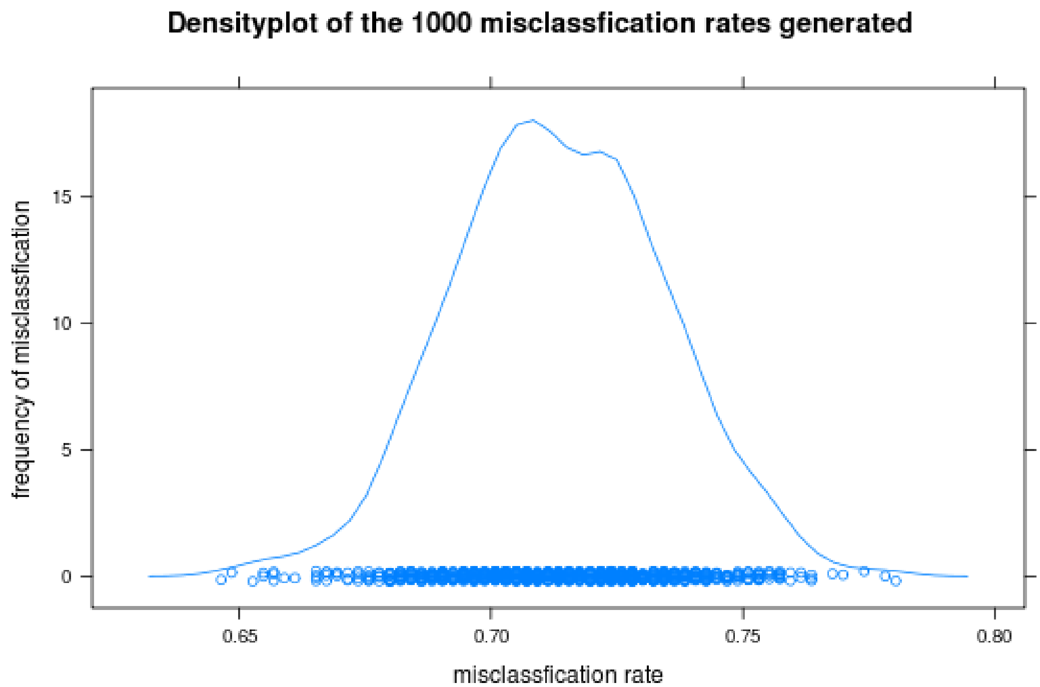


Figure 4.2.4c: The sampling distribution of 1000 misclassification rates

4.2.5 Categorizing Economics into Natural Sciences division

Although Macalester categorizes Economics as a Social Science, many debates have been raised about whether Economics should be a Natural Science instead (Nelson, 2005). From Macalester's academic website[4], the goal of the Economics department is to "develop analytical skills which contribute toward the understanding of our own and other economic systems," indicating that Economics is inclined to be a natural science since the basic elements of modern economic systems are objects rather than subjects. This section replicates the classification tree method in Section 4.2.3 except categorizing Economics into Natural Sciences division.

Figure 4.2.5a shows the plot of students' SAT Verbal and Math scores where each

[4]Refer to http://www.macalester.edu/academics/economics/majorsminors/ for more detail.

color corresponds to a division, along with the partition of the tree. Figure 4.2.5b shows the decision tree structure.

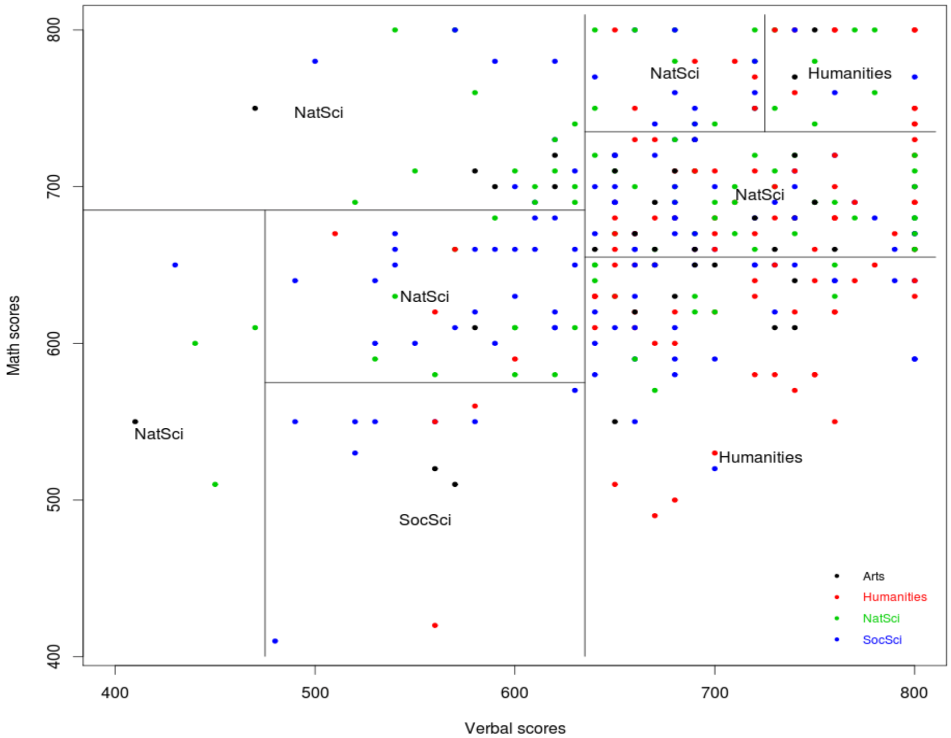


Figure 4.2.5a: Plot of students' SAT Verbal and Math scores, and the partition.

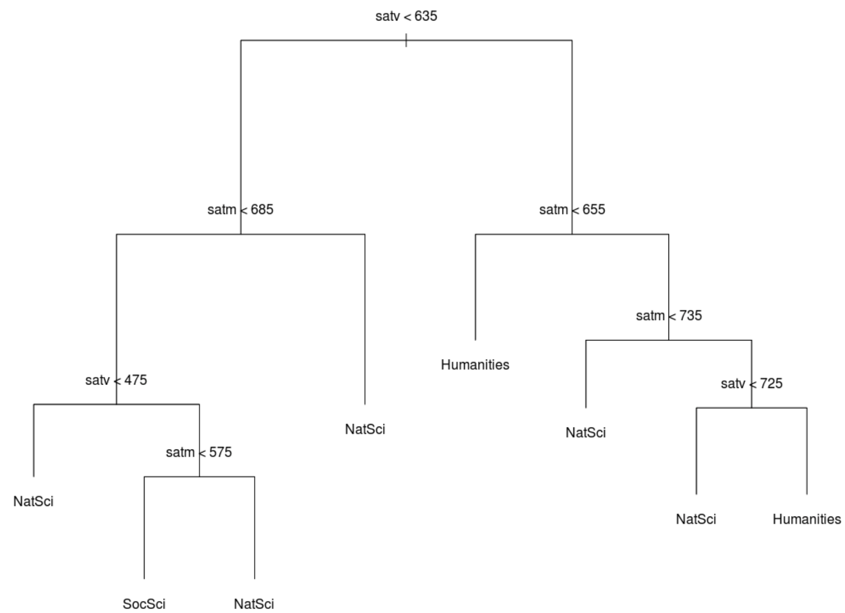


Figure 4.2.5b: Classification tree for predicting divisions from the SAT scores.

By comparing Figure 4.2.5a and Figure 4.2.5b to Figure 4.2.4a and Figure 4.2.4b, we can see that students with high Verbal scores are not only categorized into Humanities division; some students in Natural Sciences division also tend to obtain high Verbal and Math scores. Furthermore, the number of students in Social Sciences becomes fewer; only one region represents theses students with low Math scores and median Verbal scores. More students are categorized into Natural Sciences division with almost all the levels of SAT Verbal/Math scores. The recategorizing process does alter the classification due to the increased sample in Natural Sciences division. In addition, the misclassification rate is 55.67%, which is smaller than those in Section 4.2.3 and 4.2.4, indicating the best tree model with strongest SAT's prediction among the three. The figures in appendix E show that for students in Humanities and Natural Sciences, most of them are categorized correctly. However, this model does not categorize students in Social Sciences very well; most of them are misclassified into Humanities or Natural Sciences.

4.2.6 Summary

In Section 4.2, we study the SATs' potential to predict Macalester students' majors and division choice. By implementing Kullback–Leibler Divergence model and classification trees method, we find the SATs did contribute to predicting students' majors or courses' divisions. Furthermore, it seems that the Verbal scores can better predict majors than Math scores, and the prediction of the SAT scores for students' course divisions is better than that for students' major choice with more significant KL divergence values and lower misclassification rate. In addition, by recategorizing Economics into Natural Sciences division, the tree model improves, with the best misclassification rate among the three models we fit.

5. Conclusions and Future Work

5.1 Discussions and Conclusions

The SAT today serves as a widely used standardized test for college admissions which measures students' readiness for college. Although the College Board argues that the SAT is a reliable and valid predictor of college success, scholars still debate about its predictive power.

This paper mainly studies the SAT's predictive power for Macalester students' college performance defined by the overall first year grades and four year grades. The linear regression models lead to the conclusion that the SAT is not a valid predictor of either students' first year grades or four year grades, opposed to what the College Board suggested. We also investigate such predictive power of another similar standardized test, which is mainly used in the Midwest – ACT. The results show that although the ACT still cannot predict students' either first year or four year grades, it has more predictive power compared to the SAT. Furthermore, both of these tests predict students' first year grades better than four-year grades. There are two possible explanations behind the result that the individual SAT itself does not provide useful information for Macalester students' college success. First, the low predictive validity of the SAT for students' academic performance is a general trend. In other words, the SAT is not able to predict students' college success not only at Macalester College, but also at other universities or colleges. Second, it is perhaps features unique to Macalester which lead to the SAT's insignificant predictive validity. For example, it is possible that at Macalester, professors may grade students based on how much effort he or she contributes, or how much improvement the student has achieved in class, or against a set of absolute learning objects. Such different criteria make the SAT's predictions more difficult. Furthermore, stronger students might be more likely to challenge themselves; they choose harder courses and therefore gain lower grades. At the same time, relatively weaker students might be more likely to choose easier courses and get higher grades at Macalester College. As a results, due to the distinct course selection strategies, students' GPAs can not completely reflect their reasoning ability and educational achievement that the SAT can assess, so that the SAT does not possess significant predictive ability for students' grades.

Since the SAT is not a useful predictor of Macalester students' college performance, we study further whether the SAT can predict other aspects of Macalester students' academic experience, for instance, majors and divisions. We use Kullback–Leibler Divergence model and classification trees method to approach the SAT's prediction. The results show that although it does not serve as a perfect predictor, the SAT does have some valid prediction of majors and divisions, especially divisions. Such results might be explained by students' course selection strategy. We suppose that students' SAT scores may have some priming effects that make them believe which majors they should go to.

Our study is useful to give recommendations for Macalester College at the time of admission. In comparison to the prediction of academic achievement, the SAT does a much better job in predicting students' college academic choices. Thus, Macalester College may reduce the weight that it places on individual SAT scores in admissions process. Also, if Macalester intends to encourage a certain area of study, it might admit more students with SAT scores in a certain range.

5.2 Future Work

In the future we should make a more complete dataset that is representative of Macalester students. First, as indicated in the Methods section, our dataset suffers from the "restriction of range" problem; the available data only includes admitted students while excludes students who were not admitted or admitted but chose another school. In future study, we should obtain a more complete dataset with the whole applicant pool to get rid of the "restriction of range." Furthermore, our dataset only includes one-year sample of students graduated in 2012. In order to get more representative and reliable results, the sample should include students in different graduate years as what most literature did.

This study focuses on only Macalester students, however, we can probe more about the SAT's prediction of college success using other schools' data and draw more general conclusions in the future.

There are also other interesting issues related to the SAT worth studying. For instance, one might wonder whether the difference between a student's Verbal and Mathematics SAT scores is an important predictor of his or her academic choices. For example, it may be possible that a student with a high Verbal score is likely to major in Mathematics, say, if their Mathematics score is even higher than their Verbal score.

Also, the fairness of the SAT raised serious debates. For example, although ETS maintains that the SAT can identify students' potential from diverse ethnic backgrounds, the SAT has been shown to be culturally biased against African Americans, Asian Americans, and Hispanic Americans (Freedle, 2003). Such racial bias issue can be an interesting topic in the future. Similarly, a gender bias issue is also a direction for future research related to the SAT's fairness.

References

ACT (1997b). ACT Prediction Research Service. Iowa City, IA: ACT, Inc.

ACT (1998c). ACT Prediction Research Service. Iowa City, IA: ACT, Inc.

Agronow, S., and Studley, R., 2007, Prediction of college GPA from new SAT test scores - a first look. Annual meeting of the California Association for Institutional Research (CAIR), Nov 16, 2007. http://www.cair.org/conferences/CAIR2007/pres/Agronow.pdf

American Educational Research Association/American Psychological Association/National Council on Measurement in Education (1999). *Standards for educational and psychological testing.*

Baron, J. & Norman, M. F. (1992). SATs, achievement tests, and high-school class rank as predictors of college performance. Educational and Psychological Measurement, 52, 1047-1055.

Camara, W. J. & Echternacht, G. (2000). The SAT I and high school grades: Utility in *predicting success in college. Research Notes RN-10, The College Board, Office of Research and Development.*

College Board. (2001). College Bound Seniors 2001 (College Entrance Examination

Board, New York).

Donlon, Thomas F, ed. (1984). The College Board Technical Handbook for the *Scholas-tic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board.*

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. Harvard Educational Review, 73(1), 1–44.

Geiser, S. with R. Studley. (2002). "UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at theUniversity of California." Educational Assessment, vol. 8, no. 1, pp. 1-26. J. Baron & M. F. Norman in Educational and Psychology Measurement, Vol. 52, 1992.

Leonard, D. K. & Jiang, J. (1999). Gender bias and the college predictions of the SATs: A cry of despair. Research in Higher Education, 40(4), 375-407.

Morgan, R. (1989). *Analysis of the Predictive Validity of the SAT and High School Grades from 1976 to 1985. (College Board Research Report Mp. 89-7). New York: The College Board.*

Stricker, Lawrence, Rock, Donald, and Burton, Nancy (1991). Sex Differences in SAT Predictions of College Grades. College Board Report No. 91-2. New York: College Entrance Examination Board. Thorndike, R. L. (1963). The Concepts of Over- and Underachievement. New York:

Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L. (1990). Predicting College Grades: An Analysis of Institutional Trends Over Two Decades. Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1983). *A review of research on the prediction of academic performance after the freshman year* (College Board Report No. 83-2). New York: College Board.

Kobtin, J.L., & Michel, R.S. (2006). *The SAT As a Predictor of Different Levels of College Performance.* (College Board Research Report No. 2006-3). New York, NY: The College Board.

Kobrin, J.L., (Patterson, B.F., Shaw, E.J., Mattern, K.D., & Barbuti, S.M. (2008). Validity of the SAT for predicting first-year college grade point average. (College Board Research Report No. 2008-5). New York, NY: The College Board.

Loh, WY. (2011). Classification and regression trees. Department of Statistics, University of Wisconsin-Madison.

Mooney, Raymond J. (2007). Machine Learning: Decision Tree Learning. University of Texas Austin.

Nelson, Julie A. (2005). Cosmos and History: The Journal of Natural and Social Philosophy, vol. 1, no. 2, 2005.

Noble, J. (1991). Predicting College Grades from ACT Assessment Scores and High School Course Work and Grade Information. (ACT Research Report No. 91-3). Iowa City, Iowa: The American College Testing Program.

Noble, J., & Sawyer, R. (2002). Predicting different levels of academic success in college using high school GPA and ACT Composite score. (ACT Research Report 2002-4). Iowa City, IA: ACT.

Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics,* 121(1-2), 297-317.

Shlens, Jonathon. (2007). Notes on Kullback – Leibler Divergence and Likelihood Theory. Systems Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla..

Thomas M. Cover & Joy A. Thomas. (1991). Entropy, Relative Entropy and Mutual Information, Elements of Information Theory, John Wiley & Sons, Inc.

Appendix:

Appendix A: Kullback-Leibler divergence approach for majors

A.1

Table 1: Kullback–Leibler divergence – Majors[5]

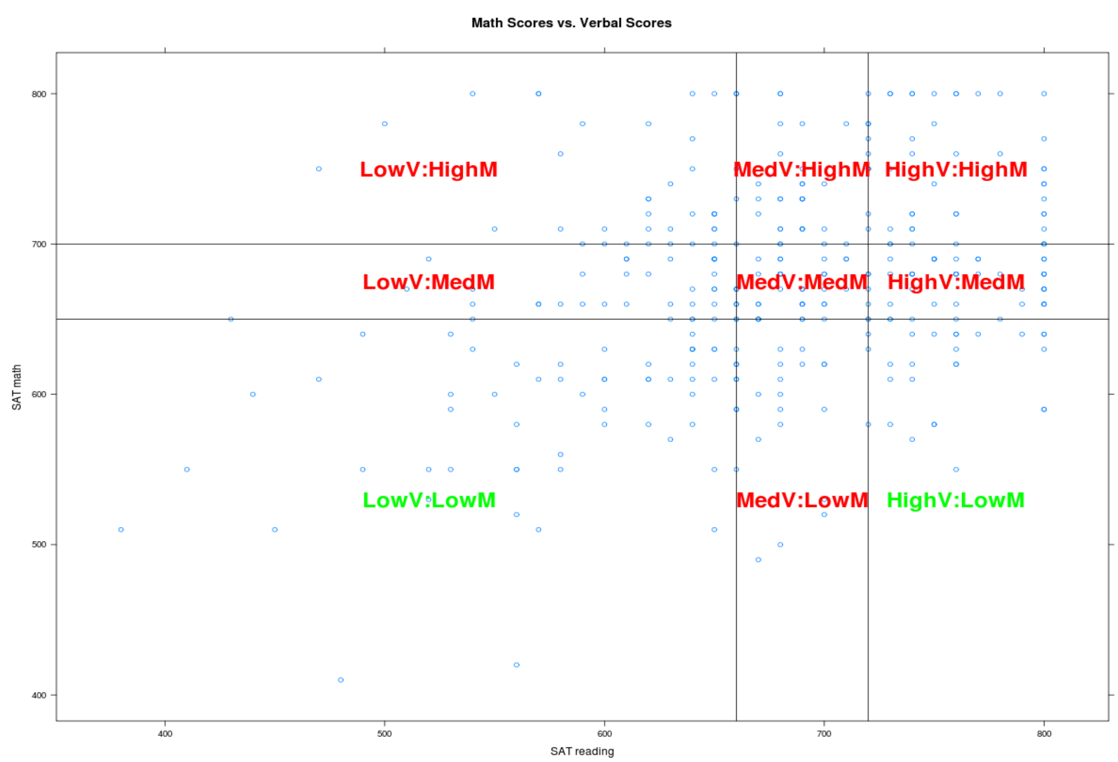
Qi	Pi	KL	x = 95% cutoff
HighV:HighM	HighV:HighM	0	0
HighV:HighM	HighV:LowM	0.32207488	0.6065216
HighV:HighM	HighV:MedM	0.48978582	0.6267838
HighV:HighM	LowV:HighM	0.3559699	0.675889
HighV:HighM	LowV:LowM	0.44861521	0.5600517
HighV:HighM	LowV:MedM	0.42315442	0.6667532
HighV:HighM	MedV:HighM	0.24165015	0.6316114
HighV:HighM	MedV:LowM	0.45756673	0.5921309
HighV:HighM	MedV:MedM	0.30595003	0.6306256
HighV:LowM	HighV:HighM	0.38980061	0.5836863
HighV:LowM	HighV:LowM	0	0
HighV:LowM	HighV:MedM	0.48294887	0.5819691
HighV:LowM	LowV:HighM	0.28908684	0.6320109
HighV:LowM	**LowV:LowM**	**0.6790374**	**0.5611349**
HighV:LowM	LowV:MedM	0.37046815	0.5706924
HighV:LowM	MedV:HighM	0.19416266	0.6211342
HighV:LowM	MedV:LowM	0.12132009	0.6064527
HighV:LowM	MedV:MedM	0.13098168	0.694872
HighV:MedM	HighV:HighM	0.38901093	0.5971383
HighV:MedM	HighV:LowM	0.53233337	0.5800403
HighV:MedM	HighV:MedM	0	0
HighV:MedM	**LowV:HighM**	**0.62103027**	**0.5920394**
HighV:MedM	**LowV:LowM**	**0.73681376**	**0.5923996**
HighV:MedM	LowV:MedM	0.48999905	0.7541643
HighV:MedM	MedV:HighM	0.53323685	0.6264873
HighV:MedM	MedV:LowM	0.41375955	0.5520005
HighV:MedM	**MedV:MedM**	**0.62929344**	**0.5902563**
LowV:HighM	HighV:HighM	0.37526458	0.6574623
LowV:HighM	HighV:LowM	0.3101976	0.5949264

[5] The bold pairs represent those are deemed significantly different.

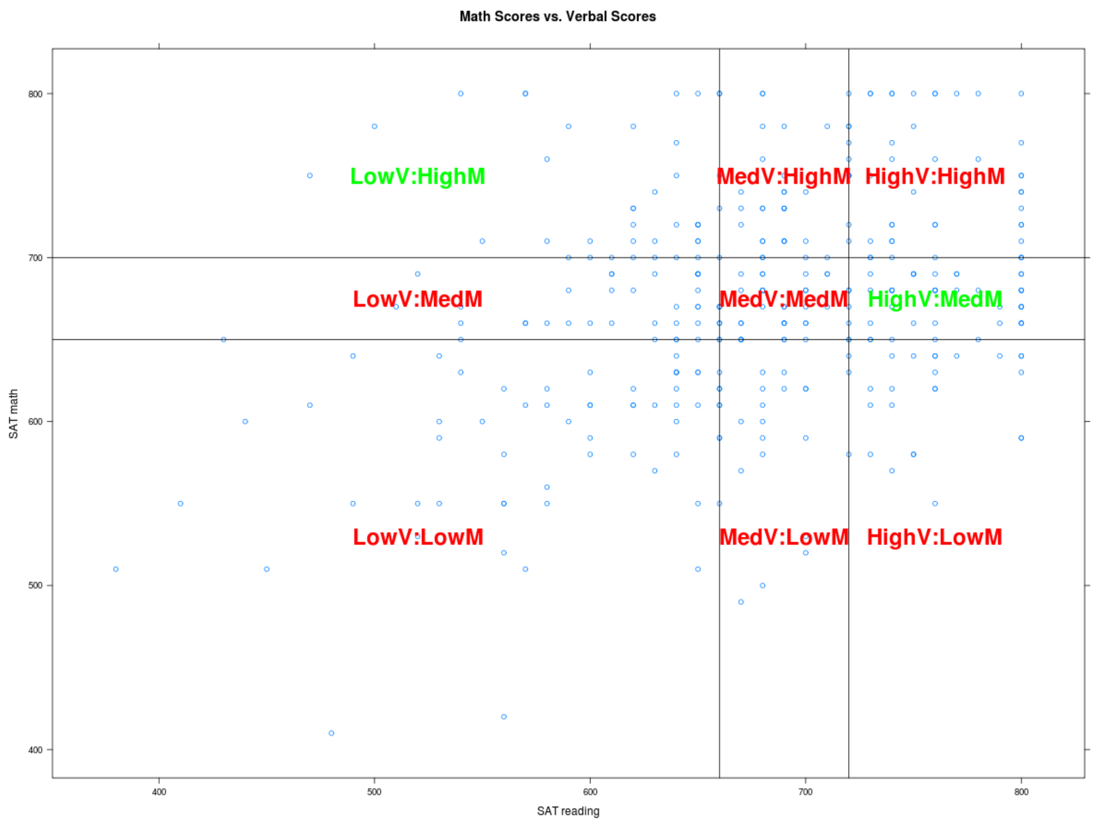
LowV:HighM	HighV:MedM	0.50716894	0.5578362
LowV:HighM	LowV:HighM	0	0
LowV:HighM	LowV:LowM	0.5516518	0.589695
LowV:HighM	LowV:MedM	0.27402853	0.524754
LowV:HighM	MedV:HighM	0.2663746	0.5834451
LowV:HighM	**MedV:LowM**	**0.5780643**	**0.5712165**
LowV:HighM	MedV:MedM	0.02333558	0.6193795
LowV:LowM	HighV:HighM	0.49011696	0.5603853
LowV:LowM	**HighV:LowM**	**0.87827118**	**0.5728334**
LowV:LowM	**HighV:MedM**	**0.83818141**	**0.5485502**
LowV:LowM	LowV:HighM	0.56900895	0.6036224
LowV:LowM	LowV:LowM	0	0
LowV:LowM	LowV:MedM	0.25411206	0.6114503
LowV:LowM	MedV:HighM	0.49509079	0.6311797
LowV:LowM	**MedV:LowM**	**0.67724178**	**0.5298502**
LowV:LowM	**MedV:MedM**	**0.70277344**	**0.5870337**
LowV:MedM	HighV:HighM	0.45565551	0.6216612
LowV:MedM	HighV:LowM	0.53511654	0.5528049
LowV:MedM	HighV:MedM	0.39267434	0.6977399
LowV:MedM	LowV:HighM	0.22097795	0.5499243
LowV:MedM	LowV:LowM	0.22673743	0.5615099
LowV:MedM	LowV:MedM	0	0
LowV:MedM	MedV:HighM	0.31209375	0.5798368
LowV:MedM	MedV:LowM	0.47779411	0.5950524
LowV:MedM	MedV:MedM	0.25506946	0.6177773
MedV:HighM	HighV:HighM	0.25242697	0.6255166
MedV:HighM	HighV:LowM	0.17627153	0.6416332
MedV:HighM	HighV:MedM	0.57645713	0.5828892
MedV:HighM	LowV:HighM	0.30250092	0.5754341
MedV:HighM	LowV:LowM	0.35519157	0.5533953
MedV:HighM	LowV:MedM	0.29196022	0.5902058
MedV:HighM	MedV:HighM	0	0
MedV:HighM	MedV:LowM	0.29486708	0.6016204
MedV:HighM	MedV:MedM	0.32303233	0.5786389
MedV:LowM	HighV:HighM	0.4580041	0.5838906
MedV:LowM	HighV:LowM	0.12222728	0.627802
MedV:LowM	HighV:MedM	0.41520178	0.5858299
MedV:LowM	**LowV:HighM**	**0.59478921**	**0.5902982**
MedV:LowM	**LowV:LowM**	**0.6064439**	**0.5251573**
MedV:LowM	LowV:MedM	0.52758026	0.6026173

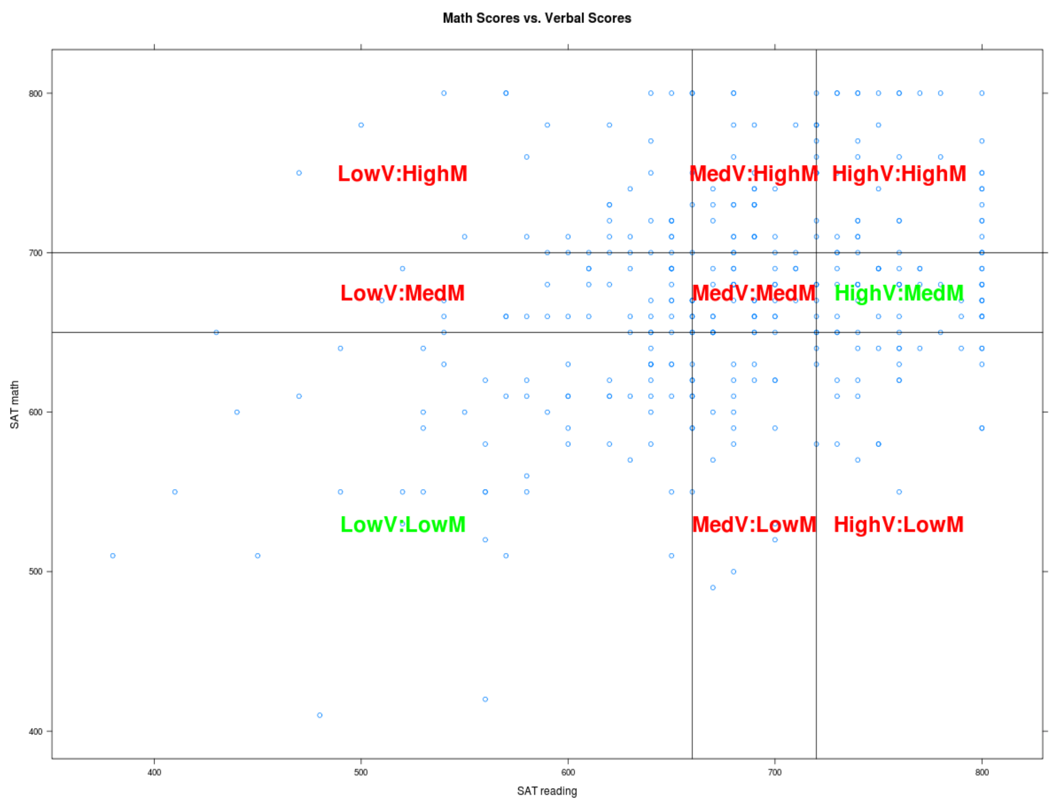
MedV:LowM	MedV:HighM	0.3319453	0.6042932
MedV:LowM	MedV:LowM	0	0
MedV:LowM	MedV:MedM	0.50978248	0.5221252
MedV:MedM	HighV:HighM	0.32744626	0.6595286
MedV:MedM	HighV:LowM	0.12890631	0.6140067
MedV:MedM	HighV:MedM	0.54156473	0.6197514
MedV:MedM	LowV:HighM	0.0236572	0.6493164
MedV:MedM	**LowV:LowM**	**0.69764356**	**0.5899409**
MedV:MedM	LowV:MedM	0.22145762	0.6206408
MedV:MedM	MedV:HighM	0.33712651	0.5637474
MedV:MedM	MedV:LowM	0.43877272	0.4860551
MedV:MedM	MedV:MedM	0	0

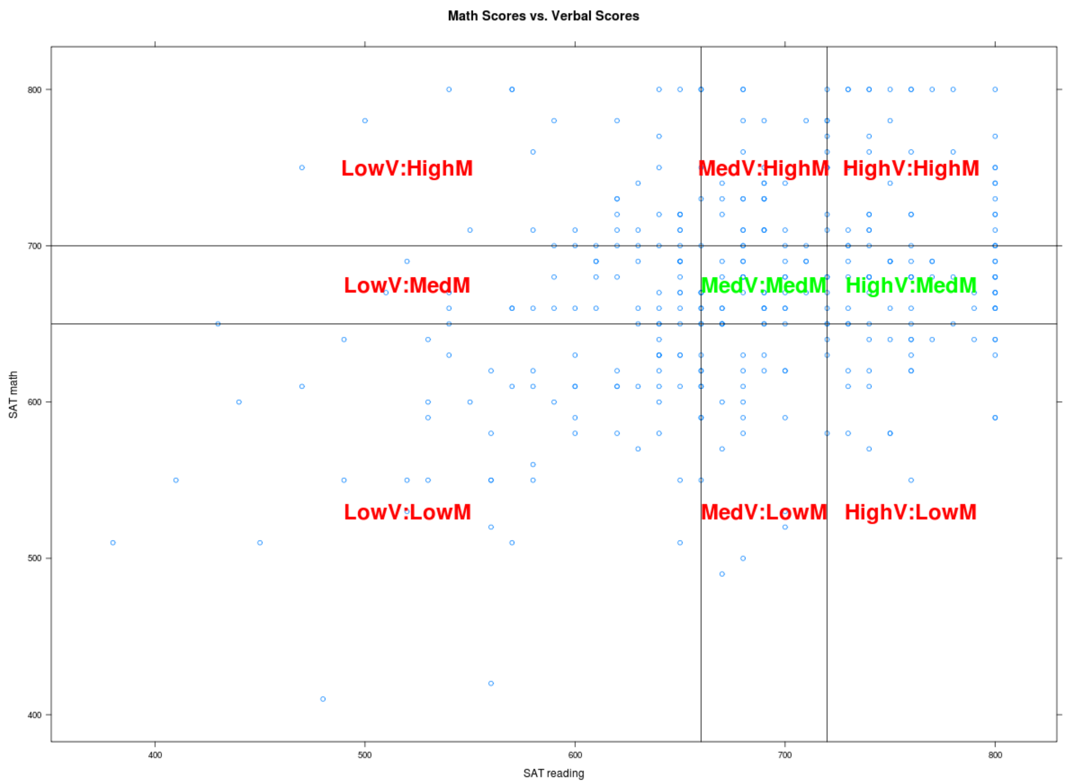
A.2: Graphic illustration of pairs with significant K-L divergence values[6]



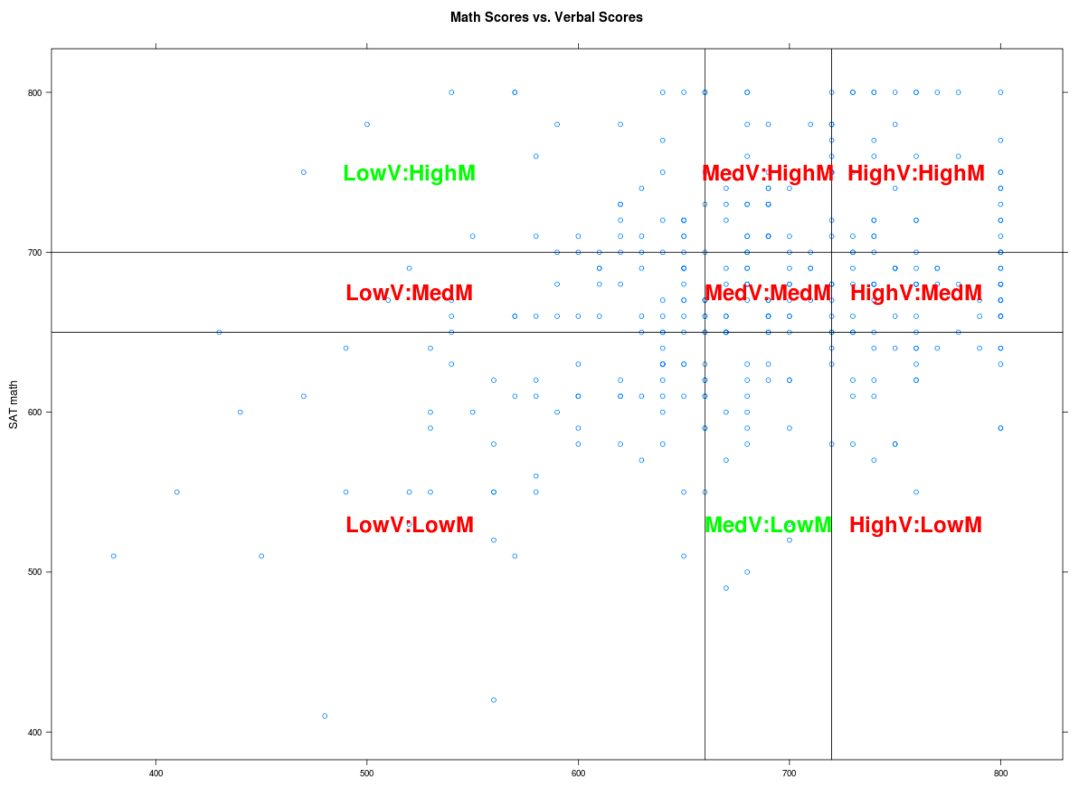
6 Green pairs represent those are deemed significantly different.


Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
800
700
600
500
400
400
500
600
700
800

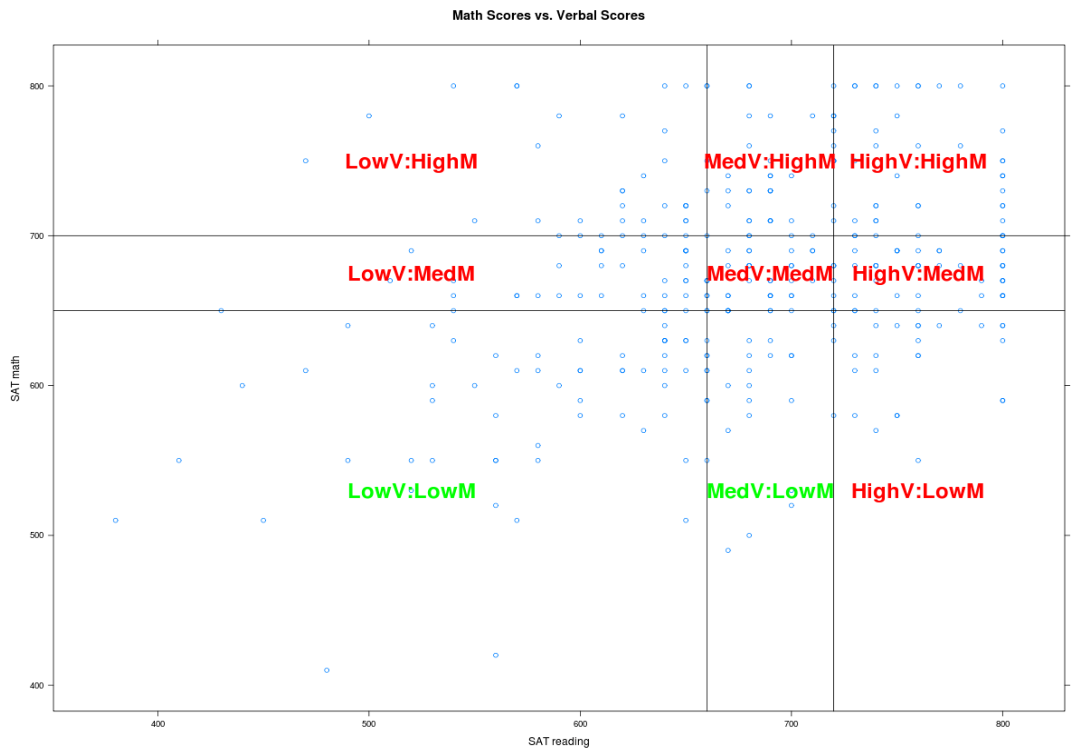

Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
800
700
600
500
400
400
500
600
700
800



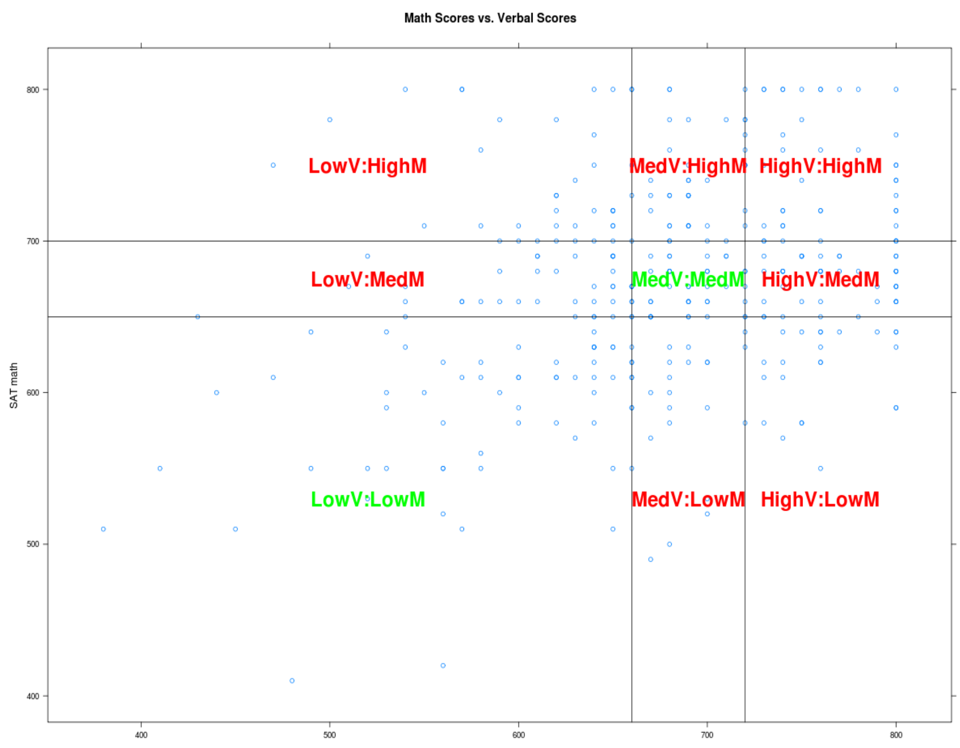
Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
800
700
600
500
400
400
500
600
700
800



Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
800
700
600
500
400
400
500
600
700
800



Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading



Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading

A.3: The 9 categories' major distributions

Category	LOWV: HIGHM	MEDV: HIGHM	HIGHV: HIGHM	LOWV: MEDM	MEDV: MEDM	HIGHV: MEDM	LOWV: LOWM	MEDV: LOWM	HIGHV: LOWM	Overall
AMST	0	0	0	0	0	2	2	0	1	5
ANTH	0	2	0	2	0	2	0	0	2	8
ASIA/ CHIN	0	1	2	0	0	0	0	0	1	4
BIOL	3	4	2	2	2	2	10	3	2	31
CHEM	2	0	1	1	0	0	2	0	0	6
CLAS	0	1	1	1	1	1	3	0	2	10
ECON	7	5	3	4	5	1	7	1	0	35
EDU	0	0	0	0	0	0	1	1	1	3
ENGL	1	4	4	1	4	2	1	6	7	31
ENVI	1	0	1	0	1	3	1	0	0	7
FREN	0	0	0	0	0	0	1	0	1	2
GEOG	0	1	1	1	3	2	4	3	2	19
GEOL	3	1	1	1	2	1	0	0	0	9
GERM	0	0	0	0	0	2	0	0	0	2
HISP/ LATI	0	2	2	0	0	0	2	1	0	7
HIST	0	0	3	1	1	2	3	2	1	13
INTL	0	0	0	0	1	0	1	0	0	3
JAPA	0	0	0	1	0	0	0	0	0	1
LING	0	1	5	0	0	1	0	0	0	7
MATH/ CS	4	2	0	1	2	4	3	2	0	20
MCST	1	0	1	0	0	0	1	1	0	4
MUSI	5	1	0	1	2	6	2	1	2	22
NEUR	0	0	3	1	2	0	0	0	0	6
PHIL	1	2	1	0	1	3	0	1	1	10
PHYS	0	3	0	0	1	0	1	0	0	5
POLI	1	0	0	3	0	0	6	1	2	13
PSYC	1	3	0	2	1	4	9	3	0	23
RELI	0	0	0	1	0	0	2	0	0	3
RUSS	0	0	0	0	0	2	0	0	0	2
SOCI	1	0	1	2	0	0	3	1	2	11
THDA	0	0	0	0	2	0	1	1	1	5
WGSS	0	0	0	1	1	0	0	0	0	2

Appendix B: Kullback-Leiber divergence approach for divisions[7]

B.1

Table 2: Kullback–Leibler divergence Approach – Divisions

Qi	Pi	KL	x = 95% cutoff
HighV:High M	HighV:High M	0	0
HighV:High M	**HighV:Low M**	**0.32723683**	**0.30269105**
HighV:High M	HighV:Med M	0.09632138	0.33789011
HighV:High M	**LowV:High M**	**0.45107779**	**0.37612875**
HighV:High M	**LowV:Low M**	**0.40324049**	**0.24566357**
HighV:High M	**LowV:Med M**	**0.54948076**	**0.32587628**
HighV:High M	MedV:High M	0.19887728	0.3474624
HighV:High M	MedV:LowM	0.20846232	0.40345528
HighV:High M	MedV:MedM	0.18215003	0.35649683
HighV:Low M	**HighV:High M**	**0.43479792**	**0.30257742**
HighV:LowM	HighV:LowM	0	0
HighV:LowM	HighV:Med M	0.25390962	0.29800975
HighV:Low M	**LowV:High M**	**0.93842951**	**0.35796354**
HighV:Low M	**LowV:Low M**	**0.36973357**	**0.25119426**
HighV:Low M	**LowV:Med M**	**0.41616525**	**0.2734547**
HighV:Low M	**MedV:High M**	**0.37211378**	**0.33797667**
HighV:LowM	MedV:LowM	0.09773652	0.34507869
HighV:Low M	**MedV:Med M**	**0.43049938**	**0.3570788**
HighV:Med M	HighV:High M	0.08757037	0.31441877

[7] The bold pairs represent those are deemed significantly different.

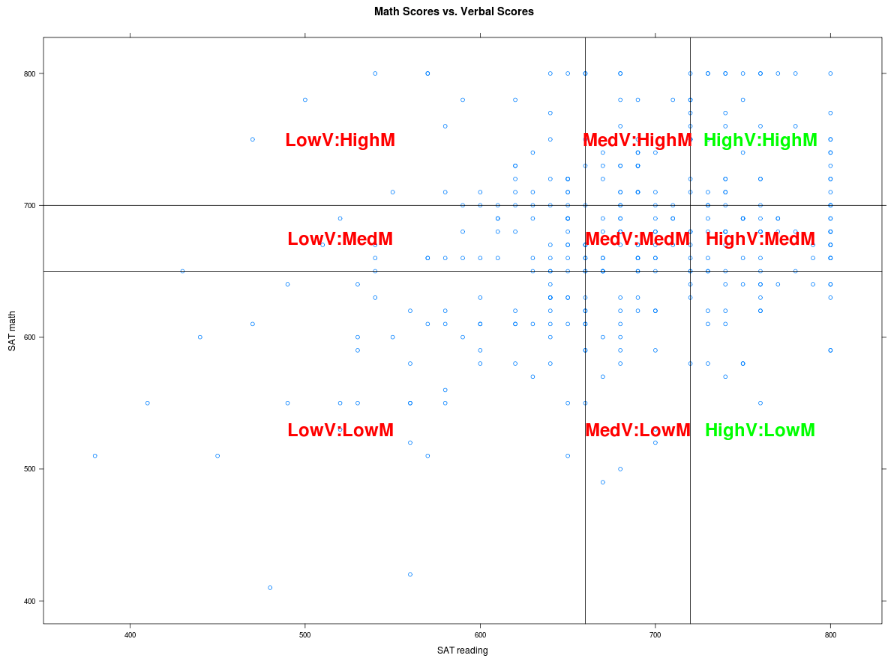
HighV:MedM	HighV:LowM	0.19272017	0.26880299
HighV:MedM	HighV:MedM	0	0
HighV:MedM	LowV:HighM	0.26896733	0.3084743
HighV:MedM	LowV:LowM	0.15698973	0.21281597
HighV:MedM	LowV:MedM	0.23230305	0.29092015
HighV:MedM	MedV:HighM	0.09129524	0.36304855
HighV:MedM	MedV:LowM	0.05611412	0.3666766
HighV:MedM	MedV:MedM	0.04696905	0.33863189
LowV:HighM	**HighV:HighM**	**0.6010551**	**0.34997732**
LowV:HighM	**HighV:LowM**	**0.93733031**	**0.35112428**
LowV:HighM	**HighV:MedM**	**0.35493282**	**0.27839131**
LowV:HighM	LowV:HighM	0	0
LowV:HighM	LowV:LowM	0.2020028	0.24503712
LowV:HighM	LowV:MedM	0.19462827	0.29225526
LowV:HighM	MedV:HighM	0.28135395	0.43816504
LowV:HighM	**MedV:LowM**	**0.4933086**	**0.33501882**
LowV:HighM	MedV:MedM	0.14942979	0.31752401
LowV:LowM	**HighV:HighM**	**0.34254148**	**0.24786178**
LowV:LowM	**HighV:LowM**	**0.35021**	**0.24889251**
LowV:LowM	HighV:MedM	0.15230814	0.19265797
LowV:LowM	LowV:HighM	0.19836237	0.25258083
LowV:LowM	LowV:LowM	0	0
LowV:LowM	LowV:MedM	0.01310125	0.2506821
LowV:LowM	MedV:HighM	0.04736446	0.2590229
LowV:LowM	MedV:LowM	0.11011572	0.25372371

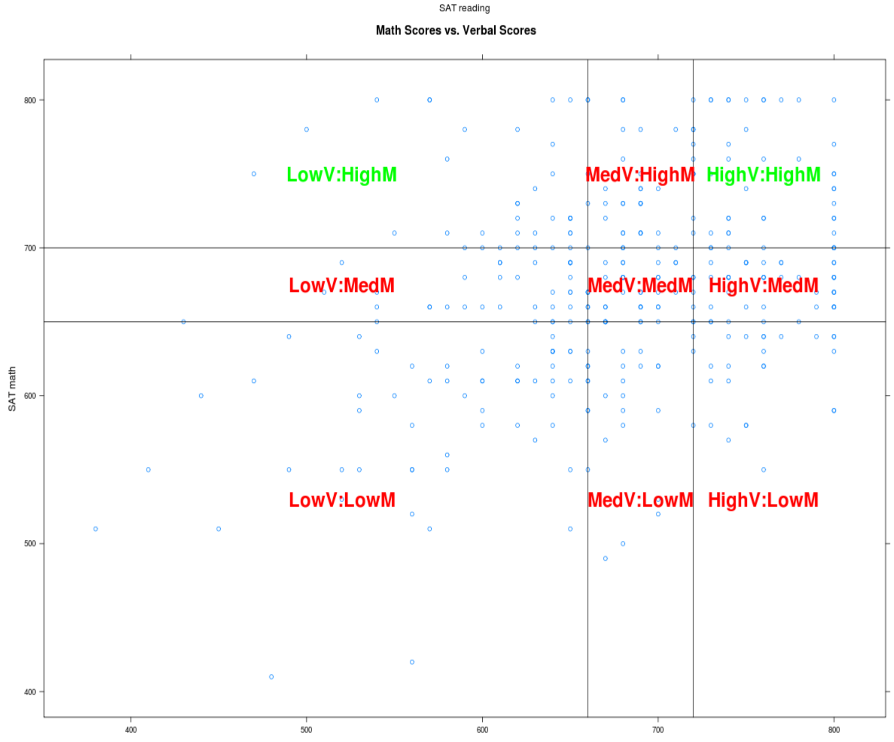
LowV:LowM	MedV:MedM	0.05936807	0.28813969
LowV:MedM	**HighV:HighM**	**0.48247982**	**0.3582301**
LowV:MedM	**HighV:LowM**	**0.44482352**	**0.29195491**
LowV:MedM	HighV:MedM	0.22574515	0.29588303
LowV:MedM	LowV:HighM	0.2012529	0.28029299
LowV:MedM	LowV:LowM	0.01342202	0.21563081
LowV:MedM	LowV:MedM	0	0
LowV:MedM	MedV:HighM	0.11158806	0.34146744
LowV:MedM	MedV:LowM	0.17731767	0.25272146
LowV:MedM	MedV:MedM	0.10509033	0.33208791
MedV:HighM	HighV:HighM	0.16843211	0.33304327
MedV:HighM	HighV:LowM	0.28623897	0.37557419
MedV:HighM	HighV:MedM	0.10741991	0.35433506
MedV:HighM	LowV:HighM	0.25226817	0.39520119
MedV:HighM	LowV:LowM	0.0477	0.26249788
MedV:HighM	LowV:MedM	0.10976654	0.3284929
MedV:HighM	MedV:HighM	0	0
MedV:HighM	MedV:LowM	0.0790364	0.38716298
MedV:HighM	MedV:MedM	0.04978257	0.3120396
MedV:LowM	HighV:HighM	0.18888355	0.38252348
MedV:LowM	HighV:LowM	0.08130615	0.31843544
MedV:LowM	HighV:MedM	0.05782606	0.32874899
MedV:LowM	**LowV:HighM**	**0.41933539**	**0.33107297**
MedV:LowM	LowV:LowM	0.10201147	0.25815982
MedV:LowM	LowV:MedM	0.15424253	0.2803911
MedV:LowM	MedV:High	0.0826197	0.36878621

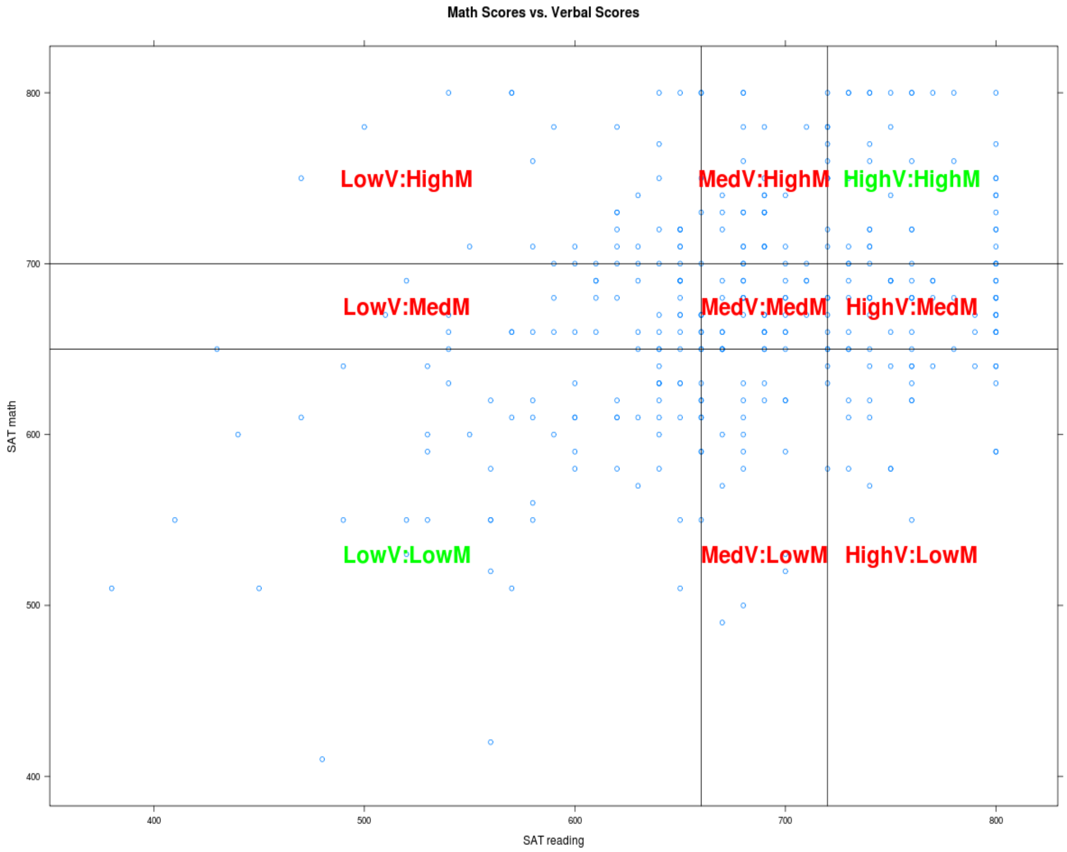
	M		
MedV:LowM	MedV:LowM	0	0
MedV:LowM	MedV:MedM	0.10750278	0.35308348
MedV:MedM	HighV:High M	0.17344261	0.29421465
MedV:Med M	**HighV:Low M**	**0.33714409**	**0.3259818**
MedV:MedM	HighV:Med M	0.0483905	0.34315456
MedV:MedM	LowV:HighM	0.11951914	0.30791673
MedV:MedM	LowV:LowM	0.06143988	0.31966145
MedV:MedM	LowV:MedM	0.11015832	0.31504774
MedV:MedM	MedV:High M	0.04123404	0.29608205
MedV:MedM	MedV:LowM	0.10093746	0.37814539
MedV:MedM	MedV:MedM	0	0

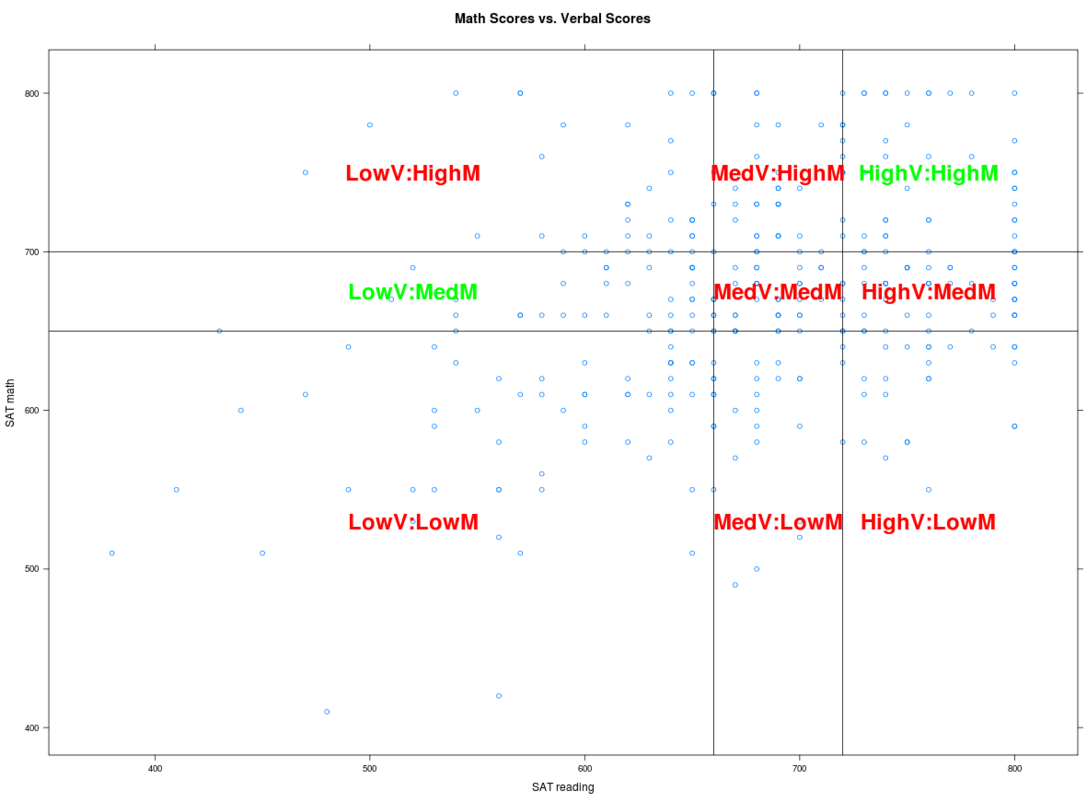
B.2 Graphic illustration of pairs with significant K-L divergence[8]

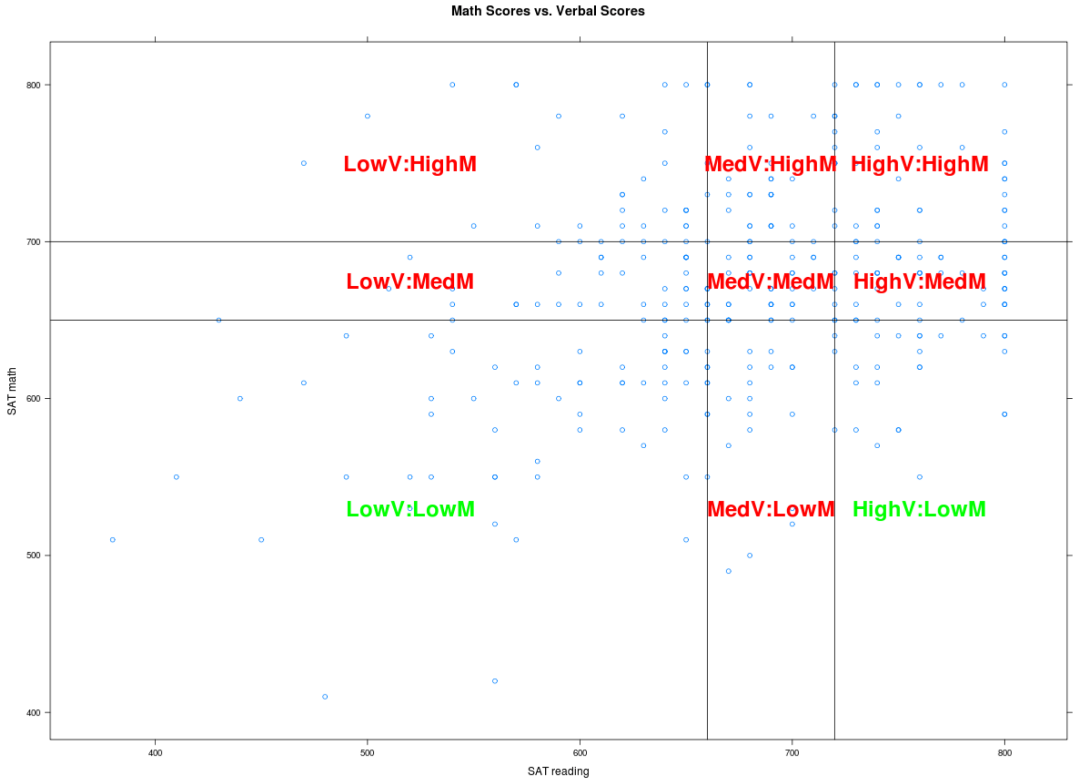
[8]Green pairs represent those are deemed significantly different.

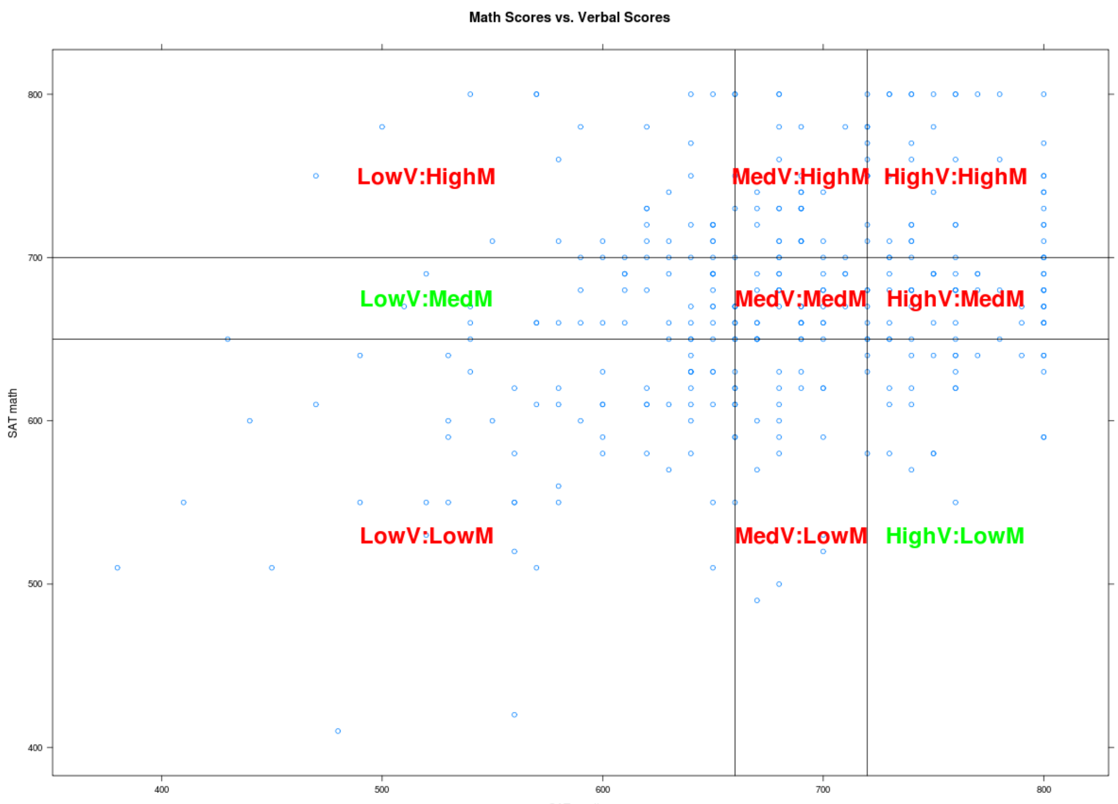

Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading


Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading

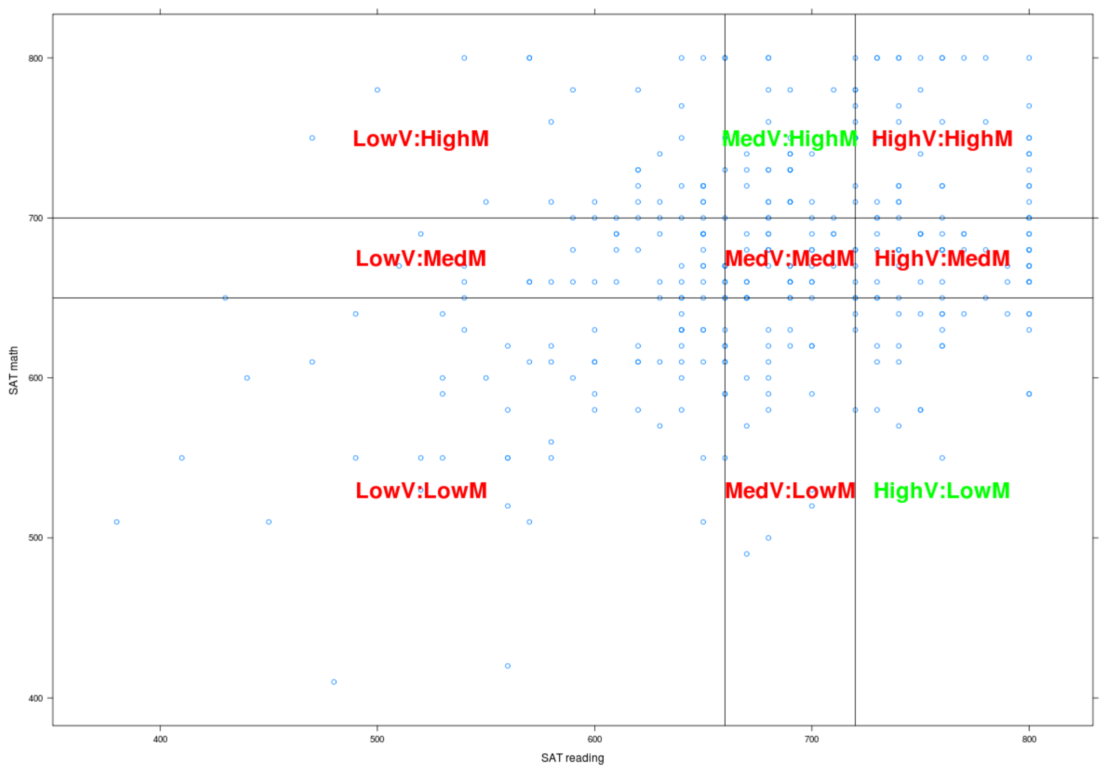

Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
400
500
600
700
800


Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
400
500
600
700
800

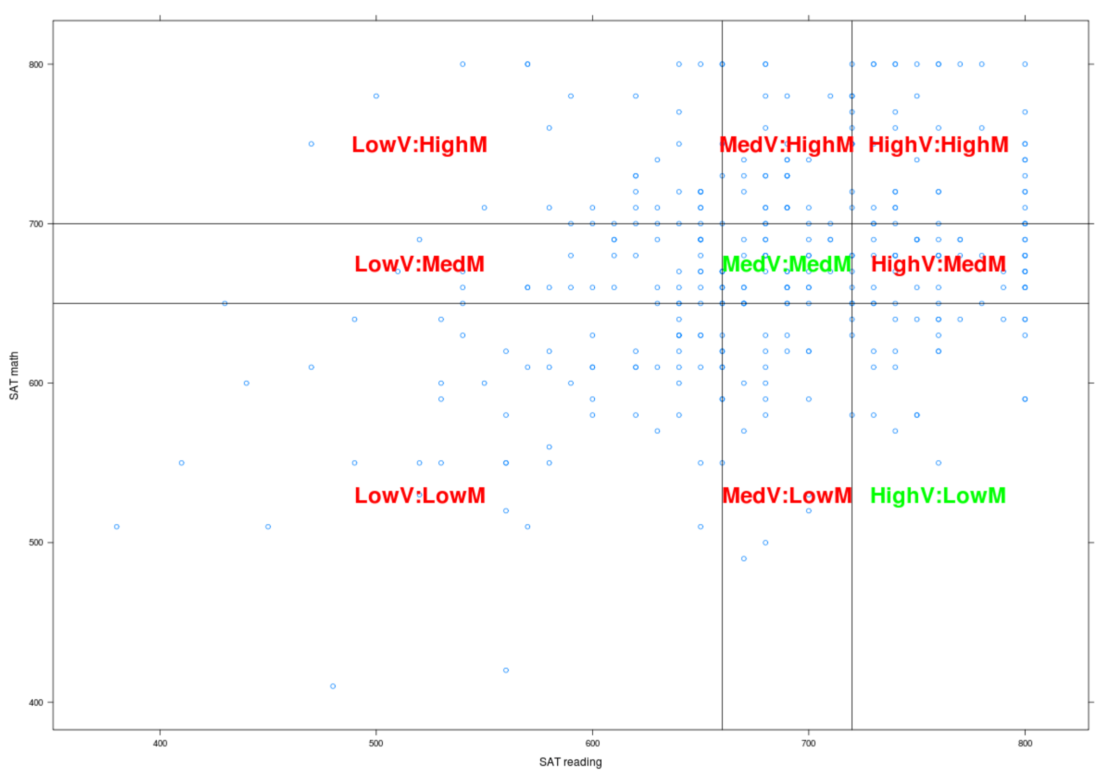

Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
400
500
600
700
800

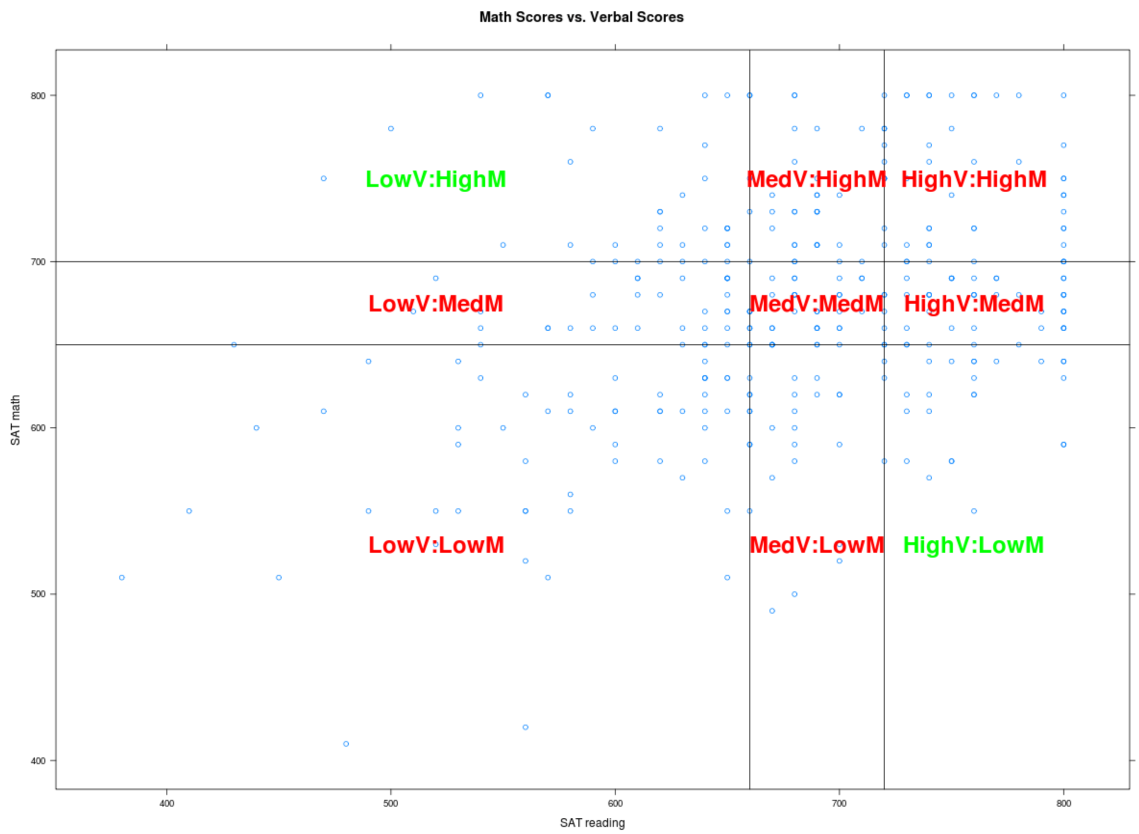

Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
400
500
600
700
800

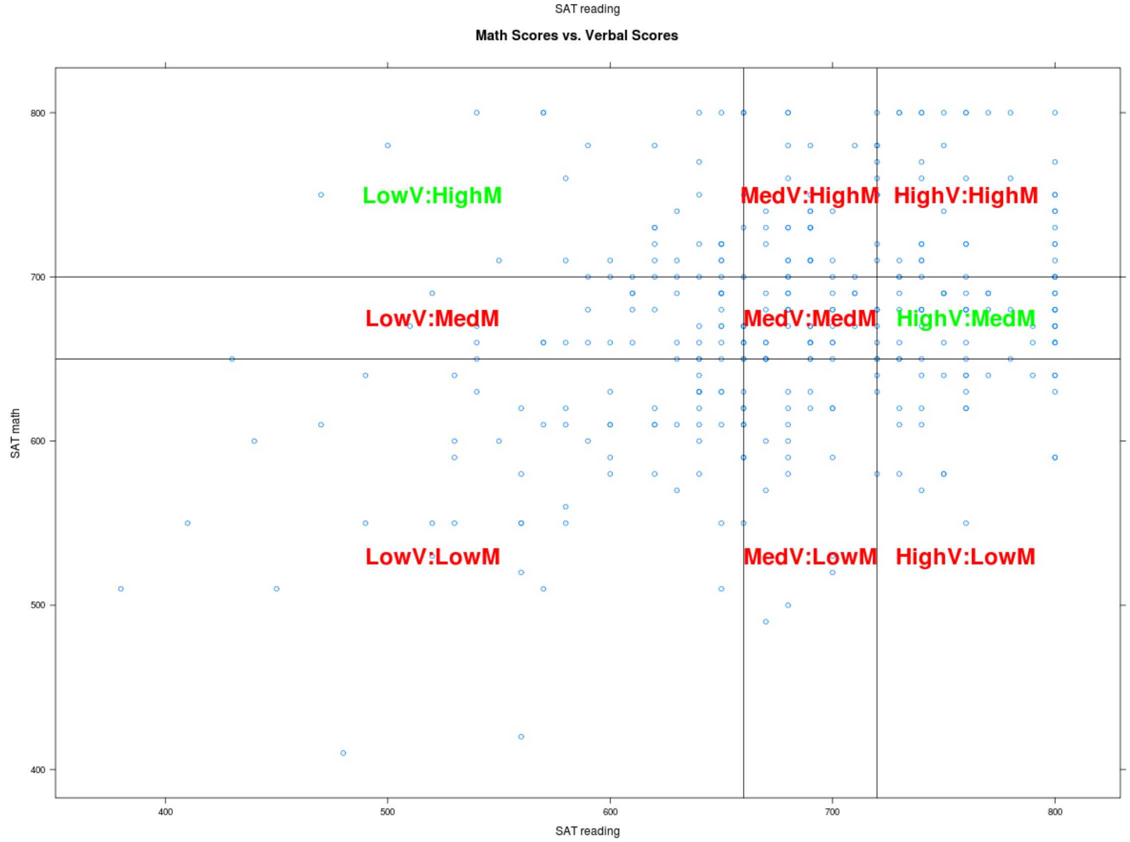
Math Scores vs. Verbal Scores

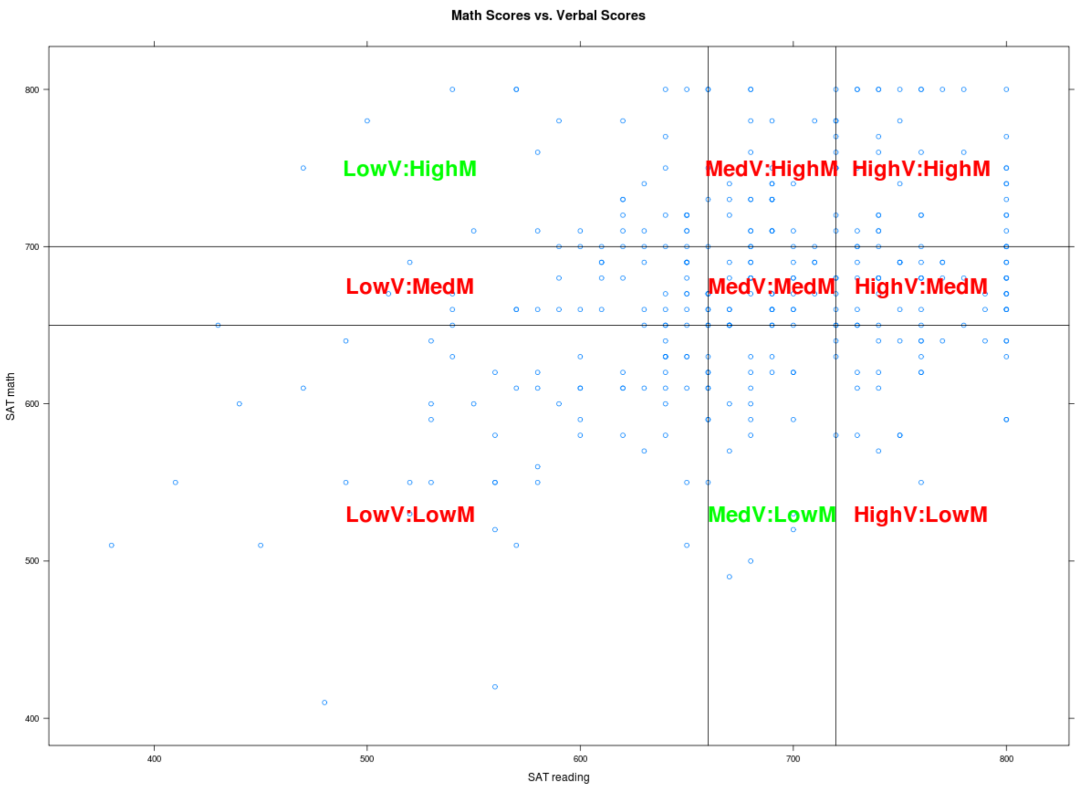


Math Scores vs. Verbal Scores




Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
400
500
600
700
800

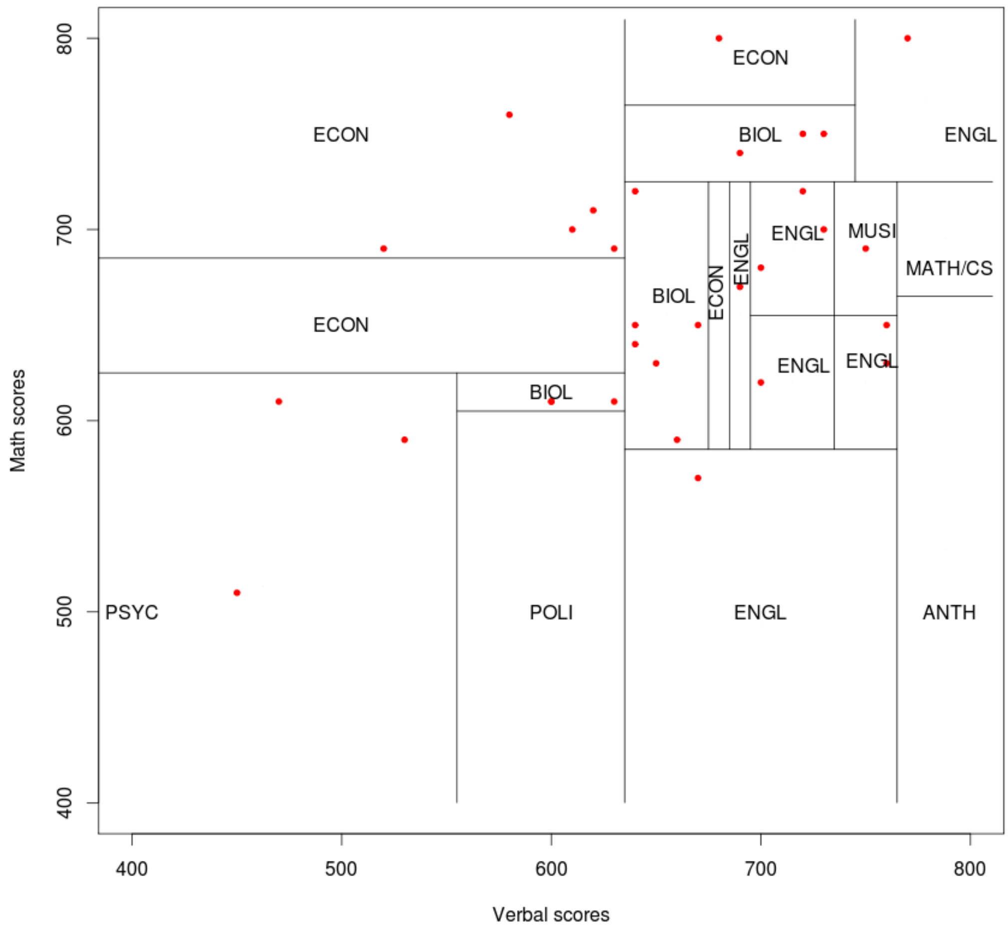

Math Scores vs. Verbal Scores
LowV:HighM
MedV:HighM
HighV:HighM
LowV:MedM
MedV:MedM
HighV:MedM
LowV:LowM
MedV:LowM
HighV:LowM
SAT math
SAT reading
400
500
600
700
800



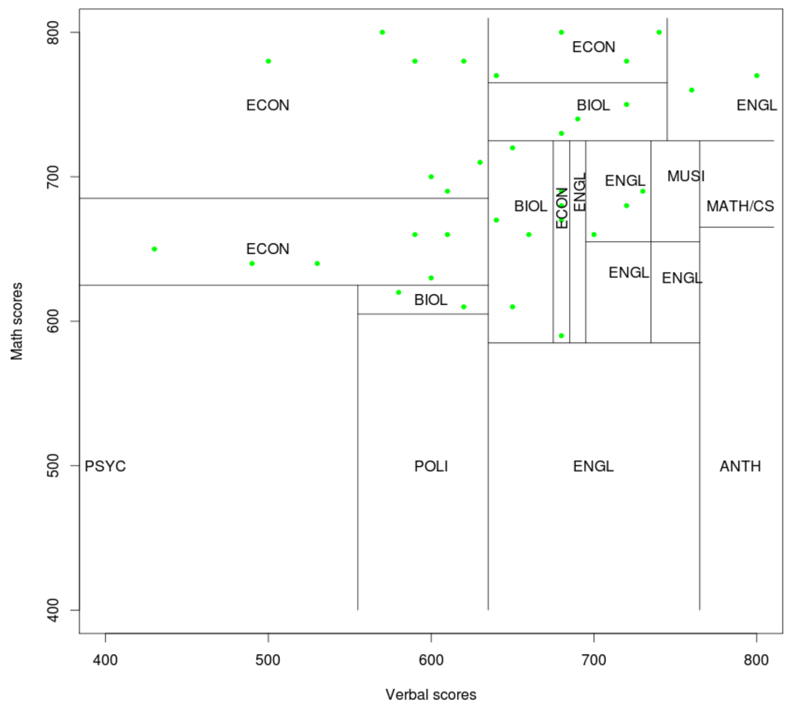
B.3: The 9 categories' division distributions

Category	LOWV: HIGHM	MEDV: HIGHM	HIGHV: HIGHM	LOWV: MEDM	MEDV: MEDM	HIGHV: MEDM	LOWV: LOWM	MEDV: LOWM	HIGHV: LOWM	Overall
Arts	5	2	3	1	4	6	6	3	3	33
Humanities	3	10	15	6	8	14	15	11	14	96
Natural Sciences	13	10	11	6	10	10	17	5	2	84
Social Sciences	10	12	5	14	10	10	31	10	9	111

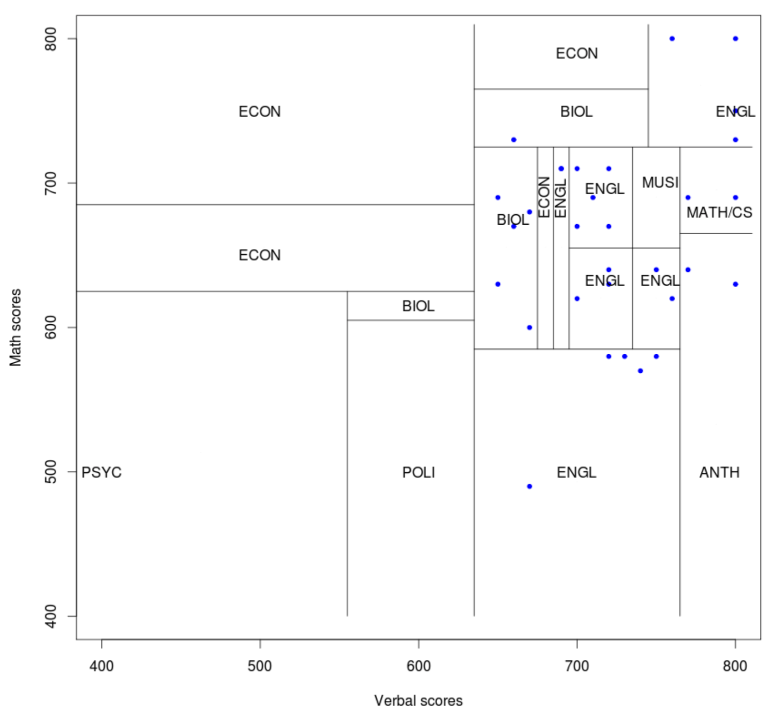
Appendix C: Plot of each major with partition



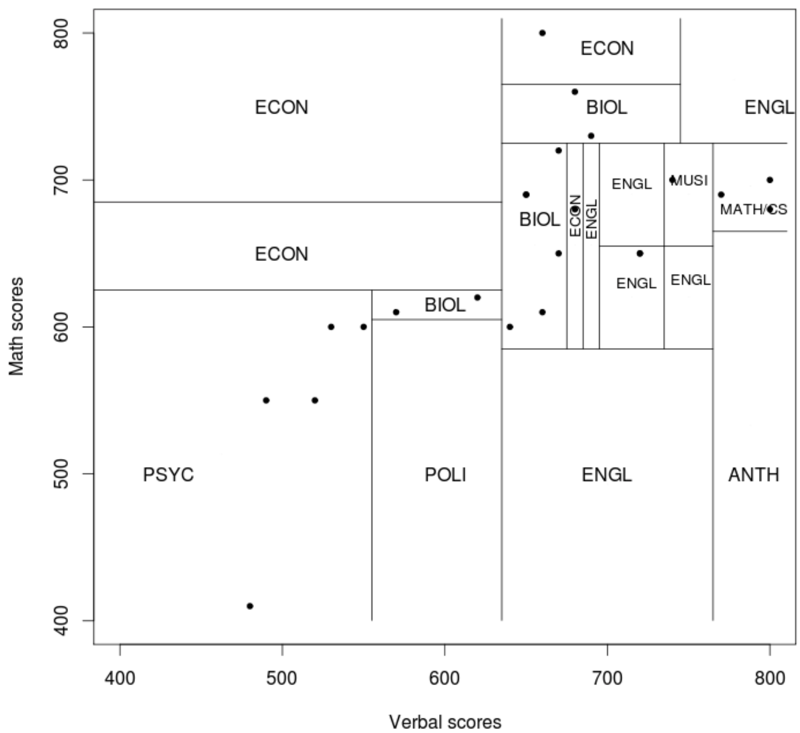
Biology



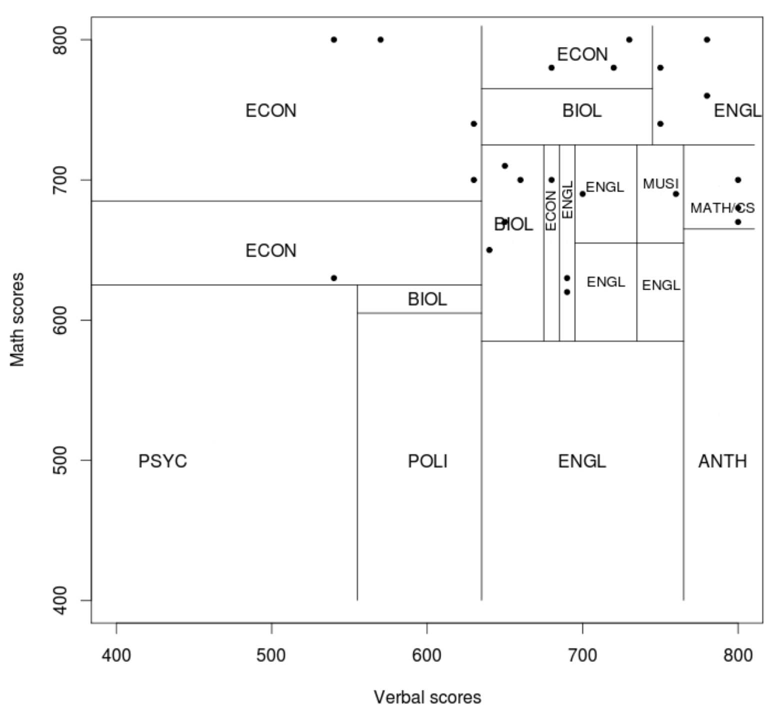
ECON



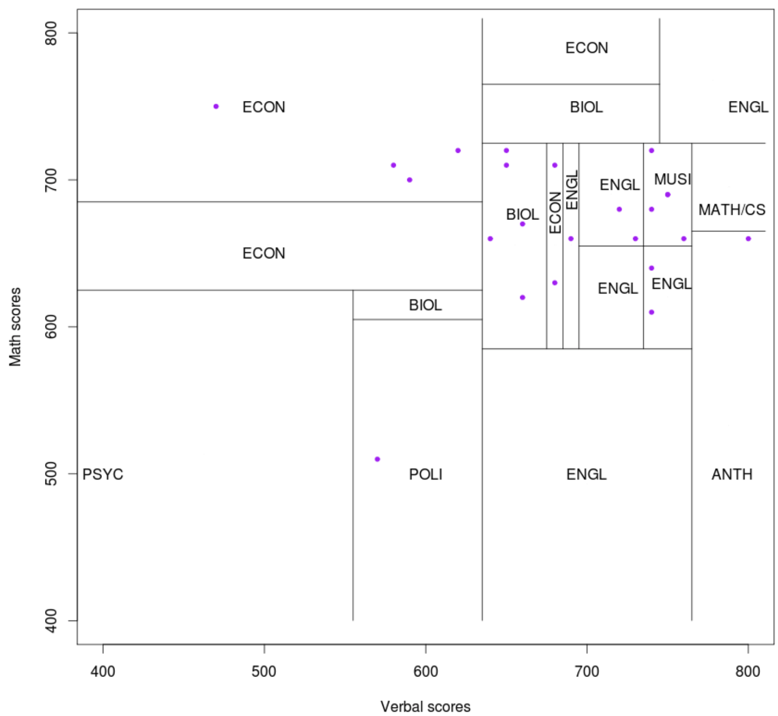
English



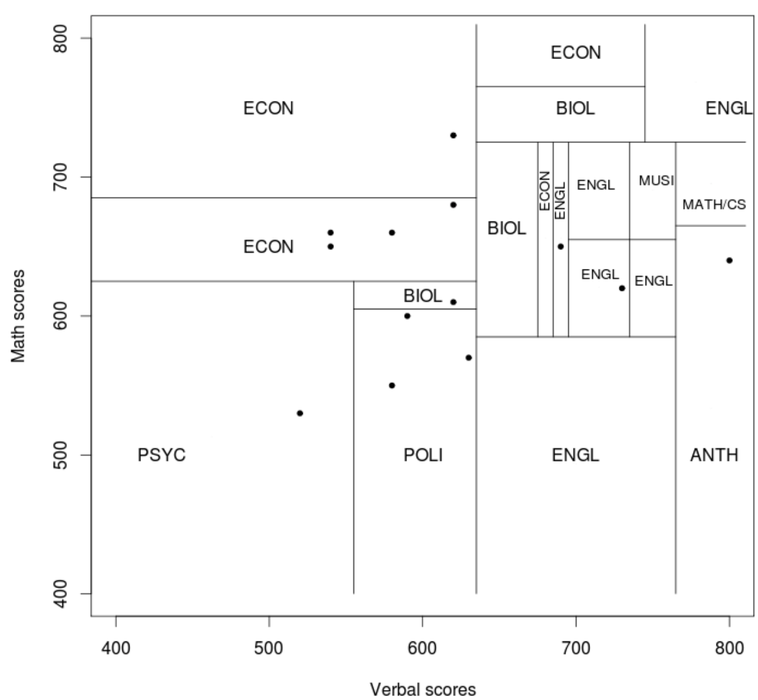
Psychology



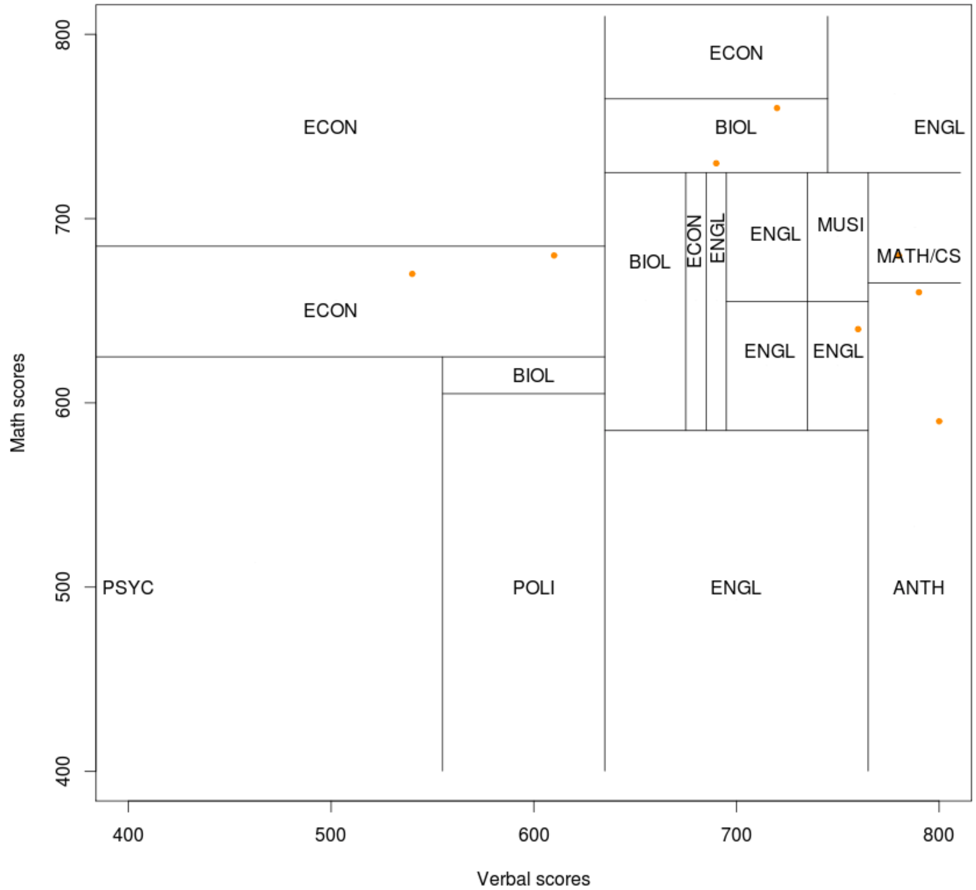
Math/CS



Music

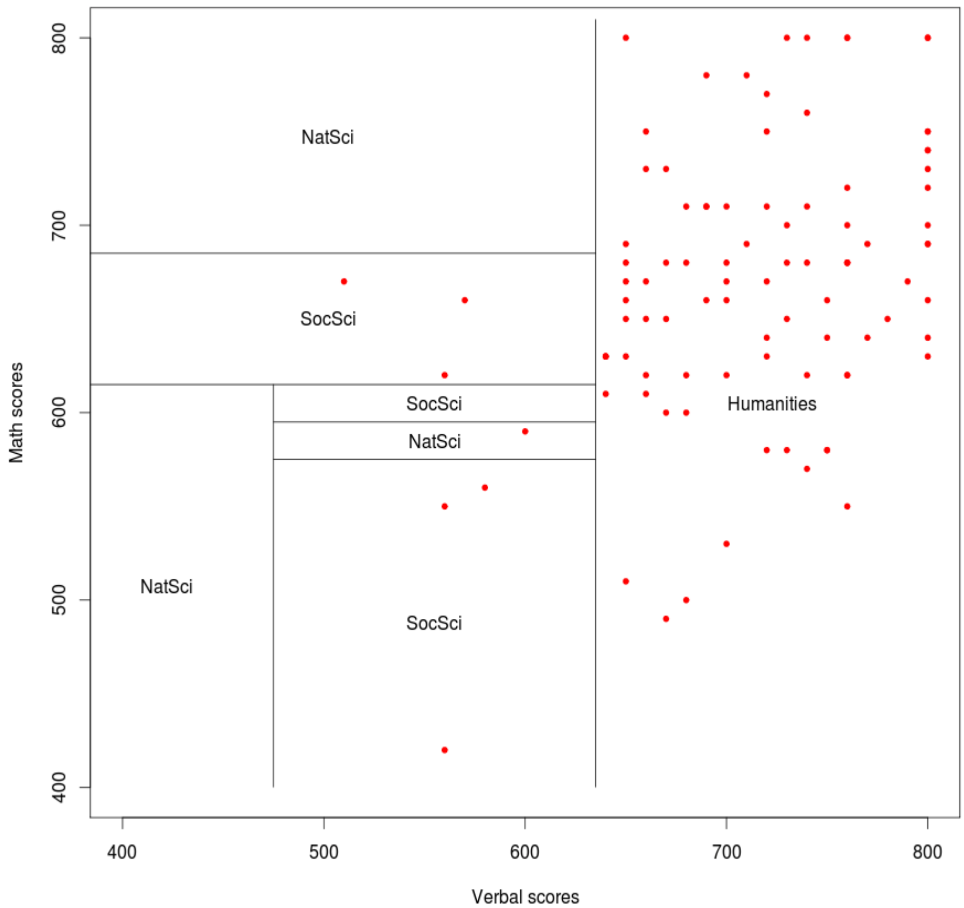


Political Science

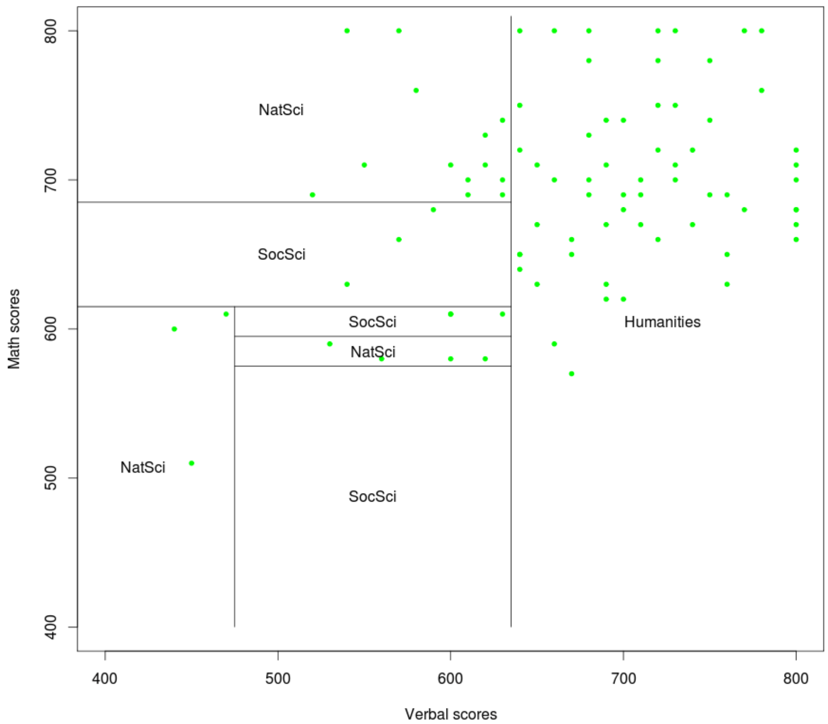


Anthropology

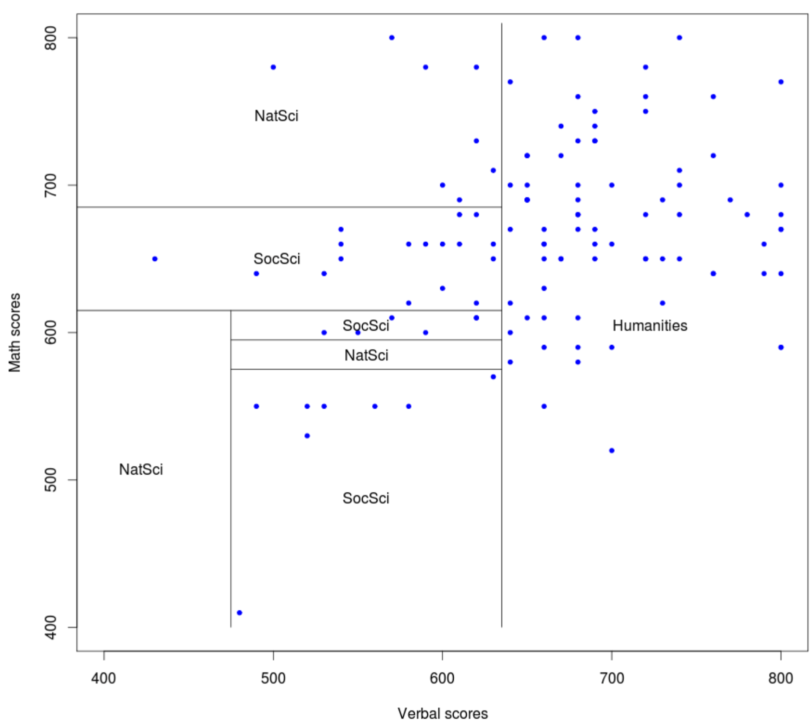
Appendix D: Plot of each division with partition



Humanities

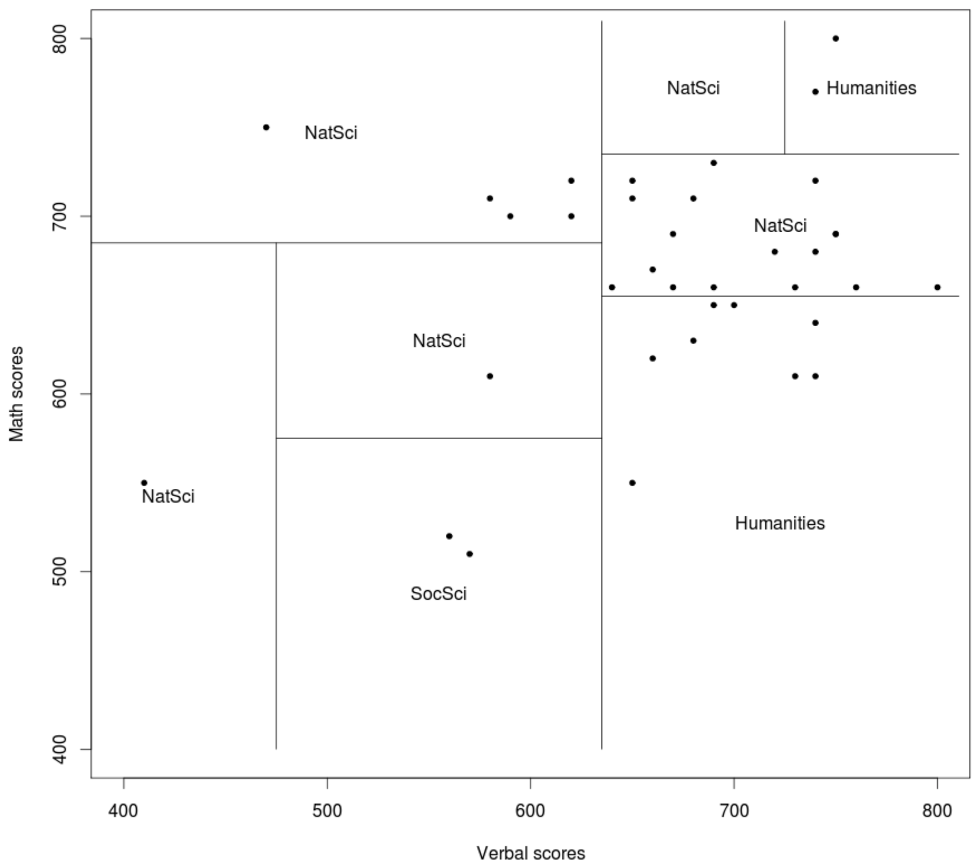


Natural Sciences and Mathematics

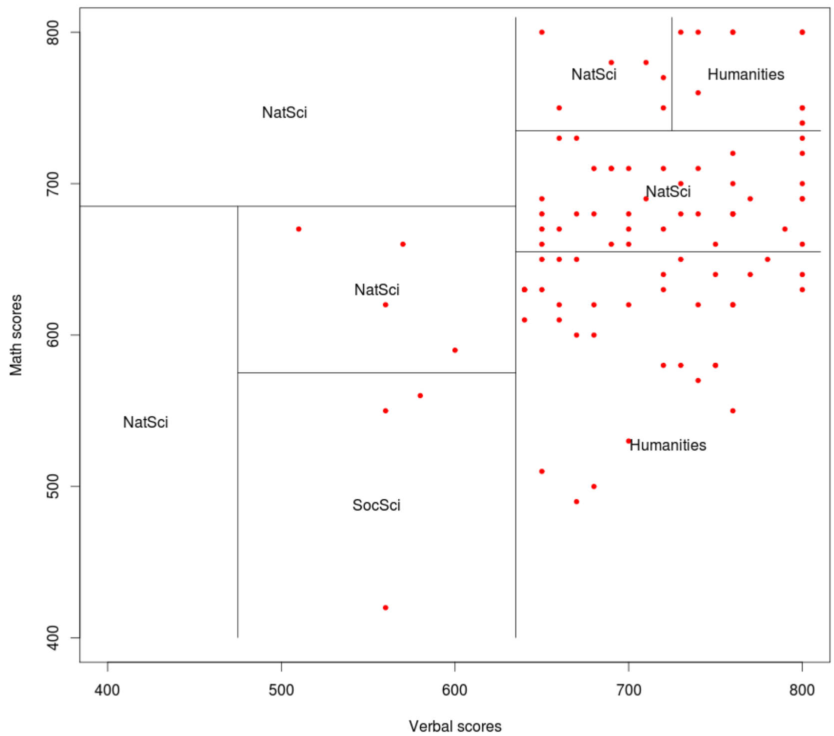


Social Sciences

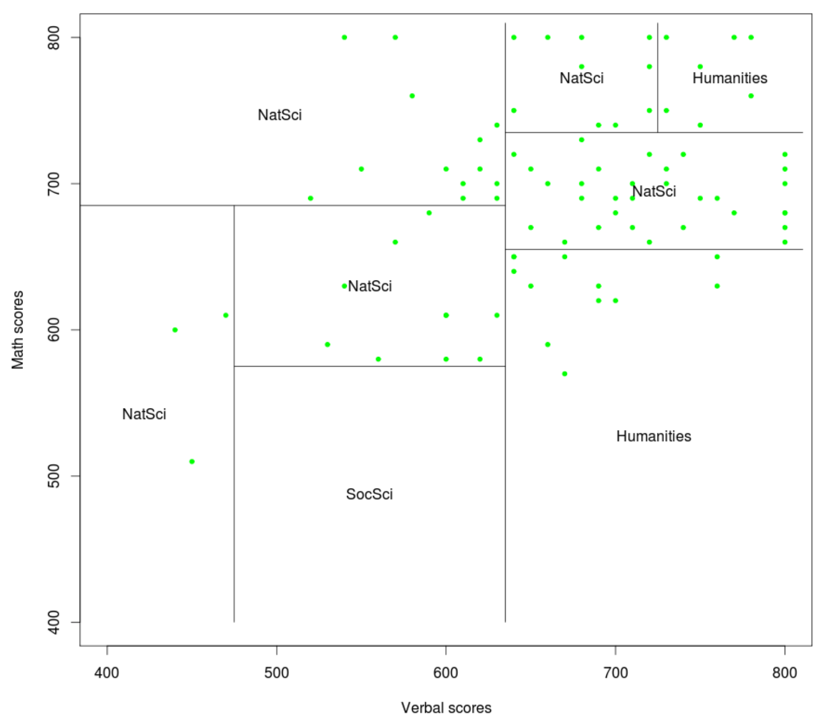
Appendix E: Plot of each division with partition when categorizing Economics into Natural Sciences division



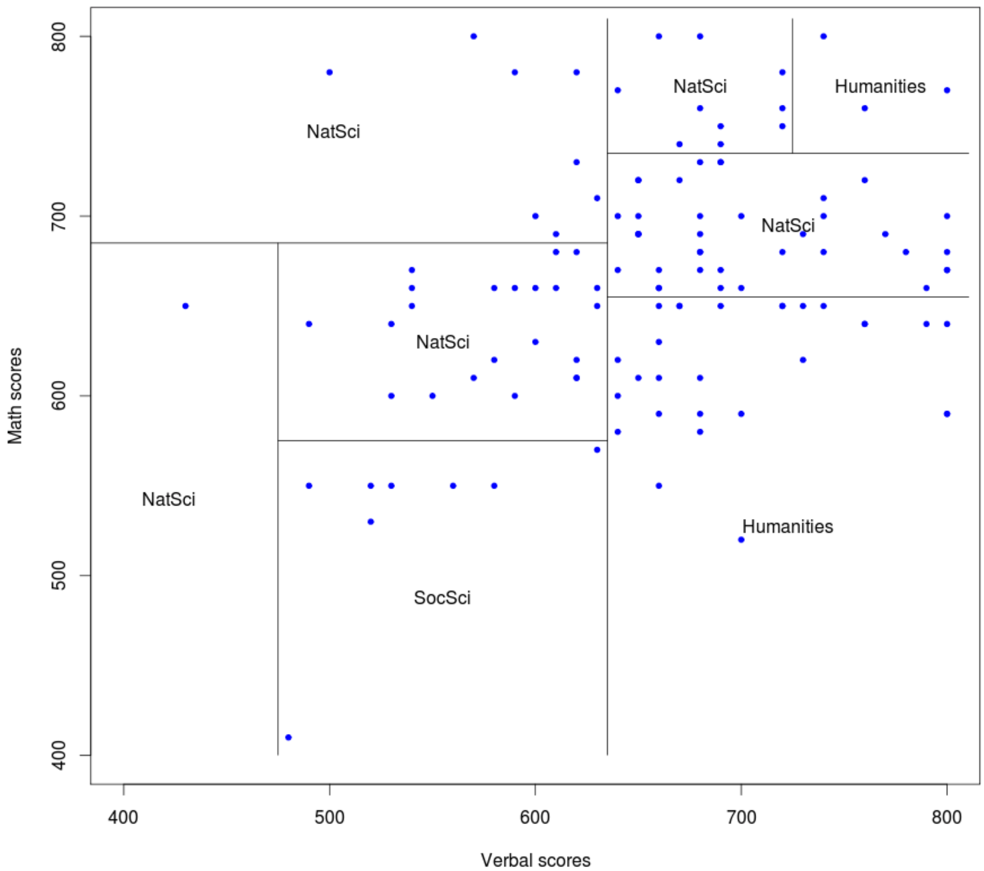
Arts



Humanities



Natural Sciences and Mathematics



Social Sciences