ABSTRACT

Title of dissertation: DETERMINANTS OF COLLEGE GRADE POINT AVERAGES

Paul Dean Bailey, Doctor of Philosophy, 2012

Dissertation directed by: Professor Judith Hellerstein
Department of Economics
Professor John Wallis
Department of Economics

Chapter 2: The Role of Class Difficulty in College Grade Point Averages.
Grade Point Averages (GPAs) are widely used as a measure of college students' ability. Low GPAs can remove a students from eligibility for scholarships, and even continued enrollment at a university. However, GPAs are determined not only by student ability but also by the difficulty of the classes the students take. When class difficulty is correlated with student ability, GPAs are biased estimates of students' abilities. Using a fixed effects model on eight years of transcript data from one university with one fixed effect for student ability and another for class difficulty, I decompose grades at the individual student-class level to find that GPAs are largely not biased. Eighty percent of the variation in GPAs is explained by student ability, while only three percent of the variation in GPAs is explained by class difficulty. This estimation is carried out using an ordered logit estimator to account for the ordered but non-cardinal nature of grades.

Chapter 3: Are Low Income Students Diamonds in the Rough?
Consider two students who earn the same SAT score, one from a lower-income household and the other from a higher-income household. Since educational expense is a normal good, the lower income student will, on average, have had a less well-resourced primary and secondary education. The lower income student may therefore be stronger than their higher income counterpart because they have earned an equally high SAT score despite a lower quality pre-collegiate environment. If this is the case, once the two students start attending the same college—and school spending becomes more similar—the lower income student's in-college performance should be relatively higher. I test this theory by using eight years of data from one university to compare the grade point averages of students from various family income levels. Results show that lower income students appear to be diamonds in

the rough: lower income students have surprisingly high outcomes, conditional on their SAT scores. However, unconditional on SAT score, the lower income students also outperform their higher income counterparts. This suggests that a single university's data is inappropriate for answering this question. I also develop how this type of regression might give insight into the production function of human capital. Specifically, a common assumption made in the economics of education literature is that first differenced human capital accumulation rates are independent of ability because ability is already represented in the test used as a base period. A "diamonds in the rough" result would contradict that assumption, and show that SAT is not a perfect measure of underlying ability.

Chapter 4: Estimation of Large Ordered Multinomial Models.

Decomposing grades data into class fixed effects and student fixed effects is difficult and the estimator's accuracy is unknown. I describe the successful application of the L-BFGS algorithm for fitting these data and propose a new convergence criterion. I also show that when the number of classes is about 32 (slightly fewer than is typical at the University of Maryland), the estimator performs well at estimating correlations and the non-parametric statistics used in Chapter 2 of this dissertation. Some issues with significance testing the sets of fixed effects are also considered and I show that when the number of classes is 32, the significance tests are not sufficiently protective against false rejection of the null hypothesis. The jackknifed likelihood ratio test is shown to be only modestly biased towards false rejection regardless of the number of classes per student.

DETERMINANTS OF COLLEGE GRADE POINT AVERAGES

by

Paul Dean Bailey

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Professor Judith Hellerstein, Co-Chair/Advisor
Professor John Wallis, Co-Chair/Advisor
Professor John Haltiwanger
Professor John Chao
Professor Paul Hanges

Dedication

This thesis is dedicated to my father, Dr. W. David Bailey.

Acknowledgments

I treasure the substantial support from the University of Maryland community and within my family that made this thesis possible.

I am forever indebted to John Wallis, who taught me how to identify and write an economics paper. In this formidable task, he was extremely generous with his time and patience.

I appreciate the assistance of Judith Hellerstein for showing me how a labor economist thinks about a paper. This is a skill for which I will always be grateful and hope to put to good use throughout my career.

I am also grateful for several important conversations with John Haltiwanger at key moments that provided useful guidance about which aspects of my research were actually interesting to an outside reader.

I also appreciate the time John Chao took with me in identifying state-of-the-art econometric techniques and providing a sounding board for my direction of inquiry for the fourth chapter.

Thanks are also due to Paul Hanges for serving on my thesis committee, taking the time to give my thesis a careful reading, and providing thoughtful responses.

I also appreciate the assistance of Kyland Howard and the Office of Institutional Research, Planning and Assessment at the University of Maryland who generously gave substantial time and energy when providing me with the data used in this thesis.

A special thanks is also due to Vickie Fletcher who was always happy to help

and watch out for me. She is an example of everything that a graduate studies coordinator can possibly be and so much more.

My colleagues and friends also provided help advice, support, and levity as the case required. I am indebted to Abby Alpert, Juan Bonilla-Angel, Teresa Fort, Carolina Gonzalez-Velosa, Aaron Szott, and many others.

A million thanks are due to my wife, Louise, who never failed to give me time and space to work on any part of my graduate degree and provided love and support without fail at every turn. I also appreciate the understanding of my son, Eli, and daughter, Keira, at times when I was busy working instead of enjoying their company.

Table of Contents

List of Figures

List of Abbreviations

BFGS	a numerical maximization / minimization algorithm
GPA	Grade Point Average
IRB	Institutional Review Board
L-BFGS	a limited memory variant of BFGS
OLS	Ordinary Least Squares
OME	Ordinal Multinomial Estimator
STEM	Science, Technology, Engineering, and Mathematics
OIRPA	Office of Institutional Research, Planning and Assessment

Chapter 1

Determinants of College GPAs

An enormous amount of human capital is developed in college with significant implications for the labor market. One of the most agreed upon empirical results in economics is that a year of education raises wages by about 7%, so that four years of college cumulatively represent a approximately 30% increase in a person's earning ability. Yet there is surprisingly little known about how to measure human capital development in college. The most obvious and typical measurement is the grade point average (GPA), an average of grades assigned in each class. Since students are enrolled in different classes and grading schemes may be subjective, it seems prudent to wonder whether GPAs are a good measure of human capital.

In addition, college is often thought of as a single "treatment" or a shared experience across students. If there are substantial differences in the average class difficulty across student ability levels, that would imply that college offers not one single treatment, but rather different treatments for different students. Intuitively, higher ability students might be expected to take more difficult classes with more rigorous grading schemes. One might also suspect that students in harder classes develop more human capital, so that high ability students trade off higher human capital accumulation for lower grades compared to what they would have earned in easier classes. This fits into a "pricing" model developed by Freeman (1999) where departments that imbue low human capital development pay for (attract) their students with high grades.[1] The marginal value (in terms of GPA) of a higher grade earned in an easier class drops for students near the top of the GPA distribution; students who typically get an 'C' or a 'B' have the possibility of getting an 'A', while students who typically get an 'A' have no possibility of getting a higher grade. The grade payoffs of easier classes are therefore of less value to these higher ability students.[2]

If each department wants students and can select the difficulty of its classes, and if students like easier classes, the Nash equilibrium is concerning: everyone gives out high grades. This type of behavior has even been observed (Eaton & Eswaran, 2008).

There are proposals to standardize post-collegiate testing to allow for a shared measure of college output. Similar to standardized testing at pre-collegiate levels, such testing would be common across schools, and therefore more objective. But this ignores that college is a time when students learn different things based on their coursework. A student who does not demonstrate significant improvement in a shared skill like critical thinking during college could have learned specific skills orthogonal to this. For example, knowing accounting standards need not increase one's critical thinking ability but may still be valuable human capital to acquire. The best test of what is taught in a class is the assessment of the faculty who know

[1] This model has been proposed by others as well (Drew, 2011).

[2] This model is like one that is often offered where STEM (science, technology, engineering and math) classes are socially desirable but fail to attract a sufficient number of students. As an example, President Obama has pushed for an increase in the number of STEM classes, presumably with the belief that the human capital developed in these classes is more beneficial to society than other classes.

the class best, and the only measure we have of this is the grade. Simply put, grades are the only quantitative measure of college performance that measure its diversity of subject material, and are therefore always an important outcome to include in analysis.

To make the most effective use of GPAs in research, one must ask: what information is carried in a GPA and how can it be used? To answer this question I decompose grades into a student fixed effect and a class fixed effect. This gives an estimate of student ability controlling for the classes the students is enrolled in and an estimate of class difficulty controlling for the ability of the students who enroll in the class.[3] Each student then has an estimated ability and it is easy to estimate the average of the difficulty of the classes the student enrolled in (the students' *average class difficulty*).

As described earlier, one might suspect that higher ability students are taking more difficult classes. One might also suspect that there is a grade "sweet spot" which could lead to students of all ability levels taking classes that they expect will lead to the same GPAs.

Chapter 2 of this dissertation measures the association between student ability and average class difficulty using correlations. These results show that student ability and average class difficulty are not strongly associated. A second set of correlations of student ability and class difficulty with GPAs shows that the correlation between average class difficulty and GPA are small and the correlations between

[3] Student ability might be better thought of as student output and class difficulty might be better thought of as anything that attenuates the function that maps the output of a student to the grade the student receives.

student ability and GPA are high. GPAs are driven by the ability of the student, not the difficulty of the classes they take.

The association of these three variables is also measured by pooling students with approximately equal ability levels and then graphing the average, within each group, of students' GPAs versus their average class difficulty. This gives a plot of the typical association between average class difficulty and GPA for each ability level and is allows any form of joint distribution and is more general than correlations, which assumes that the joint distribution is bivariate normal. These plots also show a weak association between student ability and class difficulty with most of the positive association between the two variables coming from the top and bottom of the ability distribution.

These results suggest that the basic intuition about grades being biased by student ability is wrong; higher ability students do not, in general, take significantly harder classes.

Based on the results of chapter 2, it is reasonable to wonder if an analysis of student performance should be using GPAs themselves or if more insight can be gained by using student fixed effects which net out average class difficulty. Average class difficulty is not the main driver of student grades, so it may simply be unimportant. The difference between GPAs and student fixed effects are important; GPAs are easier to calculate and are available on more datasets than the transcript data required to calculate student fixed effects.

Chapter 3 considers whether family income plays a role in students' GPAs, conditional on SAT scores. The question this chapter asks is: If two students earn the same SAT score, does the environment where they achieved that score matter? It could be that SAT fully captures the ability of these students to succeed in college. An alternative theory is that students from lower income families have been disadvantaged, had to work harder to achieve the same SAT, and will therefore outperform their higher income counterparts when they arrive at college and the playing field is leveled.[4] Running this regression on the University of Maryland data, I show that, conditional on SAT scores, higher income students receive lower grades than lower income students at the University of Maryland. However, on the same data, it is also true that unconditionally, higher income students receive lower grades that lower income students. This suggests that the result is hardly surprising—low income students at the University of Maryland are simply diamonds.

The analysis in chapter 3 is undertaken twice, once using GPAs themselves as a measure of student output, and again using the estimated ability level of each student from the decomposition of grades. Thus the analysis applies the question of the distinction between GPAs and student ability to a real world problem and answers the question of whether the student fixed effects give different results for an analysis of grades data. The result is that GPA and student ability give the same answer.

However, one specification in this chapter informs one of the larger questions of this dissertation. A regression that breaks SAT down into math and verbal components shows that a one point gain in SAT math is associated with a larger

[4] This thought is considered by Mankiw (2011) who suggests essentially the specification used in chapter 3. He later added a note from Stinebrickner who references his own article as saying the exact opposite occurs at a highly subsidized college (Stinebrickner & Stinebrickner, 2003).

increase in student ability than a one point increase in SAT verbal. At the same time a one point increase in SAT math is associated with the same increase in GPAs as a one point increase in SAT verbal. For example, given a student with a total SAT score of 1000, a breakdown of their SAT into math and verbal components is irrelevant to a prediction of their GPA, but that breakdown would help predict student ability. Another regression explains this–students who get higher SAT math scores balance out their higher student ability by taking approximately equally more difficult classes.

Chapter 4 addresses several econometric concerns regarding the decomposition of grades into student and class fixed effects. Two estimators are considered, ordered multinomial estimators (OME) and ordinary least squares (OLS). The OME has the advantage of describing the data generating process in a satisfying way, but the performance of the estimator is not well understood, while OLS does not describe the grading process well but should be accurate conditional on its assumptions. A simulation is used to show how these two estimators might perform when estimating the type of statistic used in chapter 2. The simulation shows that both estimators perform reasonably well.

Another issue is how to fit these fixed effects in a the non-linear model (the OME). The problem is that the typical methods of solving these models involve forming a huge matrix that must then be inverted. I apply the method of (Liu & Nocedal, 1989). This method has the advantage that it scales linearly in the number of regressors, both in terms of storage and computational intensity of an individual step.

Finally, chapter 4 considers the role of bias in estimating significance tests on the fixed effects. Both the jackknife estimator for significance tests and the traditional likelihood ratio test are shown to be reasonably accurate at performing the significance tests when the number of observations per student is large. When the number of observations per student is small, the jackknife remains reasonably accurate.

Many previous papers have suggested that ideally one would estimate the effect of class and student while controlling for the other (Grove & Wasserman, 2003; Eaton & Eswaran, 2008). This type of analysis has been undertaken before on a smaller dataset using a linear model for grades (Arcidiacono et al., 2011), but the relationships of student ability and class difficulty were not the focus. This dissertation describes the results of such an analysis, shows that the analysis is accurately estimated, describes the relationship between the class and student fixed effects, and uses the student ability and class difficulty to investigate a question. The results show that class difficulty matters less than one might think, but the intuition that better students take harder classes was not entirely false. Students with relatively higher *math* scores do take harder classes, though the same does not hold for relatively higher *verbal* scores.

Chapter 2

The Role of Class Difficulty in College Grade Point Averages

2.1 Introduction

Grade point averages (GPAs) are often used as a measure of college students' ability, by the university awarding them and by others judging a student's performance while at the university. GPAs, however, are determined not only by student ability but also by the difficulty of the classes the student takes, calling into question its status as a measure of ability. While there are many ways that student ability and class difficulty could be related, one that makes intuitive sense is that high ability students take harder classes than low ability students, attenuating their GPA gains. In principle, higher-ability students could even take such difficult classes that their GPAs would be lower than those of lower-ability students. Under those circumstances, GPA would be an extremely misleading measure of ability. Determining the role of class difficulty in college GPAs is therefore crucial to assessing the use of GPAs as a measure of ability.

To find the answer to this question, I estimate a model with one set of fixed effects for students and another set for classes using a large dataset of eight years of transcript data from the University of Maryland.[1] Using correlation coefficients

[1] I apply the term "ability" to student fixed effects and "difficulty" to class fixed effects. Investigating the exact nature of these variables is not a topic of this paper, so they are used despite the semantic ambiguity. What I am calling "ability" might be a product of some innate capacity to learn as well as student effort. Similarly, class difficulty could also be a measure of the quality

on the fixed effects, I come to the perhaps surprising result that GPAs *are* a good measure of student ability. In fact, class difficulty and student ability are only slightly correlated. Moreover, student ability explains 80% of the variation in GPA while the average difficulty of classes in which a student is enrolled explains only 3% of the variation in GPAs.

This result is unexpected because previous research showed that average grade varies by department at other universities and colleges (Sabot & Wakeman-Linn, 1991; Freeman, 1999; Eaton & Eswaran, 2008), a finding I reproduce for the University of Maryland datat that I have (Figure 2.2). At the same time, Salbot and Wakeman-Linn show that average SAT score does not vary by department. These two facts suggest that there is both variation in each classes' class difficulty as well as variation in each student's *average class difficulty* (the difficulty of classes a student is enrolled in averaged over a student's transcript). I am able to quantify the extend of this and find that variation in individual grades is approximately equally predicted by the difficulty of the class as the ability of the student. However, for GPA to be a biased estimate of student ability, it must also be the case that students' average class difficulty is correlated with the students' ability—for example, if higher-ability students are systematically enrolled in more difficult classes. I find that, surprisingly, this third condition is not present at the University of Maryland and, for this reason, GPAs are not very biased by student ability.

I account for grades being a limited dependent variable–an 'A' is better than a 'B', but how much better is not obvious–by using an ordered logit estimator. The

of pedagogy in each class–more "difficult" classes would simply have poorer instruction.

consistency/bias properties of the ordinal logit estimators are not well known for a fixed effects model, so chapter 4 of this dissertation explores how these estimators perform with grades data using simulations. The simulations show that the ordinal multinomial estimators do a good job of estimating the statistics used in the results of this paper and perform better than the OLS estimator.

Fitting such a fixed effects model requires large amounts of data because a greater number of observations per student and class improves the estimator's performance. Because of this I chose a lengthy time frame of administrative data from a large state university. In the fitted regression, an observation is a grade that a student receives in a particular class. There are five hundred thousand observations over eight years and nine thousand students with, on average, forty classes on their transcript.

Because the correlations used to arrive at the main result assume a joint distribution that is bivariate normal between the two variables being compared, I develop the concept of an *ability expansion path* to describe the relationship between class difficulty and student ability more thoroughly. Specifically, I group students by estimated ability level and then plot their average GPA versus their estimated average class difficulty (Figure 2.1 provides an example).

The ability expansion path graphically depicts any systematic differences between higher and lower ability students with regard to class difficulty. For example, if higher and lower ability students take classes of the same difficulty level, the expansion path is vertical and GPAs are not biased (expansion path "a" in Figure 2.1). In contrast, if higher ability students systemically enroll in harder classes, differences in GPAs underrepresent the differences in ability and the expansion path is positively sloped (expansion path "b"). When high ability students enroll in classes that are even harder still, their increase in ability is entirely masked in GPAs and the expansion path is horizontal (expansion path "c").

I observe an estimated expansion path that is largely vertical, with some deviations from this at the top and bottom of the ability distribution. This is consistent with my general finding that GPAs are not very biased. At the top of the distribution, GPAs are positively correlated with average class difficulty. Specifically, the top students (those with GPA$>$ 3.8) are systematically enrolled in harder classes. At the bottom of the distribution (students with GPAs$<$ 2.5), GPAs are again positively correlated with average class difficulty when looking at students who are on a path to graduate, but are uncorrelated when including students who are dropouts. Thus, for graduates, the positive correlation at the bottom of the ability distribution is induced by the university's GPA minimum, which eliminates those lower-ability students who were enrolled in more difficult classes at entrance.

Knowing the role of class difficulty in grades is important because low GPAs can remove a student from eligibility for scholarships and even continued enrollment at a university; GPAs are also used for graduate admissions and potentially influence labor market outcomes (Loury & Garman, 1995; Jones & Jackson, 1990). While one might initially suspect that the importance of GPAs represent a misplaced faith in their value as a measure of ability, there is an irreducible reason to use this less than transparent measure of ability–there is no substitute quantitative measure of ability in college. Because of this, and despite the real possibility that grades may be biased, they are frequently used as outcomes (Angrist et al., 2009; Klopfenstein & Thomas, 2009; DeSimone, 2008; Betts & Morell, 1998) and as regressors (Loury & Garman, 1995; Jones & Jackson, 1990).[2] In primary and secondary schools, standardized tests are specifically designed to provide a comparison external to grades. However, the diversity of students' educational experiences at college prevents this type of comparison–what standardized test could capture the material learned by a student majoring in computer science and simultaneously capture the material learned by a student majoring in music? Therefore, despite their flaws, grades remain the best available summary measure of student ability. Understanding the factors that influence GPAs is critical to their effective use both in research and practical applications.

The following section of this paper describes challenges with using correlation coefficients for discrete data, such as student grades. Sections three and four describe the data and the results. A final section concludes.

2.2 Correlations

Grade performance is described by a student ability contribution (a) and a class difficulty contribution (d). Typically, when reporting regression results, the main results of interest involve a small number of estimated regression coefficients. In the case of a fixed effect model there are thousands of regression coefficients and interpretation of all coefficients is neither possible nor desirable. The impor-

[2] In some of these papers, dropout rate is another important measure of college performance, but it is only useful when looking at students close to the margin of dropout.

tance of the regression coefficients in determining grades is measured instead by the correlation between the fixed effects estimates and the grades.

The Pearson correlation coefficient is not an appropriate measure of the correlation between student performance (a), class grading difficulty (d), and grades (G), The Goodman-Kruskal gamma is appropriate for reasons that are described in this section.

The Pearson correlation coefficient for two variables (A and B) has three salient properties that give it currency as a statistic: it is invariant to scale so that $\mathrm{Cor}(A, cB) = \mathrm{Cor}(A, B)$ for any scalar c; second it is invariant to location so that $\mathrm{Cor}(A + c, B + d) = \mathrm{Cor}(A, B)$; third, it takes on values in $[-1, 1]$ with zero indicating no covariance between A and B and larger absolute value numbers indicating stronger covariance. In the extreme, a value of 1 or -1 indicates a linear relationship of the form

$$A = c_0 + c_1 B. \tag{2.1}$$

For discrete data, which can be represented on a two-way table, the Pearson type properties cannot be preserved. Other properties are more appropriate for discrete data. For grades, operations like multiplying or adding an arbitrary constant to the data do not make sense,[3] so the first two properties of the Pearson correlation coefficient are not relevant. However, the concept that a correlation should be able to take on values on all of the values in $[-1, 1]$, does remain sensible and is maintained

[3] Adding one to a 'B' might appear to make sense, but adding π to 'B' is more difficult to interpret.

by some rank correlation coefficients. Additionally, a correlation coefficient of 1 is associated not with a linear relationship but instead with a weakly-monotonic relationship (Ghent, 1984).

Rank type correlations can be represented as operating on a two-way table so that there are cells potentially above, below, to the right, and left as well as diagonals from any given cell (X), i.e.

	variable A		
	...	...	...
variable B	...	X	...
	...	...	...

For data tabled this way, the Goodman-Kruskal statistic uses all possible pairs of data and defines *concordance* between a reference cell value and another value when, relative to the reference cell, it is above and to the right or down and to the left. These relationships suggest A and B are associated because larger values of A are associated with larger values of B. *Discordance* is the name given to cells above and to the left or below and to the right. These relationships suggest A and B covary negatively. The table below shows these two types of cells labeled C and D for concordance and discordance with cell X, respectively for grade data and SAT scores.

			Grade	
		F	B	A
	high	D	. . .	C
SAT score	mid	. . .	X	. . .
	low	C	. . .	D

The remaining relationship is those cells that are in the same row or column as X and are not included in the calculation of the Goodman-Kruskal statistic because they are uninformative. They are ties where the data provide no information.

Using these definitions, the Goodman-Kruskal correlation coefficient is

$$\tau - G = \frac{C - D}{C + D}. \tag{2.2}$$

In the case of two variables A and B being bivariate normally distributed, when the Pearson correlation coefficient is c, then A can be said to explain c^2 of the variation in B. This results from a maximum sum of squares for Pearson correlations of a multivariate normal distribution being one. This means that if $\mathrm{Cor}(A, B) = c$ and $\mathrm{Cor}(B, C) = 0$ then the largest possible value of $\mathrm{Cor}(A, C)$ is $\sqrt{1 - c^2}$. However, for rank correlation coefficients, there is no such hard and fast rule and thus the correlation coefficients cannot be interpreted with the same “percent explained” interpretation.

2.3 Data

The primary data used in this paper is transcript data from the University of Maryland from the years 2003 to 2010. Each individual observation consists of a class identifier, an anonymized student identifier, and the grade the student received in that class. In addition, students' application information is linked and used for baseline characteristics. These variables are summarized Table 2.1.

The data were provided by the the University's Institutional Research Planning and Assessment group as two files, one with transcript entries and another with application information.[4] Because the data were used to produce transcripts and for applications, it arrived relatively "clean" with few missing values. The transcript file itself would be sufficient for constructing most of the variables necessary to run the regressions in this paper. However, some cleaning was necessary. More information on the data is given in the Data Appendix.

Some of the variables included in the results for an individual student are derived from more than one transcript entry. Information was aggregated to the following levels:

Individual class level this is the transcript (one student in a single class who receives a grade) and represents no aggregation; regressions were run at this level;

Semester level the total number of courses completed to date, and number of credits

[4] In compliance with our IRB application, student identification numbers were scrubbed from the file, but a new version that was not related to the actual student identification number was placed on the new file so that a transcript could be built. For the purpose of this paper, these new identification numbers serve all of the purposes of a student identification number.

enrolled was calculated by aggregating data to the semester level;

Student level the GPA and demographic information from the application were calculated by aggregating each student's entire transcript.

Students who receive GPAs of exactly 4.0 or very low GPAs are excluded from analysis. Students receiving 4.0s represent about 100 total observations in the sample, and low GPAs are those where the student never earned more than a 'C'. This is done because the simulation results (section 4.1) show that the estimated statistics are unchanged or improve when boundary cases are removed. A few percent of the students meet these criteria, almost all of them because they never passed any classes.

In addition, internships and classes in which fewer than five students enrolled over the seven years studied are not included–these classes are not shared experiences in the same way that most other classes are. This restriction represents a small number of total credits, but a substantial number of classes.[5]

The remaining data used is filtered to include only those students who are between ages 18 and 25 when they entered college, started in 2003 to 2005, were matriculated at some point,[6] and stayed enrolled for at least eight semesters or completed 120 credits (the minimum for graduation) by 2010. I call this the "degree" sample ($n = 9,410$) because these are the students who completed or are likely on the road to completing a degree at the University. The entry and exit profile for

[5] When these classes are dropped, GPAs calculated based on the remaining classes need not meet the minimum for graduation, even for graduating students. Some of the figures include students who apparently received lower than 2.0 GPAs, but this is just for the non-excluded classes.

[6] *Matriculated* is the status of a student who was enrolled as a degree-seeking student.

these students is shown in Table 2.2.

For robustness checks, two expanded samples are considered. One is all students who entered between 2003 and 2005, as long as they completed at least five classes since enrolling and were matriculated at some point. This sample is called the “enter” sample ($n = 18,400$) because this is the entering cohort of the “degree” sample. The robustness of the results to the inclusion of students that eventually left the program is tested using this sample. The entry and exit profile for students in the “enter” sample is shown in Table 2.3.

The final expanded sample includes all students who have at least five transcript entries, regardless of when (within 2003-2010) they entered the university or whether they matriculated. This is called the “full” sample because it contains almost all of the students at the university during this time range.

The “degree” and “enter” samples are similar in their covariates. Of those who took the SAT, the “degree” sample’s average was higher by only seven points on the verbal and seven points on the math test. This represents an approximately two percentile difference in test scores. Relative to the “degree” sample, the “enter” sample had a lower average GPA by 0.14, five percentage points fewer took the SAT test, twelve percentage points more transferred. By definition, every person in both the “degree” and “enter” samples was a matriculated student.

The “full” sample includes students who did not have sufficient time to complete a degree because they enrolled in the last 5 years. The only qualification to be in the sample is to have entered the University of Maryland, but the time range is longer and the potential tenure shorter. Because of that, they more closely resemble the "enter" sample but have completed fewer classes and credits.

Histograms show the distribution of average grade points for departments, students, and classes. For each of these

$$\bar{GP} = \frac{1}{n_i} \sum GP_i$$

where GP_i is the grade points for grade i associated with the unit (department/student/class), and are taken from the typical GPA calculation ('A' = 4.0, 'B' = 3.0, etc.), and n_i is the total number of observations associated with the unit.

Previous studies have found substantial variability in average grade by department (Sabot & Wakeman-Linn, 1991; Freeman, 1999; Eaton & Eswaran, 2008); this is also the case with this sample (Figure 2.2). In addition, student and class average grades show substantial and similar variation (Figures 2.3 and 2.4, respectively). Note that peak is nearly flat over an entire grade point.

2.4 Results

The decomposition of grades into a set of fixed effects for student and another set for classes is first estimated with the OME, with no additional regressors, on the "degree" sample using the equation

$$y^*_{ij} = a_i - d_j + \epsilon_{ij} \ . \tag{2.3}$$

Where i indexes students, j indexes classes, and each class in which a student receives a grade is an observation. Robustness of the results is then tested to the addition of time dependent student specific regressors, the estimator used (ordered logit, probit, and OLS), and the selection criteria for the sample.

2.4.1 Decomposing grades using the ordered multinomial estimator

Using the fitted values from the main regression equation (eq. 2.3), correlations are then calculated between the fitted ability (a_i), GPA, and a student's average class difficulty ($\bar{d_i}$), where a student's average class difficulty is simply the average of the class difficulties of the classes in which they are enrolled. This is given only the third by

$$\bar{d_i} = \frac{1}{n_i} \sum_{j \in \{i' \text{s classes}\}} d_j \ , \tag{2.4}$$

where the set $\{i$'s classes$\}$ is the classes in which student i is enrolled, and n_i is the number of classes in which the ith student is enrolled.

GPAs are only slightly biased. The correlation between student ability and class difficulty is only 0.16. This is a small correlation, student ability and class difficulty are only slightly related. GPA and class difficulty are also not strongly related with a correlation coefficient -0.19; squaring the correlation coefficient gives 0.036, or 3.6% of the variation in GPA is explained by class difficulty. Higher ability students are taking only slightly harder classes. The main driver of GPAs is ability, with a correlation coefficient of 0.90, meaning that about 80% of the variation in

GPA is explained by student ability alone (Table 2.4).

The ordered multinomial estimator generates “intercepts” for each grade boundary. These are useful for interpreting the magnitude of the effects shown above, which are on a logistic scale. These values show that the “width” of a grade increases with the grade, so that ‘B’ is 2.47 units wide (Table 2.5, far right column) and encompasses a larger range of observed ability than ‘C’ (1.65 units wide) or ‘D’ (0.59 units wide). These can be used to judge the magnitude of values in these units. For example, a ‘B’ grade is about 2.47 wide in these units, so a student whose ability places them at the lower edge of ‘B’s (receiving about 1/2 grades ‘B’ or higher and 1/2 grades ‘C’ and lower) has 2.47 lower ability *ceteris paribus* than a student whose ability places them at the upper edge of ‘B’s (receiving about 1/2 ‘A’s and 1/2 grades ‘B’ and lower). Because grades have a maximum and minimum value, the GPA of these students would be expected to be less than 1.0 apart.

2.4.2 The ability expansion path

The correlation coefficients between student ability, GPA, and average class difficulty are parametric tests and assume that the two variables being correlated are bivariate normally distributed. The ability expansion path makes no such assumption and shows the relationship between all three of these variables most clearly (Figure 2.5). This plot shows that for the vast majority of students with GPAs between about 2.5 and 3.8, the ability expansion path is essentially vertical and there is no GPA bias. However, for the highest ability students (those with GPA

> 3.9), the average class difficulty is substantially harder, and for students with GPAs below about 2.5, GPA and average class difficulty are positively associated. This suggests that what little bias there is in GPAs is focused in these two groups.

To graphically demonstrate the tradeoff between grades and difficulty level, each student's individual estimated average difficulty and observed GPA are plotted in 20 color bands where each band represents students with approximately similar fitted values of a_i (Figure 2.6). This shows that students with a common level of student ability (sharing a single color) are, in fact, trading off between difficult classes and higher grades.

Having described the properties of the fitted a_i and d_j regressors, it is important to know that the estimated coefficients are statistically significant. The LR test for the fixed effects in eq. 2.3 shows that addition of class and students fixed effect in any possible order is highly significant (Table 2.6).

2.4.3 Variation in class difficulty

Earlier in this chapter I mentioned that there are three conditions for grades to be biased: (1) there must be variation in class difficulty, (2) there must be variation in average class difficulty, and (3) the variation in average class difficulty must be correlated with student ability. The results so far raise the question of which of the three conditions for bias in grades actually hold, since grades are only very slightly biased. A skeptic might wonder if the graph of average grade points by department was misleading and was actually the result of random variation and not fundamental dispersion in grades. But this is not the case. The graph gives an accurate impression, because only the third condition for bias is not present; student ability and average class difficulty are not correlated. This occurs even though there is substantial variation in class difficulty and average class difficulty.

The first condition holds. There is substantial variation in difficulty across classes. The standard deviation of class fixed effects (d_j) is 1.79 (Table 2.7, column A), or about 3/4 of the width of the 'B' range. The standard deviation of student fixed effects is approximately equal at 1.76. Both of these ways of interpreting the variation in class difficulty show that the variation in class difficulty is large.

Another way to look at the variation in class difficulty is to calculate raw correlations between class difficulty, student ability, and grades awarded in each class. The Goodman-Kruskal correlation coefficient between grades and student fixed effect (τ-$G_{g,a}$) is 0.43 while the correlation coefficient between grades and class fixed effect (τ-$G_{g,d}$) is 0.34 (Table 2.8).[7] These two coefficients are of approximately the same magnitude, meaning that the ability of the student taking a class and the difficulty of the class are approximately equally important in predicting the awarded grade.

The substantial variation in class fixed effect and the relative size of the Goodman-Kruskal correlation coefficient between classes and grades shows that classes vary in difficulty about as much or slightly less than students vary in ability.

[7] The remaining correlation coefficient τ-$G_{a,d} = -0.06$ simply mirrors the low Pearson correlation coefficient between average class difficulty and grade—ability and class difficulty are not strongly related.

Results from these ways of viewing the difficulty show that the first condition is readily met; students have a very wide selection of class difficulty levels to chose from, and do so.

The second condition also holds. There is substantial variation in students' average class difficulty ($\bar{d}_i$). The observed standard deviation in average class difficulty ($\sigma_{\bar{d}_i}$) is 0.66. Because it is an average, the standard deviation of $\bar{d}_i$ has to be smaller than the standard deviation of d_j, but how much smaller is an empirical question.[8] One way to put this value into context is to compare it to what would happen if the values of d_j were random draws from the available classes. If this were the case, then the central limit theorem says that, given that each student in the sample takes about $n_i \approx 40$ (Table 2.1) classes, then the standard deviation of average class difficulty ($\bar{d}_i$) should be $1.79/\sqrt{40} = 0.28$. Thus 0.28 can be used as a yardstick for measuring the observed variation in $\bar{d}_i$. Values higher or lower than 0.28 indicate that something is raising or lowering the variation in average class difficulty relative to random assignment to classes. The observed standard deviation of 0.66 is over twice the random assignment value of 0.28–there is substantial variation in average class difficulty.

Putting these two together, only the final condition for bias in grades does not hold–ability and average class difficulty are not strongly correlated with each other. In other words, essentially all students are enrolled in an equivalent mix of hard and easy classes.

[8] This result follows from the mild assumption of finite support for d_j.

2.4.4 Robustness to other regressors

These results so far are based on the decomposition of grades into a student and a class fixed effect (eq. 2.3). But other things also change within students' careers at a university. I conduct a robustness check of these results to additional regressors of the form

$$y^*_{ijt} = a_i - d_j + x_{it}\beta + \epsilon_{ijt} \tag{2.5}$$

where x_{it} are student and semester specific regressors and β are the associated regression coefficients.

Previous studies have observed grade inflation, so year fixed effects are added (Sabot & Wakeman-Linn, 1991). This model also includes semester of the year (Spring, Summer I, Summer II, Fall, and Winter) fixed effects to account for possible time of year variation in grading. These are statistically significant (Table 2.7, column B) and trend in the right direction (Table 2.9, column B). Fall and Spring are 15 week semesters and the other three terms are shorter terms where students often focus on a single class. Grades are substantially higher in the shorter terms; this could be because of changes in faculty posture towards grades or students' concentration or interest level when taking one class at a time. Despite the improvement in the fit when these regressors were added, the main results do not change when adding these regressors (Table 2.7).

Another possible issue is that student fixed effects (a_i) might not be so fixed and might change through time.[9] Adding fixed effects for class year is statistically significant (Table 2.7, column C) and does not change the correlation coefficients appreciably. However, despite the theoretical reasons to believe that there should be a positive trend in class year, there is a clear negative trend (Table 2.9).

One might suspect that students taking more or fewer classes could do better or worse, so registered (graded) credits and total (graded + ungraded) credits regressors are added (Table 2.7, column D). The addition is statistically significant. From this table it is apparent that the correlations of interest are unaffected by this addition. The coefficients are very small; adding a typical graded class with three credits (registered and total) decreases the estimated value of the (latent) prediction by 0.0097, about a third of a percent of the distance between the cutoff between a 'C+' and 'B-' and the cutoff between 'B+' and 'A-'. This could be because students are at the optimum number of classes for their allocation of time to studying college material, or because additional classes do not tend to have negative overflows, perhaps because time spent studying in college is very low so that students' effort is not constrained by time (Babcock & Marks, 2010).

The ability expansion path is very robust to inclusion of these controls with almost no change as they are added (Figure 2.7).

Finally, the intercepts for the grade boundaries are very stable across specification (Table 2.5).

[9] For example, students might be accumulating human capital.

2.4.5 Robustness to estimator

The OME describes many possible estimators. *Ordered logit* is the name for the OME when the error term is drawn from the extreme value distribution, while *ordered probit* is the name for the OME when the error term is normally distributed. Which one to use is sometimes resolved *a priori* based on the data generating process. Another way to decide which estimator to use is to treat it as an empirical question by adding a parameter that adjusts between zero and one when the error term is extreme value distributed or normally distributed, respectively (McCullagh & Nelder, 1989). A LR test on this parameter tests whether logit is better than probit. For this sample, the logit is a substantially better fit than probit, with a chi-squared of 240 on 1 degree of freedom. The associated p-value is essentially zero. Because of this, I use the ordered logit exclusively when estimating the OME.[10]

An alternative to the OME estimator is the OLS estimator, which assumes that the residual in the fixed effects estimate are normally distributed. When using the OLS estimator, the main specification changes from (eq. 2.3) to,

$$GP_{ij} \quad = \quad a_i - d_j + \epsilon_{i,j},$$

where GP_{ij} is the grade points awarded to student i in class j.

The stylized facts of the results do not change with the OLS estimator. In the OLS fit, the standard deviation of the class fixed effects is 0.48 and the stan-

[10] Other alternatives to the ordered probit and ordered logit that were tested include the asymmetric error terms from log-log and complementary log-log link functions described in McCullagh & Nelder (1989). Both were extremely poor fits for these data.

dard deviation of student fixed effects is 0.62. Similar to the OME estimate, these two standard deviations are of about the same magnitude. Qualitatively, the regression coefficients on the models with additional controls are similar (not shown). The correlations between student fixed effect, class fixed effect, and grade are also qualitatively similar (Table 2.10, first column).

The ability expansion path for the OLS results have a different interpretation than in the OME results, because the x-axis is a latent variable in the OME and it is not when using OLS. However, the resulting curves are remarkably similar nonetheless (Figure 2.8). The principal difference is at the top of the grade distribution where the simulation in chapter 4 shows that OLS performs poorly.

2.4.6 Robustness to sample selection criteria

Looking across samples, the results are qualitatively similar. The correlations coefficients for these three samples are shown in Table 2.10. These results suggest that the sample selection criteria does not change the conclusions.

The ability expansion path for these groups is qualitatively similar at every level except at the bottom (Figure 2.9). This suggests that, for students in the "Degree" sample, the lower tail's leftward shift is the result of removal of lower-ability students who were enrolled in more difficult classes at entrance.

2.5 Conclusion

GPAs and student ability are strongly correlated, and the GPA is not strongly biased. A GPA is therefore a reasonable proxy for student ability. An interesting result of this chapter is that although a student's GPA strongly reflects their ability, the same cannot be said of any particular grade a single student receives in an individual class. In fact, student ability and class difficulty contribute equally to any individual grade. However, because students of varying ability are enrolled in very similar mixes of easy and difficult classes, the influence of class difficulty on individual grades disappears in the overall GPA.

A necessary condition for GPAs to be biased is that individual classes must vary in difficulty. This condition holds–the variation in individual class difficulty (class fixed effects) is so large that ability (student fixed effects) and class difficulty are approximately equally important in determining any individual grade. Individual grades, even conditional on the class, are dominated by noise. Unconditionally, individual grades are not comparable; ability is not strongly predictive of the grade. Thus there is a strong distinction between the influence a student has on their overall GPA and their grade in each individual class.

Bias in GPA also requires that within-student-average class difficulty (average class difficulty) varies. This condition also holds–variation in average class difficulty by student is substantial.

Finally, bias in GPA requires that the variation in average class difficulty is associated with student ability. This condition does not hold–average class difficulty and student ability are only slightly correlated. Therefore, GPAs are largely not biased at the University of Maryland. Student ability predicts 80% of the variation in GPAs and the empirical ability expansion path is nearly vertical.

The shape of the ability expansion path also informs the mechanism by which human capital or ability is increasing human capital production–students with higher ability are taking classes of the same difficulty level as the low ability students.

There are two exceptions to the vertical expansion paths–students at the top and at the bottom of the ability distribution, where ability and difficulty are positively associated.

Students at the top of the ability distribution appear to enroll in harder classes and get mostly 'A's in these classes. Because of this, the difference between a student who earns a 3.9 GPA versus a 3.91 is much larger than the difference between a student who earns a GPA of 3.0 versus 3.01.

At the bottom of the ability distribution, students who are continuously enrolled take easier classes. This appears to be a mechanical effect of a minimum GPA for graduation.

When using grades as a measure of ability, especially for students who are in the middle of the GPA distribution, GPAs are a good predictor of ability. Average class difficulty (the variation in class difficulty by student) explains little (about 3%) of the variation in GPAs. There is also 17% residual variation, so one cannot rule out that some other confounding factor could affect grades. When studying students that are near the 2.0 GPA cutoff, the difficulty of the classes that they take is important, and an indicator for dropping out could be a better measure of outcomes than GPA, which is kept near the minimum grade by university policy.

Because these results are from only one university, they might be observed as “local” to one portion of the ability distribution–that observed at a top state university. But two facts suggest that the results are relatively broad: consistent with universities studied in other research, there is substantial variation in average grade by department, and the range of SAT scores observed is quite wide. Nevertheless, reproducing (at other universities) the main results of this paper–that GPAs are a good measure of student ability–would be valuable.

Table 2.1: Descriptive statistics of student transcript data

	Rect	Enter	Full
GPA	2.94(0.599)	2.80(0.747)	2.82(0.717)
SAT verbal	599(85.5)	592(88.7)	594(89.2)
SAT math	632(85.7)	625(87.7)	623(87.4)
took SAT	0.683	0.635	0.475
HS GPA	3.80(0.493)	3.75(0.508)	3.75(0.491)
has HS GPA	0.881	0.790	0.785
trans. GPA	3.13(0.464)	3.14(0.461)	3.15(0.458)
has trans. GPA	0.202	0.325	0.300
female	0.485	0.494	0.483
white	0.569	0.565	0.575
age at entry	18.4(1.31)	18.8(1.64)	19.1(1.72)
n terms	9.94(2.11)	8.25(3.14)	6.23(3.19)
n classes	41.6(7.91)	34.7(12.7)	26.2(12.9)
total credits	121(20.1)	101(35.6)	76.4(37.0)
degree seeking	1.000	1.000	0.992
n	12,995	19,489	63,875

Note: Standard deviations are in parentheses for all non-binomial variables.

Table 2.2: Entry/exit for “degree” data

	2003	2004	2005
n	3,909	4,528	4,558
2004	1	—	—
2005	118	9	—
2006	476	218	14
2007	2,738	740	212
2008	410	2,916	800
2009	116	497	3,007
2010	50	148	525

Note: Each entry shows the number of students in “degree” data (see text) who enter in 2003-2005 (columns) and who exit in 2004-2010 (rows). The total number of entrants for each year is listed in the first row.

Table 2.3: Entry/exit for “enter” data

	2003	2004	2005
n	5,709	6,819	6,961
2003	60	—	—
2004	512	183	—
2005	741	626	165
2006	913	1,071	650
2007	2,881	1,174	1,095
2008	417	3,080	1,282
2009	123	520	3,215
2010	62	165	554

Note: Each entry shows the number of students in “enter” data (see text) who enter in 2003-2005 (columns) and who exit in 2004-2010 (rows). The total number of entrants for each year is listed in the first row.

Table 2.4: Main results

	GPA_i	$\bar{d}_i$	a_i
GPA_i	—	-0.19	0.90
$\bar{d}_i$	-0.19	—	0.16
a_i	0.90	0.16	—

Note: Pearson correlation coefficients (ρ) between column and row quantities. The quantity $\bar{d}_i$ is the average class difficulty level (d_j) aggregated to the student level.

Table 2.5: Grade boundary cut points from regressions

	Model				Grade width			
	A	B	C	D				
F/D	0.00	0.00	0.00	0.00	0.59	D	0.59	D
D/C-	0.59	0.59	0.59	0.59	0.36	C-	1.65	C
C-/C	0.94	0.95	0.95	0.95	0.99	C		
C/C+	1.94	1.95	1.95	1.94	0.33	C+		
C+/B-	2.25	2.27	2.27	2.27	0.54	B-	2.47	B
B-/B	2.79	2.81	2.81	2.81	1.35	B		
B/B+	4.13	4.16	4.16	4.16	0.60	B+		
B+/A-	4.73	4.76	4.76	4.76	0.97	A-		
A-/A	5.69	5.73	5.73	5.73				

Note: Grade widths (last column) are based on model D. Because the 'F' and 'A' ranges are unbounded, the total width of the bins is always infinite.

Table 2.6: Likelihood ratio tests for addition of class and student fixed effects

Small	Large	k	LR test statistic	Jackknife LR
intercept	class	4,809	193,578	342,195
intercept	student	12,994	143,692	262,861
intercept	class + student	17,803	369,029	661,791
class	class + student	12,994	225,336	398,929
student	class + student	4,809	175,450	319,596

Note: All p-values are indistinguishable from zero. See chapter 4 for a description of the jackknife method applied.

Table 2.7: Standard deviations, correlation coefficients, and significance tests for various specifications, estimated with the OME

	Model			
	A	B	C	D
student + class	X	X	X	X
year + semster	–	X	X	X
class year	–	–	X	X
semester credits registered and total	–	–	–	X
σ_a	1.60	1.63	1.65	1.66
σ_d	1.70	1.68	1.61	1.62
$\sigma_{\bar{d}_i}$	0.63	0.63	0.62	0.62
τ-$G_{d,g}$	0.33	0.33	0.35	0.33
τ-$G_{a,g}$	0.41	0.44	0.39	0.42
τ-$G_{a,d}$	-0.07	-0.06	-0.05	-0.05
ρ_{a_i,GPA_i}	0.902	0.905	0.906	0.907
$\rho_{\bar{d}_i,GPA_i}$	0.161	0.165	0.164	0.164
$\rho_{a_i,\bar{d}_i}$	-0.185	-0.176	-0.176	-0.175
additional controls	–	11	3	2
LR test stat.	–	$5,461$	329	75
χ^2 p-value	–	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
observations	$523,151$	$523,151$	$523,151$	$523,151$
students	$12,995$	$12,995$	$12,995$	$12,995$

Note: The top block shows the regressors included in the columns. The second block shows correlation coefficients for classes (measured with the Goodman-Kruskal correlation coefficient τ-G) and transcripts (measured with ρ). Subscripts indicate ability (a), class difficulty (d), average class difficulty ($\bar{d}_i$), and grade g. The final block shows likelihood ratio tests for inclusion of the additional regressors.

Table 2.8: Goodman-Kruskal correlation coefficients

	g_{ij}	a_i	d_j
g_{ij}	—	0.43	0.34
a_i	0.43	—	-0.06
d_j	0.34	-0.06	—

Note: Each Goodman-Kruskal correlation coefficient (τ-G) is between column and row quantities, grades g_{ij}, student fixed effect (a_i); and class difficulty (d_j).

Table 2.9: Ancillary regression coefficients

	Model			
	A	B	C	D
year=2003	–	−0.15	−0.63	−0.60
2004	–	−0.23	−0.65	−0.62
2005	–	−0.20	−0.56	−0.54
2006	–	−0.15	−0.42	−0.41
2007	–	−0.16	−0.32	−0.31
2008	–	−0.17	−0.24	−0.24
2009	–	−0.05	−0.08	−0.08
2010	–	–	–	–
term=Spring	–	−0.04	−0.08	−0.07
Summer I	–	0.86	0.84	0.84
Summer II	–	0.84	0.84	0.83
Fall	–	–	–	–
Winter	–	1.01	1.05	1.04
freshman	–	–	–	–
sophomore	–	–	−0.08	−0.06
junior	–	–	−0.20	−0.19
senior	–	–	−0.36	−0.34
registered credits	–	–	–	5.9×10^{-4}
total credits	–	–	–	-38.2×10^{-4}

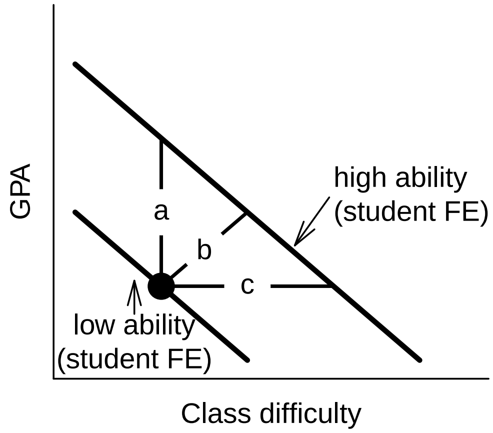
Note: A dash indicates a regressor is the omitted level, or that the whole set was was not included in the specification.

Table 2.10: Robustness of results to sample selection

		"Degree"	"Enter"	"Full"
OME	σ_a	1.76	2.41	2.95
	σ_d	1.80	1.67	1.61
	$\sigma_{\bar{d}_i}$	0.66	0.66	0.71
	τ-$G_{a,g}$	0.43	0.45	0.47
	τ-$G_{d,g}$	0.33	0.32	0.30
	τ-$G_{a,d}$	-0.05	-0.04	-0.05
	ρ_{a_i,GPA_i}	0.86	0.87	0.79
	$\rho_{\bar{d}_i,GPA_i}$	0.19	0.23	0.20
	$\rho_{a_i,\bar{d}_i}$	-0.10	-0.02	-0.04
OLS	σ_a	1.61	0.81	0.79
	σ_d	0.46	0.46	0.45
	$\sigma_{\bar{d}_i}$	0.21	0.22	0.24
	τ-$G_{a,g}$	0.43	0.45	0.45
	τ-$G_{d,g}$	0.31	0.31	0.30
	τ-$G_{a,d}$	-0.06	-0.06	-0.04
	ρ_{a_i,GPA_i}	0.94	0.96	0.95
	$\rho_{\bar{d}_i,GPA_i}$	0.09	0.18	0.16
	$\rho_{a_i,\bar{d}_i}$	-0.24	-0.08	-0.11
	observations	$537,606$	$677,634$	$1,699,053$

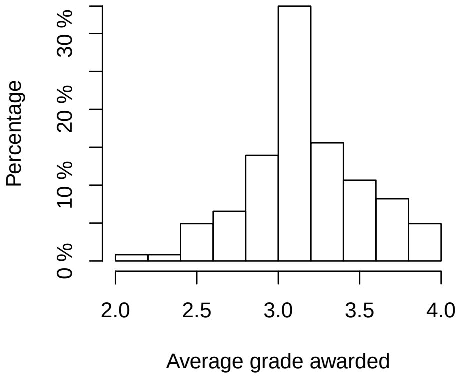
Note: This table shows the standard deviation of student ability (a), class difficulty (d), and average class difficulty ($\bar{d}_i$); Goodman-Kruskal (τ-G) type correlations at the transcript entry level between student ability (a), class difficulty (d), and grade (g); and Pearson type correlations between GPA, ability, and the average of all class difficulties ($\bar{d}_i$) for OME and OLS estimators.

Figure 2.1: Ability expansion path



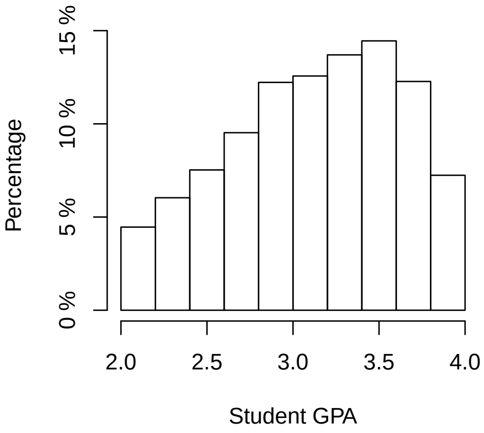
Note: see text.

Figure 2.2: Histogram of average grade awarded by departments



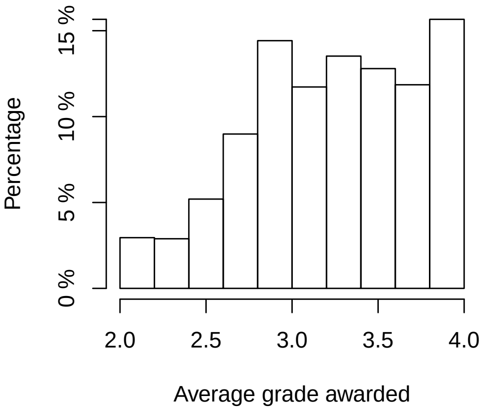
Note: Each department is counted once; the histogram is not weighted by enrollment. Only departments with average graded awarded of 2.0 and higher are shown. This figure uses the “Degree” sample.

Figure 2.3: Histogram of student GPAs



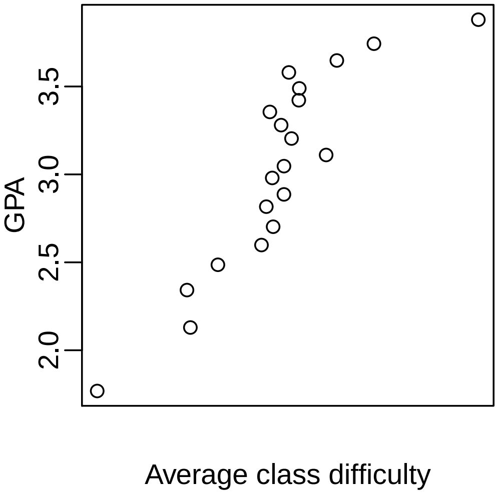
Note: The data used for this histogram is the "degree" sample and only students with GPA 2.0 and higher are shown.

Figure 2.4: Histogram of average grade awarded by class



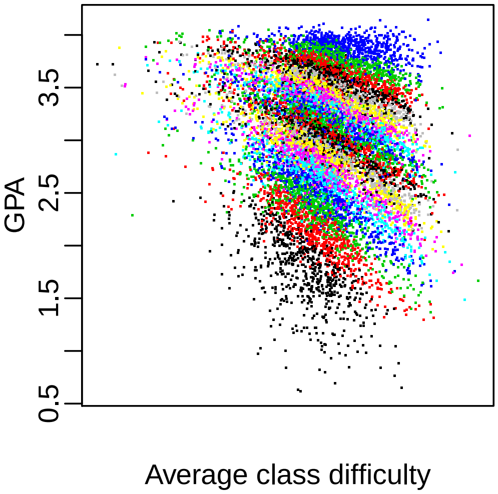
Note: The data used for this histogram is the "degree" sample and only classes with average grade awarded 2.0 and higher are shown.

Figure 2.5: Observed ability expansion path



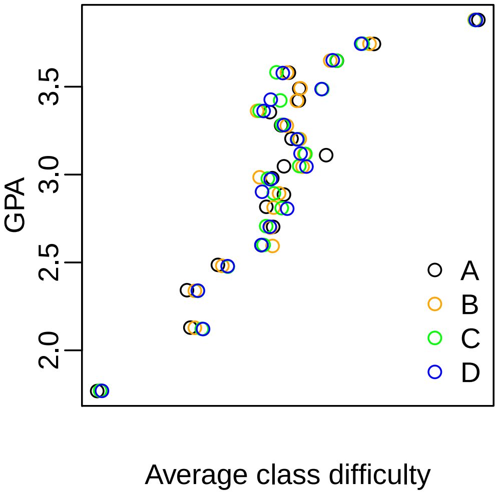
Note: Each point is a single vigentile (twentieth), showing the average class difficulty (as a deviation from the grand mean) on the x-axis and the average GPA on the y-axis.

Figure 2.6: Raw GPA versus average class difficulty



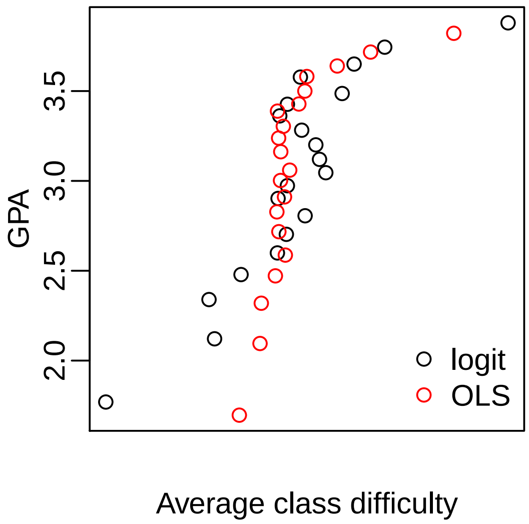
Note: Individual students' GPA vs average class difficulty, with color used to indicate student ability (student fixed effect) by vigentile (twentieth).

Figure 2.7: Ability expansion path with various additional regressors



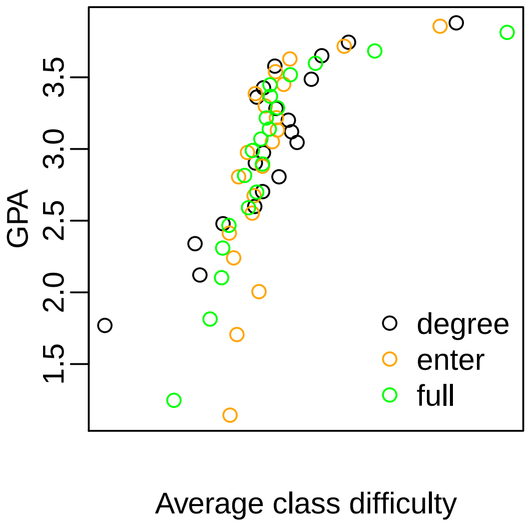
Note: Each point is a single vigentile (twentieth), showing the average class difficulty (as a deviation from the grand mean) on the x-axis and the average GPA on the y-axis. Model A is shown in black, model B in red, model C in green, and model D in blue.

Figure 2.8: Ability expansion path from OME and OLS estimators



Note: Using two estimators: OLS (red) and OME (black, also called logit). Each point is a single vigentile (twentieth), showing the average class difficulty (as a deviation from the grand mean) on the x-axis and the average GPA on the y-axis.

Figure 2.9: Ability expansion path from the three datasets



Note: Using three datasets: “degree” (black), “enter” (orange), and “full” (green). Each point is a single vigentile (twentieth), showing the average class difficulty (as a deviation from the grand mean) on the x-axis and the average GPA on the y-axis.

Chapter 3

Are Low Income Students Diamonds in the Rough?

3.1 Introduction

Consider two students who receive the same SAT score, one from a lower income family and another from a higher income family. If investment in schooling is a normal good, then the student from the higher income family is the product of a higher investment education than the student from the lower income family.[1] Because of this one might suspect the individual from the lower income family has a higher "innate ability" while the individual from a higher income family has been the benefactor of a higher investment/higher output educational environment.[2] When these two students enroll at the same university and the difference in their family investment inputs into education are reduced or eliminated, who will perform better?

This question is important for college admission, where a debate continues[3] as to whether low income students with lower observable ability at the time of application should nonetheless be granted entrance, because their past has included fewer opportunities and "held them back." In this view, students from low income

[1] Note that investment could be pecuniary, or time. The mechanism is not important or investigated in this chapter.

[2] Here family income is measured using the median income by zip code.

[3] See, for example popular press pieces such as Leonhardt (2011) and the response of Mankiw (2011). Others have focused on simply measuring collegiate performance as a function of prior inputs, with many authors focusing on the predictive power of standardized tests as an ends (Betts & Morell, 1998; Cohn et al., 2004; Grove & Wasserman, 2003; Bettinger et al., 2011; Geiser & Studley, 2001; Rothstein, 2004).

families are *diamonds in the rough*.

Because I am using administrative data from one school (the University of Maryland) the students in the sample are not a random sample of college students. The college admissions process has multiple steps that make each university's students a mutually selected group and thus using data from a single school turns out to be problematic for this research design. A model that focuses on family income and SAT score (as a measure of student ability) at one university is not sufficient to explore the "diamonds in the rough" hypothesis.

In addition, at the University of Maryland, high income students perform slightly worse than their lower income counter parts unconditionally. When the same regression is conditioned on SAT score, low income students continue to outperform their higher income counterparts, but by a lower margin.

According to the traditional model for human capital development, at any given time an individual has a particular level of human capital that determines how quickly he or she can accumulate new human capital. In the canonical example, Ben-Porath describes human capital accumulation as a product of innate ability and current human capital (Ben-Porath, 1967), but ability and accumulated human capital are essentially identical since they always appear multiplied by one another. In a stylized form, production functions based on the Ben-Porath model could be written

$$\frac{dHC_t}{dt} = f(HC_t, I_t), \tag{3.1}$$

where HC_t is accumulated human capital for an individual at time t, I_t is inputs applied to accumulating additional human capital, $\frac{dHC_t}{dt}$ is the accumulation rate of human capital, and $f(\cdot, \cdot)$ is a function that has positive partial derivatives everywhere in both arguments. Here the concept of ability manifests as initial levels of human capital and is otherwise not present. In this type of model, a measure of aptitude/ability is a sufficient statistic to explain future performance. Two students with similar test scores are expected to be equally productive regardless of whether one was initially (at birth) high ability but saw low productivity increases before the test and the other was initially low ability but saw large productivity increases before the test.

Education research uses this type of model for value added modeling, with human capital accumulation functions of the form (Hanushek, 2006)

$$E(\widehat{HC}_t - \widehat{HC}_{t-1}) = f(I_t), \tag{3.2}$$

where $E(\cdot)$ is the expected value operator, $\widehat{HC_{t-1}}$ is a pre-test score, taken before the time period of interest, $\widehat{HC}_t$ is a post-test score taken after the time period of interest, and I_t is inputs of interest during the time period of interest.[4] The underlying assumption is that by subtracting a measure of human capital from the previous time period, all inputs to human capital that occurred before the time of interest are captured in the measure of human capital measure taken in the pre-time-period, $t - 1$.

[4] This is a stylized version of the model presented by Hanusheck.

An alternative to the assumptions of these models is that ability is always important to production. Regardless of the current level of human capital accumulated, there is an innate ability to learn that varies among individuals. In this view, a test of aptitude will not capture ability to learn and so is inadequate for predicting output once at college. Human capital accumulates according to

$$\frac{dHC_t}{dt} = f(HC_t, I_t, \alpha_0), \tag{3.3}$$

where α_0 is innate ability.[5] [6] In this model, greater innate ability can compensate for lower prior investment. Students with lower pre-collegiate educational investments who are able to acquire the same degree of human capital as students with higher pre-collegiate educational investments achieved this by virtue of greater innate ability. Once enrolled at the same university, these lower-investment (i.e., lower income) students will outperform their higher-investment (i.e., higher income) peers, conditional on having the same level of human capital upon entrance.

Finally, one might suspect that non-cognitive skills play a role in human capital formation in a way that is not completely captured in test scores. It would also make sense that higher income families imbue their children not only with higher cognitive skills but also higher non-cognitive skills so that family income will be positively correlated with college output, even when conditioning on test scores.

Because the sampled population is college students at the University of Mary-

[5] A subscript zero is used on alpha to distinguish it from the later use of a constant in regressions.

[6] The Ben-Porath model fits into this specification, but is multiplicatively not identified for the first and third arguments so that $f(\lambda \cdot HC_t, I_t, \alpha_0) = f(HC_t, I_t, \lambda \cdot \alpha_0)$.

land, students must have been admitted to the University and then have chosen the University of Maryland for college.[7] To understand the role of the selection process, it is important to know that the main specification is a regression of the form

$$[\texttt{student output}] = \alpha + \beta_1 \cdot [\texttt{SAT}] + \beta_2 \cdot [\texttt{family income}] + \epsilon \ . \tag{3.4}$$

I make the assumptions that SAT is an ability measure and that the the SAT score is the only measure of pre-collegiate human capital observable to the admissions office. Making these strong assumptions, β_2 is interpretable as an effect from ability itself. The source of these assumptions is a null hypothesis that the model described by Hanusheck's (eq. 3.2) is correct, with the SAT treated as a pre-test (Hanushek, 2006).

Surprisingly, running a bivariate regression of GPA on family income yields a larger negative coefficient than one gets in equation 3.4 suggesting that low income students at the University of Maryland are unconditionally stronger applicants—they are simply diamonds. Adding in the SAT, per equation 3.4, mitigates the negative coefficient on family income. This is not what the diamonds in the rough hypothesis would predict. The negative coefficient on family income, which would have been surprising if the unconditional regression coefficient on β_2 were positive, is not surprising when low income students are unconditionally outperforming their high income peers.

[7] Dale & Krueger (2002) identify three steps of selection where students select schools to which they apply, schools admit students, and then students select schools from those they were admitted. Because I am not modeling the selection process, its exact form is less important here.

In the the main specification (eq. 3.4) the outcome variable (student output) is measured in two different ways: GPA and *student ability*. Student ability should not be confused with *innate ability* (α_0); it is measured as the student fixed effect in a decomposition of grades into student and class effects and best interpreted as exactly that. In chapter 2, I found that GPA and student ability are highly correlated, with a correlation coefficient of 0.90, suggesting that both measures should tell a similar story in this chapter. The main results of this chapter are approximately the same across the two specifications, consistent with my previous chapter's findings. However, I find that there is a difference between GPA and student ability. in one specification check, I find that a one point higher SAT math score is associated with higher student ability than a one point higher SAT verbal score. At the same time, a one point higher SAT math score is associated with the same GPA as a one point higher SAT verbal. Because SAT math and SAT verbal are approximately Z-scores, this implies that for students with higher SAT math scores, GPA underestimates actual ability.

The next section describes the data used in this chapter, the sample selection criteria applied and its impact on the covariates. The third section presents the results and the final section concludes.

3.2 Data

The data for this chapter, like those described in section 2.3 are drawn from transcripts of University of Maryland college students between 2003 and 2010. The main sample is a subset of the “enter” sample, which focuses on matriculated students who entered between 2003 and 2005 and thus had five years in which to complete their degree–though it is not a requirement that the students in the sample did, in fact, complete a degree. Later specification checks use the “degree” sample, which adds a requirement that the student appears to have graduated, and the “full” sample which removes the admission year requirement of 2003 through 2005. In contrast with the previous chapter, an observation in this chapter is a student. Student performance is measured by GPA and by student ability (a_i) as measured from the decomposition in the chapter 2 of the form

$$G_{ij} = a_i - d_j + X\beta + \epsilon_{ij} \tag{3.5}$$

where G_{ij} is the grade student i received in class j, a_i is a fixed effect for the ith student and d_j is a fixed effect for the jth class, and there are some other variables such as year fixed effects in X. The exact specification used is specification D from chapter 2 (Table 2.7), which includes controls for semester, class year, and number of classes the student is signed up for in the semester.

Two methods are used to fit this decomposition in chapter 2, ordered multinomial logit and OLS; in this chapter I use the OLS results because they are readily interpreted as changes to GPA while the ordered logit results are in terms of a latent parameter. For example, when using the OLS results, a student with an ability that is 1.0 higher than another student would be expected to get one letter grade higher when he or she takes the same class. In contrast, when using the ordered logit results, a similar statement cannot be constructed for a student with an ability 1.0 higher than another student.

Another statistic used to describe student's college career is the *average class difficulty* ($\bar{d_i}$) defined as

$$\bar{d_i} = \frac{1}{n_i} \sum_{\text{student } i} d_k \tag{3.6}$$

where the sum is over the classes student i enrolled in, of which there are n_i. Each student has his or her own average class difficulty, and it is the average of the difficulty of the classes the student took. In eq. 3.5, the difficulty enters with a negative in front of it so that more difficult classes are associated with lower grades.

The median income of the zip code listed as a permanent address is used as a proxy for the student's family income (and this variable is called called *family income* throughout this chapter). This measure can be thought of as typical family income in the community the student is from. In addition, while school boundaries and zip code boundaries are not identical, they are both based on geographic proximity and so membership in a community may not represent identical incomes, but does represent access to a similar level of public primary and secondary school.[8]

The main specification includes GPA, SAT score, student ability, and zip code with a published median income on the 2000 Census standard file 3 (U.S. Census Bureau, 2001). Only students for which all these variables are present are included

[8] See, for example, Oates (2005) for an excellent review of the concept of the market for local government goods. One example is Hamilton, who argues that zoning and property taxes can homogenize communities with respect to public service demand (Hamilton, 1975).

in the sample. From a baseline of all students who qualify for the "enter" sample (n_0=19,500), removing students with zip codes that are not tabulated (presumably to maintain the privacy of those living at the address) removes about 1,000 individuals; removing students without a valid SAT score removes 7,300 students (n=10,621). The other requirements do not remove any students because they are derived from the transcript data itself.

Students selected into the sample are similar to the entire "enter" sample. A comparison of the raw and selected groups is shown in Table 3.1 (columns 1 and 2) and shows that the two groups are approximately balanced on race and gender. The academic achievement variables are slightly higher for the selected sample–for example, GPA is 0.03 higher and student ability is 0.05 higher–reflecting a small increase in typical student quality. The exception to the increase in academic achievement is that the students in the sample have a slightly lower average high school GPA. Without controlling for the high school the student attended, it is not possible to know if this represents increased difficulty at their particular high schools or a lower level of achievement. In any case, the difference is small.

Not all of the students on the sample have taken the SAT. Some of the students who did not take the SAT took the ACT instead. It is possible to use a conversion factor to estimate an SAT score from an ACT score. Wainer (1986) observes that while such conversion is possible, it is inaccurate for individuals, even if accurate when averaged over large groups. Wainer published his conversion factors but since the publication date Educational Testing Service (the owner of the SAT) decided to "recenter" the SAT periodically, essentially updating the scores to be normally distributed with a relatively invariant mean and standard deviation by applying periodic adjustments to the raw score on the SAT scale (Dorans, 2002).[9] This recentering renders conversion factors based on pre-recentering data (such as Wainer's) inappropriate. Using a subsample of the students who took both the SAT and the ACT and ignoring possible selection bias associated with using this self-selected group, it is possible to develop a conversion factor based on a linear model (Table 3.2). These conversion factors are not ideal. The R^2 is about 0.7 for the math and verbal imputation scheme so only about 70% of the variation in SAT scores explained by the ACT scores.

The observable academic measures (including their SAT imputed from ACT scores) are different from the selected group (Table 3.1, column 3). The students who took only the ACT have lower (imputed) SAT scores, higher family income, and higher ability than the students who took the SAT. In light of this, an additional regression is run with these students included as a specification check.

Subsamples of the "degree" and "full" samples are also used as a specification check. The subsamples are created using the same sample selection as the "enter" group described above. The "degree" sample is created in a similar way to the "enter" sample except that an attempt is made to winnow it to students who probably graduated by limiting the sample to students who were continuously enrolled for the entire sample or who completed over 120 credits.[10] Even more than the "enter" sample, the "degree" sample and the selected "degree" sample are very similar to

[9] The scores are centered so the mean is about 500 and the standard deviation is about 110 for each test.

[10] A full description appears in chapter 2.

each other (Table 3.3).

The "full" sample is similar to the "enter" sample but removes the requirement that the student started taking classes by 2005. For the "full" sample there is a non-trivial increase in student ability (a_i) in the selected sample, with an increase of 0.10, this is much larger than in the other samples (Table 3.4). While larger than the other changes, it is still not very large, and has no obvious effect on the regression results shown.

3.3 Results

The main regression is

$$Y = \alpha + \beta_1 \cdot [\texttt{SAT}] + \beta_2 \cdot [\texttt{family income}] + \epsilon \tag{3.7}$$

where Y is the outcome of interest (GPA or student ability). The regression asks whether, controlling for SAT scores, family income is associated with higher or lower GPA or student ability. The family income coefficient from this regression shows that higher incomes are associated with slightly lower GPAs and student ability (Table 3.5). The estimated change in GPAs from a \$10,000 higher family income is -0.013. The estimated change in student ability from a \$10,000 higher family income is -0.012.

As an example, consider two students, one with average family income, the other with a two standard deviation higher income family (\$27,000 higher) and both students have combined SAT scores of 1210. The lower income student will be expected to earn 0.035 higher GPA, or have an estimated 0.032 higher ability. A third student from the same lower income family but with 24 points lower SAT score (a 1186) would be expected to match the GPA of the higher income student.[11]

Running this regression by fitting just the family income regressor (dropping SAT) reveals a simpler explanation for the negative coefficient on β_2: lower income students are also unconditionally higher performing (Table 3.6). The regression coefficient of family income unconditionally is -0.021 when GPA is the outcome and -0.022 when student fixed effect is the outcome. The R-squared on this regression is almost exactly zero, suggesting that family income is not a strong determinant of grades.

When similarly running the regression with just SAT the regression coefficients are almost unchanged relative to the results when including family income.

An apparent reason for this negative coefficient on β_2 when SAT is dropped might be that while nationally SAT and family income are positively correlated (College Board, 2009), at the University of Maryland, the family income and SAT scores are almost uncorrelated. In fact, there is a small negatively correlation coefficient (-0.04). This means that lower income students have higher SAT scores. This raises the possibility that they also have higher unobservable quality and this is what drives the negative coefficient on family income when conditioning on SAT.

In specification 3.7, I lumped together SAT math and SAT verbal scores as if they test two skills that are equally useful for increasing student output. However,

[11] Differences in SAT scores can only be denominated in units of 10, so a technically correct statement would be that a group of students from the same lower income family with average SAT scores of 1186 would be expected to perform the same as this hypothetical third student.

they might be different and this is tested by separating out the SAT scores in the regression

$$Y = \alpha + \beta_{1v} \cdot [\texttt{SAT verbal}] + \beta_{1m}[\texttt{SAT math}] + \beta_2 \cdot [\texttt{family income}] + \epsilon \qquad (3.8)$$

Adding in SAT math and verbal scores separately does not change the results in a statistically significant way for the GPA estimates, meaning that the simper model with the two tests lumped together cannot be rejected as less explanatory. The best estimates from regression equation 3.8 suggests that an increase of 100 points in SAT verbal scores increases GPAs by 0.155 while a 100 point increase in SAT math scores increases GPA by 0.139 (Table 3.7). The insignificant t-test on the contrast between the two regressors indicates that SAT math and SAT verbal do not have different effects on GPA.

In the same specification, using student ability as the outcome variable, the addition of SAT math is statistically significant and has a much stronger association with student ability than SAT verbal scores. An increase of 100 points in SAT verbal scores is associated with an increase in ability of 0.421 while an increase of 100 points in SAT math scores is associated with an increase of 0.580 in student ability (Table 3.7).

This result is striking. Despite the previous chapter's result that GPAs are a good measure of student ability, GPAs are not affected differently by SAT math and SAT verbal scores, but student ability is.

For both possible outcome variables, the main results—changes in GPA based on family income—is not affected by breaking SAT down into math and verbal scores.

The most obvious explanation for why these two differ is that the difficulty of the classes that students take varies by family income and SAT math score. To verify that, I regressed students' average class difficulty ($\bar{d_i}$) on the same regressors using the same specifications (Table 3.8). These results show that average class difficulty matters. Students with relatively higher SAT math scores are taking harder classes than their counterparts with lower SAT math scores. An increase of 100 points in SAT math increases average class difficulty by 0.093.[12] Higher total SAT verbal scores are associated with a slight decrease the average class difficulty of 0.023. Both the coefficients on SAT math and SAT verbal are statistically significant at the 1% level, as is the contrast between the two.

An increase in family income increases difficulty, if very slightly, by 0.003, regardless of the specification and this result becomes statistically significant at the 10% level when including SAT math and SAT verbal separately. This suggests that it is possible that higher income students are taking more difficult classes, on average, even if just slightly.

To test the robustness of the main results, I performed additional checks. First, family income might not enter entirely linearly, for example, if the effect accrues largely at the top or bottom of the income distribution. I ran a regression with the family income binned by quintile, and the bins are included as dummy

[12] Note that one might have hoped that the increase in ability minus the increase in difficulty might exactly equal the decreased GPA but it does not. However, a null hypothesis that they are equal cannot be rejected.

variables in lieu of the linear estimator (Table 3.9). When this is done, the effects are approximately linear for GPA and noisy but still approximately linear for ability. In the case of GPA the estimated change in GPA by income quintile decrease monotonically with an average decrease of 0.014 from bin to bin. When using student ability as an outcome, the estimated effects decrease monotonically, with only one exception at the lowest income group. Overall, the non-linearities appear to be relatively subtle, making an assumption of linearity reasonable.

In the previous chapter, I used ordered logit as well as OLS to estimate the decomposition in equation 3.5 but there are theoretical reasons to prefer the ordered multinomial logit to the OLS results even though they were very similar in final fit. Running the regression on the ability measure shows slightly different results. The estimated regression coefficient is still negative at -0.019 but is not statistically significant, with an associate t-statistic of 1.6. Note that the different absolute value of the estimated coefficient between the OLS-based student ability and ordered multinomial logit-based student ability is not itself interpretable because the two ability measures are not measured on the same scale.

Students who did not take the SAT but did take the ACT are not in the sample, despite the fact that it is possible to impute their SAT scores using the ACT. Imputed SAT scores are more imprecise and that can bias a regression coefficient towards zero. However, there is a competing concern that these students were not matched exactly on their baseline characteristics. Running the regression including these students not only does not move the estimate closer to zero; instead it decreases it slightly to -0.013, and it becomes significant at the 1% level (Table 3.10). This suggests that removing these students did not bias the sample towards negative and more significant results.[13]

The "enter" sample includes most students who started college at the University of Maryland in 2003 through 2005 regardless whether they ultimately graduated. Running the regressions using just those who graduate gives very similar results (the "degree" sample), except that when GPA is the outcome, the estimate for family income increases slightly in absolute terms to -0.015 and remains significant (at the 1% significance level) when predicting the student fixed effect (Table 3.14). Using the "full" sample, which includes students who entered after 2005 also shows a nearly identical result as the "enter" sample, with an estimated coefficient on family income of -0.012, a result that is again statistically significant at the 1% level.

The loss of significance using the ordered logit-based estimate of ability may have been a result of marginal significance of the original regression coefficient. Rerunning the regressions on the "degree" and "full" samples using the logit shows that both have a statistically significant estimate of family income (Tables 3.14 and 3.15, respectively).

3.4 Discussion and Conclusion

These results show that, at the University of Maryland, conditional on SAT scores, lower income is associated with higher college output. However the same is also true unconditional on SAT scores. In any case, the result is relatively modest;

[13] When a noisy measure of a regressor is used in place of a true value the regression coefficient is biased towards zero. What is worse, as other regressors are added, the bias increases (Griliches, 1977).

a $10,000 increase in household income (approximately one standard deviation) is associated with a decrease of about 0.01 in GPA–suggesting that the family income effect identified here, even if causal, would not dominate students' college performances.

Methodologically, this result holds when college output is measured directly via GPA and ability, which controls for the difficulty of the classes in which students enroll. This suggests that GPA and ability could have been used interchangeably to get the same results in this chapter. However, in one specification, a weakness of measuring collegiate output with GPA was revealed. An increase in SAT math is associated with an increase in class difficulty and an approximately balancing increase in student ability. In contrast, an increase in SAT verbal is associated with a smaller increase in student ability and a slightly negative change in average class difficulty. For this chapter, this asymmetry was irrelevant to the conclusion, but one could imagine a case where it is important. Because of this, using GPA as a proxy for ability still may make sense, but additional consideration should be given as to whether relative math ability might play a role in the particular question being asked.

At the beginning of this chapter I posited two students who received the same SAT scores, one from a lower income family and another for a higher income family. I then asked if innate ability, previously unrealized, would shine through at the University of Maryland. However, data from the University of Maryland proved inappropriate for answering this question because of the non-random nature of enrollment.

Table 3.1: Sample description

	Enter	Enter sample	Enter ACT
GPA	2.80(0.747)	2.83(0.731)	2.90(0.716)
SAT verbal	5.91(0.864)	5.91(0.873)	5.71(0.617)
SAT math	6.24(0.878)	6.18(0.892)	6.05(0.837)
family income /$10,000	4.13(1.35)	4.13(1.30)	4.19(1.43)
a_i	-1.17(0.748)	-1.12(0.733)	-1.11(0.698)
$\bar{d}_i$	-0.198(0.221)	-0.199(0.218)	-0.170(0.215)
took SAT	0.599	1.000	0.000
HS GPA	3.75(0.508)	3.72(0.512)	3.79(0.481)
has HS GPA	0.790	0.950	1.000
transfer	0.407	0.319	0.189
female	0.494	0.488	0.652
white	0.565	0.574	0.545
n terms	8.25(3.14)	8.60(3.15)	8.77(2.86)
n classes	34.7(12.7)	36.4(12.4)	39.1(11.2)
total credits	101(35.6)	107(34.6)	113(30.9)
degree seeking	1.000	1.000	1.000
n	19,489	11,183	514

Note: The first column is the entire "Enter" sample, before sample selection criteria for this chapter are applied. The second column is the main sample for this chapter. The third sample is the students who have an ACT score but no SAT score. Their imputed SAT scores are reported. Each entry in the table is an average of a variables, named in the row, followed by standard deviations in parentheses for non-binomial variables. Some of the sample selection criteria force a variable to be exactly 0 or 1.

Table 3.2: Predicting SAT with ACT scores

	SAT verbal	SAT math
	(1)	(2)
Intercept	234^{***}	203^{***}
	(5)	(3)
ACT reading	6.82^{***}	—
	(0.25)	
ACT English	7.16^{***}	—
	(0.29)	
ACT math	—	15.7^{***}
		(0.1)
R^2	0.67	0.73
obs	2,657	6,402

Note: Standard errors appear in parentheses below each regression coefficient. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. The first column shows regressions of SAT verbal scores with various ACT tests. The second two columns show regressions of SAT math scores on ACT tests.

Table 3.3: Sample description: degree dataset

	Degree	Degree sample
GPA	2.94(0.599)	2.96(0.583)
SAT verbal	5.97(0.831)	5.98(0.838)
SAT math	6.32(0.858)	6.26(0.871)
family income /$10,000	4.12(1.32)	4.12(1.27)
a_i	-1.03(0.635)	-0.988(0.620)
$\bar{d}_i$	-0.221(0.213)	-0.219(0.212)
took SAT	0.643	1.000
HS GPA	3.80(0.493)	3.78(0.495)
has HS GPA	0.881	0.972
transfer	0.289	0.239
female	0.485	0.482
white	0.569	0.571
n terms	9.94(2.11)	10.1(2.19)
n classes	41.6(7.91)	42.3(7.48)
total credits	121(20.1)	123(19.0)
degree seeking	1.000	1.000
n	12,995	8,017

Note: Means and standard deviations are in parenthesis for non-binary variables. This table shows the degree sample before and after sample selection criteria are applied.

Table 3.4: Sample description: full dataset

	Full	Full sample
GPA	2.82(0.717)	2.85(0.693)
SAT verbal	5.89(0.833)	5.91(0.864)
SAT math	6.22(0.873)	6.18(0.882)
family income /$10,000	4.15(1.35)	4.12(1.30)
a_i	1.55(0.727)	1.67(0.709)
$\bar{d}_i$	-0.111(0.241)	-0.0792(0.238)
took SAT	0.399	1.000
HS GPA	3.75(0.491)	3.69(0.502)
has HS GPA	0.785	0.945
transfer	0.396	0.284
female	0.483	0.479
white	0.575	0.588
n terms	6.23(3.19)	7.07(3.29)
n classes	26.2(12.9)	28.7(13.2)
total credits	76.4(37.0)	84.4(37.8)
degree seeking	0.992	0.997
n	63,875	24,435

Note: Means and standard deviations are in parenthesis for non-binary variables. This table shows the full sample before and after sample selection criteria are applied.

Table 3.5: Predicting outcomes with SAT and income

	I	
	GPA	a_i
SAT/100	0.147***	0.187***
	(0.005)	(0.005)
family income	−0.013**	−0.012**
/$10,000	(0.005)	(0.005)
R^2	0.10	0.16
obs	11,183	11,183

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. These are predictions of GPA and student fixed effect (described in body) using a the sum of verbal and math SAT and median income by zip code and are fit on the "enter" dataset (see text).

Table 3.6: Predicting outcomes with SAT or income

	GPA	a_i	GPA	a_i
SAT/100	0.148^{***}	0.187^{***}	—	—
	(0.005)	(0.005)		
family income	—	—	-0.021^{***}	-0.022^{**}
/$10,000			(0.008)	(0.008)
R^2	0.10	0.19	0.00	0.00
obs	11,183	11,183	11,183	11,183

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. These are predictions of GPA and student fixed effect (described in body) using a the sum of verbal and math SAT or median income by zip code and are fit on the "enter" dataset (see text).

Table 3.7: Predicting outcomes with SAT math and verbal separately

	I		II	
	GPA	a_i	GPA	a_i
SAT/100	0.147***	0.187***	—	—
	(0.005)	(0.005)		
SAT verbal/100	—	—	0.155***	0.143***
			(0.009)	(0.009)
SAT math/100	—	—	0.139***	0.229***
			(0.009)	(0.009)
faimly income	−0.013**	−0.012**	−0.013**	−0.011**
/$10,000	(0.005)	(0.005)	(0.005)	(0.005)
	contrasts			
SAT math/100	—	—	−0.016	0.086***
− SAT verbal/100			(0.018)	(0.018)
R^2	0.10	0.16	0.10	0.16
obs	11,183	11,183	11,183	11,183

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. Specification I is reproduced here for easy comparison to specification II.

Table 3.8: Predicting average class difficulty

	I	II
	$\bar{d}_i$	$\bar{d}_i$
SAT/100	0.036^{***}	—
	(0.001)	
SAT verbal/100	—	-0.023^{***}
		(0.003)
SAT math/100	—	0.093^{***}
		(0.003)
family income	0.003	0.003^{*}
/$10,000	(0.002)	(0.002)
	contrasts	
SAT math/100	—	0.116^{***}
− SAT verbal/100		(0.005)
R^2	0.07	0.11
obs	11,183	11,183

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. This table estimates the first three specifications (previous two tables) to explain average class difficulty for each student.

Table 3.9: Predicting outcomes with SAT: linearity of income

	I		IV	
	GPA	a_i	GPA	a_i
SAT/100	0.147***	0.187***	0.147***	0.187***
	(0.005)	(0.005)	(0.005)	(0.005)
family income	−0.013**	−0.012**	—	—
/$10,000	(0.005)	(0.005)		
family income lowest	—	—	0.020	0.006
			(0.027)	(0.025)
family income low	—	—	0.012	0.011
			(0.029)	(0.027)
family income middle	—	—	—	—
family income high	—	—	−0.013	−0.008
			(0.035)	(0.030)
family income highest	—	—	−0.038	−0.044*
			(0.026)	(0.024)
R^2	0.10	0.16	0.10	0.16
obs	11,183	11,183	11,183	11,183

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. This shows a test of the linearity of the zip code median income by breaking it down into five bins (the middle bin is the omitted group). Specification II is reproduced here for easy comparison with specification IV.

Table 3.10: Predicting outcomes with SAT including SATs imputed from ACT scores

	I	
	GPA	a_i
SAT/100	0.148^{***}	0.188^{***}
	(0.005)	(0.005)
family income	-0.014^{***}	-0.013^{***}
/$10,000	(0.005)	(0.005)
R^2	0.10	0.16
obs	11,753	11,753

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. This shows results when including those students who took the ACT and not the SAT but had their SAT imputed based on their ACT score.

Table 3.11: Predicting outcomes with SAT and income, ordered multinomial logit based ability estimate

	I	
	GPA	a_i
SAT/100	0.147^{***}	0.502^{***}
	(0.005)	(0.011)
family income	-0.013^{**}	-0.019
/$10,000	(0.005)	(0.012)
R^2	0.10	0.19
obs	11,183	11,183

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. Predictions of GPA and student fixed effect (described in body) using the sum of verbal and math SAT and median income by zip code. Fit on the "enter" dataset (see text).

Table 3.12: Predicting outcomes with SAT and income, degree dataset

	I	
	GPA	a_i
SAT/100	0.140***	0.189***
	(0.004)	(0.004)
family income	−0.015***	−0.013***
/$10,000	(0.005)	(0.005)
R^2	0.13	0.21
obs	8,017	8,017

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. This is the same model as was fit in Table 3.5 but it is fit on the smaller "degree" dataset (see text).

Table 3.13: Predicting outcomes with SAT and income, full dataset

	I	
	GPA	a_i
SAT/100	0.123^{***}	0.171^{***}
	(0.003)	(0.003)
family income	-0.012^{***}	-0.012^{***}
/$10,000	(0.004)	(0.004)
R^2	0.08	0.14
obs	24,435	24,435

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. This is the same model as was fit in Table 3.5 but it is fit on the larger "full" dataset (see text).

Table 3.14: Predicting outcomes with SAT and income, degree dataset, ordered multinomial logit ability

	I	
	GPA	a_i
SAT/100	0.140***	0.525***
	(0.004)	(0.011)
family income	-0.015^{***}	-0.025^{*}
/$10,000	(0.005)	(0.013)
R^2	0.13	0.23
obs	8,017	8,017

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. This is the same model as was fit in Table 3.5 but it is fit on the smaller "degree" dataset (see text) and the fixed effects taken from the ordered multinomial fit.

Table 3.15: Predicting outcomes with SAT and income, full dataset, ordered multinomial logit ability

	I	
	GPA	a_i
SAT/100	0.123^{***}	0.469^{***}
	(0.003)	(0.007)
family income	-0.012^{***}	-0.018^{**}
/\$10,000	(0.004)	(0.009)
R^2	0.08	0.17
obs	24,435	24,435

Note: The standard errors in parentheses reflect clustering at the zip code level. Stars indicate significance at the 1% (***), 5% (**), or 10% (*) levels. This is the same model as was fit in Table 3.5 but it is fit on the larger "full" dataset (see text) and the fixed effects taken from the ordered multinomial fit.

Chapter 4

Estimation of Large Ordered Multinomial Models

When estimating an equation with grades on the left hand side there are theoretical reasons to think an ordered multinomial estimator (OME) is preferable to the simpler ordinary least squares (OLS) estimator. An OME acknowledges that there is a maximum grade, so when a good students takes an easy class the prediction will not be higher than the an 'A' grade (an 'F' is treated in the same way for low predictions). At the same time, an OME acknowledges that spacing between grades need not be identical, so that the range of performance encompassed by 'B' grades need not be the same as 'C' grades. This is in contrast with OLS using the awarded grade points as the outcome variable which can, for example, predict that a student's grade be worth 5 grade points on a 4 point scale, and cannot accommodate different widths of grades.

I assume grades are generated as follows. Each student has a true ability (a_i). This ability level is then observed in an individual class as a noisy signal

$$\tilde{a}_{ij} = a_i + \epsilon_{ij}, \tag{4.1}$$

where ϵ_{ij} accounts for noise introduced into the observation in the grading process. The observed ability is not a grade but can be thought of as either a number in a ledger or possibly a yet unquantified assessment of a final paper. This is then mapped to a grade using “cutoffs” that separate observed ability levels ($\tilde{a}_{ij}$) into grades. For example, the ‘A’/‘B’ cutoff is the location where students with observed ability above the cutoff will receive ‘A’s and student just below the cutoff will receive a ‘B’s or lower grades.

Imagine two students, named 1 and 2, with ability levels a_1 and a_2, respectively (Figure 4.1, top scale). In a class c their observed ability level is given by a_{1c} and a_{2c} (Figure 4.1, middle scale). Based on their observed ability levels, these students are then assigned grades ‘F’ and ‘A’, respectively (Figure 4.1, bottom scale).

Theoretically, each class could have its own cutoffs for each grade–one class might have a small cutoff for each grade while another might allow a large range of observed ability to fall into an individual grade. However, the concept of difficulty is cleanest if all the classes share the size of the ranges for the values for each grade, but all the ranges move up and down together (Figure 4.2), and for tractability I will assume that this is true. Intuitively, this model allows for a university wide agreement on the size of the ‘B’ range as well as changes in the exact position of each of the grades.

Given their locations, one would expect student 1 to receive mainly ‘D’ grades and student 2 to receive mainly ‘B’ grades. The likelihood function for student 1 is given by the location of a_1, the distribution of ϵ, and the location of the grade boundaries. To illustrate this, Figure 4.3 shows the distribution of a_1, shading the area that is proportional to the probability of student 1 receiving a ‘D’ grade.

The OME estimates the model by maximizing the likelihood function where the probability of each grade is given by:

$$\begin{aligned} L\left(G|a,d\right)_{ij} &= \Pr\left(\text{grade } G_{ij} \text{ observed, conditional on } a_i, d_j\right) && (4.2)\\ &= \int_{G_{ij}} \Pr\left(\tilde{a}_{ij} - d_j = z\right) f(z)dz && (4.3)\\ &= \int_{G_{ij}} \Pr\left(a_i + \epsilon_{ij} - d_j = z\right) f(z)dz && (4.4)\\ &= \int_{G_{ij}} \Pr\left(\epsilon_{ij} = z - a_i + d_j\right) f(z)dz && (4.5) \end{aligned}$$

where the bounds of the integral are the cutoffs for the actual grade assigned to student i in class j, and $f(z)$ is the density function of the distribution of ϵ (Figure 4.3).[1] This leads to a regression of the form

$$G_{ij} = a_i - d_j + \epsilon_{ij} \tag{4.6}$$

where G_{ij} is the grade, a_i is the ability of the student, d_j is the difficulty of the class, i indexes the student and j indexes the class.

When the grading is more “difficult,” the observed ability level required to get each grade is higher. This corresponds to a rightward movement of the bottom scale in Figure 4.2. In the figure, class B is more difficult than class A because, for example, some observed performance levels that are assigned a ‘D’ in class A are assigned a ‘F’ in class B.

The terms “ability” and “difficulty” are used somewhat loosely here, because

[1] The models that describe these models are the ordered multinomial probit when ϵ is normally distributed and ordered multinomial logit when ϵ is Weibull distributed. To create a general term, I call these estimators ordered multinomial estimators (OME) and, treat the distribution of ϵ as a fitted parameter as in McCullagh & Nelder (1989).

so many factors play into them. For example, a class with good pedagogy might improve every student's output and thus would appear as a less difficult class *ceritis paribus*. Similarly, a student who spent more time on school work could potentially improve his or her "ability."[2]

Despite the advantage that this model describes the data generating process in a satisfying way, the OME can have biased estimates in the fixed effects context if the number of observations per unit (student) is not "large" (McCullagh & Nelder, 1989). However, the number of observations required to meet the threshold of being "large" is not well known.[3] Because of this, the OME is not obviously the best estimator.

Both the OLS and OME estimators have theoretical problems—OLS incorrectly models the grading process as continuous, which it is not, while the OME does not have an applicable consistency proof—so how exactly to proceed when estimating a grade decomposition is not obvious. The remainder of this chapter explores questions surrounding bias of the estimators in simulations (section 2), and the role of bias in significance tests for the OME (section 3), and the difficulty in estimation of the OME (section 4). A final section concludes.

4.1 Simulation of estimators

It is well known that a regression of the form (eq. 4.6) need not be consistent when there are fixed effects that increase the sample size. A literature has devel-

[2] This is something that appears to be possible given the low time investment made in college (Babcock & Marks, 2010).

[3] See Green (2002) for a review.

oped around estimating this type of problem (Ferrer-i-Carbonell & Frijters, 2004; Chamberlain, 1980). One related estimator is described by Chamberlain (1980). His method is intended for estimating parameters of interest in the presence of fixed effects for binomial and multinomial estimators, treating the fixed effects as nuisance parameters.[4] The method substitutes a term for each fixed effect by unit with conditioning on the margins/totals of the outcomes by unit.[5] A problem with this approach is that the conditional term involves calculating 10^6 to 10^{32} of terms per fixed effect–an infeasible task.[6] Substantial work has been done to minimize the computational effort while allowing for very relaxed assumptions about the data generating process by Ferrer-i-Carbonell & Frijters (2004). Even the least intensive of the non-linear methods described by Ferrer-i-Carbonell and Frijters requires far too much computation for problems with as many observations as a typical student in this data (circa 40 observations)

Another complicating factor is that unlike in the existing literature, the results herein are not regression parameters themselves, but relationships between them. The main results of interest include correlations between all possible pairs of GPA, ability, and difficulty, and the ability expansion path. Since the correlations are location and scale invariant, a bias in the estimator that doubled one set of fixed

[4] An issue that hampers but does not exclude the method from consideration, as described below.

[5] Chamberlain shows that the resulting estimator is consistent for the remaining parameters. However, herein, the fixed effects *are the parameters of interest.* Nonetheless, the method described by Chamberlain could be used to estimate the student fixed effect netting out the class fixed effects and then estimate the class fixed effects netting out the student fixed effects.

[6] A second method discussed by Chamberlain (1980) is a probit with random effects. While the model does not assume that the fixed effects are uncorrelated, it does treats the main results of this thesis (the relationship between the fixed effects) as a nuisance parameter problem in a way that is integral to its nature.

effects would not change the correlations. Because of this the exact properties of the estimator of the statistics I use are not well known. When faced with an estimator that is known to not have all the desirable properties, one possible way of quantifying its performance is simulation (Heckman, 1981; Green, 2002).

To investigate the properties of the OME and OLS, three cases are simulated by generating student and class fixed effects randomly using the OME model described above. Estimates of the fixed effects $\hat{a}$ and $\hat{d}$ using the OME and OLS estimators and statistics later calculated in the results are compared to their true value (in the simulation). The exact simulations are chosen to show how well these estimators perform at identifying bias in grades in simulations where there is and is not bias. The three different cases test these estimators in different situations (Table 4.1). In the first simulation, student ability and class difficulty are uncorrelated ("uncor"); in the second simulation, student ability and class difficulty are strongly correlated ("cor"); in the third simulation, the first simulation parameters are modified so that a large number of students receive 4.0 GPAs ("high GPA").

The parameters used to generate these simulations are also shown in Table 4.1. An individual class's grade can be described by its Grade Points (GP) and the average of GP is determined by the difference in $E(a)$ and $E(d)$,[7] and because of upper and lower limit effects, to a lesser extent by the variance covariance matrix of a and d. The cutoffs between grades are kept linear–the space between the b values are always exactly one–so the effect of unequal spacing in grades is not investigated.[8]

[7] Here expectations are taken over the population of transcript entries.

[8] Some investigation of the importance of non-linearities showed little difference between OLS and the OME in estimating when they were present.

In the results, I summarize the data with Pearson correlation coefficients between student ability (a_i) and average class difficulty ($\bar{d_i}$); the performance expansion path (a plot of GPA versus average class difficulty); and Goodman-Kruskal correlation coefficients ($\tau - G$) between class difficulty (d_j) and student ability (a_i). The relationship between estimated derived statistics (i.e., Pearson correlation coefficients, performance expansion path, and the Goodman-Kruskal correlation coefficients) and the true (simulated) values is used to judge the estimators.

Looking at the simulated results, both the Pearson type correlation coefficients (ρ) and the Goodman-Kruskal ($\tau - G$) are reliably estimated in almost every instance (Table 4.2). When the correlation between ability and GPA is high, the estimated correlation is high, and the converse is also true.

However, the standard deviations of ability and class difficulty are not as reliably estimated. For the "uncor" and "cor" simulations, the OME estimates of the standard deviation on student ability (a_i) are slightly high while none of the OLS estimates are accurate, with all of them being too small.

Removing the approximately 2,000 students (∼6%) with 4.0s (Table 4.3) does not negatively impact the estimates in the "uncor" and "cor" cases, but does improve the OME estimates of the parametric standard deviation and Pearson correlation.

Another summary of the data used in the results is the performance expansion path. The estimated and true performance expansion paths are plotted for the "uncor" simulation (Figure 4.4), the "cor" simulation (Figure 4.5), and the "High GPA" simulation (Figure 4.6). These figures are generated by separating the students into vigentiles (equally spaced twentieths) by true/estimated student ability

(a_i / $\hat{a}_i$) and the average GPA and average class difficulty ($\bar{d}_i$) are calculated within each bin. The 4.0s are always segregated into a single bin, so that there are 21 total bins.

These figures show that when grades are unbiased ("uncor") the estimated ability expansion path is vertical and that when the grades are biased by higher ability students taking harder classes ("cor"), the estimated ability expansion path is slanted to the right. In the results, these shapes indicate these outcomes.

In the "uncor" and "cor" simulation, both estimators find the true vertical line when uncorrelated (Figure 4.4) and the slanted line when correlated (Figure 4.5). The only exception is the high estimate from the OLS estimator near 4.0 for the "uncor" simulation and a mild attenuation of the slope in the correlated case.

In the case of the "High GPA", similar to "uncor", both estimators again perform well, correctly identifying the vertical performance expansion path (Figure 4.6). However, both estimators overestimate the size of a hook to the left, near 4.0. Interestingly, the OME then accurately estimates the average difficulty of classes for those with 4.0s.

These simulations show that both estimators are reliable, though there are situations where the OME is accurate while the OLS estimator is not. Removing 4.0s from the analysis is required to get accurate Pearson type correlation coefficients and standard deviation estimates, though a sufficiently large number will also bias the estimated performance expansion path near the top of the GPA scale.

4.2 Significance testing

Typical significance tests for multiple regressors rely on an F-test. This is not possible for the OME. Instead, a likelihood-ratio test (LR) can be used

$$LR = 2 * (\ell_l - \ell_s) \tag{4.7}$$

where ℓ is the log-likelihood of the fitted models, and subscripts l and s are for a larger and a nested smaller model, respectively. Under the null hypothesis that the k additional variables in a larger model are all equal to zero,[9] the likelihood-ratio (LR) test statistic is chi-square distributed with k degrees of freedom under the mild assumption that there are many observations for each of the k variables being tested (McCullagh & Nelder, 1989). Unfortunately, the term *many* is not well defined.

When there are not sufficient observations the likelihood ratio test statistic will not produce accurate p-values. This is because panel data analysis generates biased estimates of β when there are individual fixed effects and the outcome is discretized (Hahn & Newey, 2004). For estimates of the type

$$Y^*_{it} = X_{it}\beta + FE_i + \epsilon_{it}, \tag{4.8}$$

where i subscripts are used to indicate individuals in the panel and t subscripts indicate the time variable, FE_i is a fixed effect for each unit, and the observed outcome Y_{it} is a discretized version of Y^*_{it}. This bias is often described in a Taylor

[9] It is also possible to test a hypothesis that the values are not equal to zero, but I use the simple (and applicable) example of zero here.

series of the form

$$\hat{\beta} = \beta + \frac{B}{T} + o\left(\frac{1}{T^2}\right) \tag{4.9}$$

For related non-linear models, methods of removing the first order bias term $\left(\frac{B}{T}\right)$ exist, for example Hahn & Newey (2004).

The correlation coefficients between fixed effects are not biased by first order bias terms in the estimates of β. However, for looking at significance tests, this issue cannot be dismissed. Bias of the form in eq. 4.9 makes the likelihood ratio test non-central χ^2 distributed instead of central χ^2 distribution and the associated tests would be biased towards rejection. This would make the actual type I error rate higher than intended.

The remainder of this section details existing methods in the literature and their applicability to testing grades data.

One method for estimating a statistic of unknown distribution is to use the bootstrap (Efron, 1982) resampling scheme to estimate a confidence interval for any statistic, largely free of assumptions about the distribution of the error term in equation 4.8. In particular one might bootstrap the test statistic for the joint hypothesis that a set of fixed effects are all zero (Fox, 2008).

There are a few problems when using the bootstrap. First, the sample must be redrawn to represent the original sampling scheme–but when there is a census, as in this dissertation, it is not obvious how this can be done. Another problem with using the bootstrap is that it requires refitting the data about one hundred times, compounding the already difficult task of fitting the regressions.

Because of these difficulties, I did not use the bootstrap to attempt to perform the likelihood ratio (significance) tests.

Another method for constructing confidence intervals is empirical likelihood (Owen, 2001) which does not use normal theory in developing confidence intervals. However, this method requires solving a number of linear equations that increases with the sample size (Jing et al., 2009) and is therefore infeasible for a large dataset.[10]

There are also analytic methods of removing the bias in equation 4.9 (Hahn & Newey, 2004; Fernandez-Val, 2009), but these authors do not propose a method of testing joint significance of several parameters.

Finally, the jackknife is a well known method of removing the bias term from the estimates. There is very little theory for the jackknife for non-linear estimators, but as Wolter (1985) points out, since most non-linear systems behave like linear systems in the neighborhood of the solution, presumably the theorems regarding linear estimators suggest that the jackknife will still be helpful for non-linear estimators. In addition, both of the analytic methods papers also include simulations where the jackknife is included. In once case the jackknife estimator has lower bias than any other estimator but a higher standard deviation (Fernandez-Val, 2009). In another, the jackknife estimator always has the lowest bias *and* standard deviation (Hahn & Newey, 2004). While simulations of this type are only suggestive, certainly the jackknife is not an obviously inferior choice.

[10] The method presented in Jing et al. (2009) need not apply to the question at hand without additional proofs and so is not as useful for estimating a potentially misspecified model.

Using a block jackknife reduces the number of times that the statistic must be recalculated, and speeds calculation further. The test statistic is analogous to a ANOVA analysis where the variance-covariance matrix for the estimated values is used to test the joint hypothesis that several variables are simultaneously equal to zero (Duncan, 1978; Matloff, 1980). Constructing this test statistic is problematic in this context because they variance-covariance matrix would be very large and dense, and therefore require massive amounts of memory to store and use when there are many observations.

Instead, the suggestion of Fox (2008), presented in the context of bootstrapping, is to use each resample to generate the test statistic instead of the fitted values, and this idea can also be used for jackknife estimators.

The traditional jackknife estimator, applied to a maximum likelihood estimator for a statistic θ, is

$$\widehat{\theta}_{JK} \quad = \quad n\widehat{\theta}_{MLE} + \frac{(n-1)}{n}\sum_{i=1}^{n}\tilde{\theta}_{(i)} \tag{4.10}$$

where the subscript MLE is used to indicate the maximum likelihood estimator, and the subscript (i) indicates $\widehat{\theta}_{MLE}$ estimated with the ith jackknife replicate.[11]

In the context of this thesis, the "ith jackknife replicate" is the entire sample, but with one of the school years (indexed by i) of data removed. Thus one jackknife replicate would remove the 2004-2005 school year and reestimate the regression as if that year never occurred.

[11] For a more complete description of the jackknife, see Efron (1982).

The jackknife estimator for the likelihood ratio test is

$$\begin{aligned}\widehat{LR}_{JK} &= n\widehat{LR}_{MLE} + \frac{(n-1)}{n}\sum_{i=1}^{n}\widetilde{LR}_{(i)} && (4.11)\\ \widetilde{LR}_{(i)} &= 2(\ell_{l(i)} - \ell_{s(i)}) && (4.12)\end{aligned}$$

where $\ell_{l(i)}$ and $\ell_{s(i)}$ are the likelihoods of the large and small models with the ith jackknife replicate removed.

The issue of how to choose blocks persists since, similar to the bootstrap, blocks should be drawn according to the sampling scheme. The most defensible choice is to use an academic class year as a block. This also has the desirable property of leading to a small number of jackknife replicates.

4.2.1 Simulation

The theoretical critiques of the use of the likelihood ratio statistic state that there is a problem with the test but not the importantance nor magnitude of the problem. The question remains how poor the p-values will be for actual testing.

To answer this, I use a simulation. In the simulation data is generated according to

$$Y^*_{ij} = \epsilon_{ij} \tag{4.13}$$

where the observed ϵ_{ij} is independently and identically normally distributed with mean zero and standard deviation 1, Y is a discretized version of

$$Y = f(Y^*) \quad = \quad \begin{cases} F \ \texttt{if} & Y^* \leq -1 \\ C \ \texttt{if} & -1 < Y^* \leq 0 \\ B \ \texttt{if} & 0 < Y^* \leq 1 \\ A \ \texttt{if} & 1 < Y^* \end{cases} \tag{4.14}$$

I then perform two hypothesis tests, sharing a null but with different alternatives

$$\begin{aligned} H_0 &: Y^*_{ij} = \epsilon_{ij} \\ H_{A1} &: Y^*_{ij} = x_i + \epsilon_{ij} \\ H_{A2} &: Y^*_{ij} = z_j + \epsilon_{ij} \end{aligned}$$

Thus, in this simulation, the null hypothesis is true. There are I values of i and J values of j, and the parameter I is varied between eight and thirty-two while J is fixed at 500. In the simulation I test the hypothesis that all of the xs are zero jointly using the likelihood ratio test

$$LR_1 = 2 * (\ell_1 - \ell_0) \tag{4.15}$$

$$LR_2 = 2 * (\ell_2 - \ell_0) \tag{4.16}$$

where ℓ_0 is the likelihood when the fit take the form

$$Y^*_{ij} = \epsilon_{ij}, \tag{4.17}$$

ℓ_1 is the likelihood when the fit take the form

$$Y_{ij}^* = x_i + \epsilon_{ij}, \tag{4.18}$$

and ℓ_2 is the likelihood when the fit take the form

$$Y_{ij}^* = z_j + \epsilon_{ij}. \tag{4.19}$$

The cutoffs for the discretization are fit in each model as well.

Under the null, this test statistic should be χ^2 distributed with I or J degrees of freedom. These tests are run using the likelihood ratio test based on the maximum likelihood estimator (eq. 4.7) and the same statistic jackknifed (eq. 4.11).

Usually, a simulation can be criticized because the specific form of the simulation affects the outcomes. However, in this case the idea is to test the Type I error rate of an estimator, so the simulation generates data where the null hypothesis is true and finds the rate at which the significance test rejects the null hypothesis. There is only one way for the null hypothesis to be true, so the simulation captures this situation completely.

In the simulation, one thousand runs are done where data are generated and then the tests performed with a significance level of 0.1. I then tabulate the fraction of the time that the test was rejected. Ideally, this would be 10% of the time, or a 0.1 in the table. Values larger than this indicate a test that rejects the null hypothesis too often, that is, the test is not sufficiently conservative. A value smaller than this indicates a test that does not reject the null hypothesis often enough, the test is too conservative.

Because there are one thousand replicates, if the true value were 0.1, the standard error on the mean would be about 0.01, so values between 0.08 and 0.12 would be in a 95% confidence interval built around the null hypothesis. Thus values outside of this can be considered to not agree with an ideal value of 0.1.

The results of the simulation show that the likelihood ratio test behaves as one might suspect, it is too large (not appropriately protective against Type I error) while I is small, but then as I increases, it does work correctly as I becomes "large" and even becomes too conservative at $I = 32$. In contrast, the bias corrected jackknife tests generally reject about 10% of the time (Table 4.4), but this is rarely within the confidence interval, which extends about 0.02 above and below every value in the table.[12] The jackknife test statistics are close but probably not exactly distributed as described in the test-statistic, however they provide much more robust protection for small values of I.

4.3 Computation

Estimating fixed effects in a non-linear framework poses a computational challenge (Green, 2002; Heckman, 1981). The basic problem is to find the maximum

[12] The confidence interval extends 0.02 above and below 0.10 when the actual value is 0.10. In other cases on the table, this value is approximately correct.

likelihood estimator using the log-likelihood function for the OME

$$\ell(\beta; Y, X) = \sum_{mn} = Y_{mn} \log(\mu_{mn}(\beta; X)) \tag{4.20}$$

where m indexes the observations[13] and n indexes the possible grades; β is the regression coefficients (including the fixed effects); Y_{mn} is 1 if a student received the grade indexed by n in the observation indexed by m and is 0 otherwise; X are the regressors (such as the class and student fixed effects); and $\mu(\beta; X)$ is the probability of observing a particular grade, given X and β.

The algorithm used to maximize the likelihood is a Newton method optimizer. The basic algorithm finds roots (zeros) in $f(\cdot)$ by iteratively solving

$$\beta_{k+1} = \beta_k - [f'(\beta_k)]^{-1} f(\beta_k), \tag{4.21}$$

where k is the iteration index, $f(\beta) = \frac{\partial \ell(\beta)}{\partial \beta}$, the vector of first derivatives of the likelihood with respect to the regression coefficients[14] and $f'(\beta) = \frac{\partial^2 \ell(\beta)}{\partial \beta^2} = H(\beta)$, the Hessian matrix of second derivatives of the likelihood function with respect to the regression coefficients. This equation is easily verified as solving OLS in a single step by plugging in the first and second derivatives $f(\beta) = X^T(y - X\beta)$ and $f'(\beta) = -(X^T X)$.

In the case of the OME, assuming information equality (Nelder & Wedderburn, 1972), the Hessian matrix is a function of β, and so no such simplifications occur to

[13] An observation is a student taking a class and receiving a grade.

[14] Also called the first order conditions.

eq. 4.21

$$f'(\beta) = X^T[\Omega(\beta)]X, \tag{4.22}$$

where $\Omega(\cdot)$ is a block diagonal weighting matrix that varies with β.[15]

The typical method of maximizing this equation, essentially due to Fisher (Bliss, 1935), is to calculate the Hessian $H(\beta_k)$ at every step (Nelder & Wedderburn, 1972). However, this method requires calculating and inverting the Hessian matrix, a tasks that grows cubicly with the number of regressors. For the majority of the problems presented in this dissertation, a sparse matrix package provided with the R programing language makes this a tractable task. There are two ways this package improves performance, by decreasing storage space for a matrix from $o(m^2)$ to $o(k)$ where k is the number of non-zero entries, and secondly by using automated methods to ideally row reduce the Hessian matrix. However, some fits are too large for even these methods to handle. Storage of the inverse Hessian is an issue not improved by using sparse matrix methods because it is a dense matrix (there is no entry that is necessarily zero). For the largest problems in this paper this is not feasible so an alternative method must be used.

There are several methods proposed to get around this. Green proposes a method based on the binomial matrix inversion theorem that does not require inversion of or storage of a large matrix (Green, 2002). Both of these apply to a single set of fixed effects. Green's method cannot be readily extended to multiple

[15] The exact form of $\Omega(\cdot)$ appears in McCullagh & Nelder (1989).

sets of fixed effects. Another method proposed by Heckman estimates the equation in two steps (Heckman, 1981). In one step the likelihood function is maximized with respect to the fixed effects, and in the other the fixed effects are maximized with respect to all other parameters. Heckman's method has been successfully extended to a multiple fixed effects problem by Arcidiacono et al. with linear grades (Arcidiacono et al., 2011).

A third alternative is used by Abowd et al. who invert the Hessian matrix using graph theory (Abowd et al., 2002). While there is no general theory of the complexity of this problem, this method is too memory intensive for this application, perhaps because classes and students weave a denser interrelationship web than workers and firms.[16]

The L-BFGS[17] both avoids constructing the full inverse Hessian and keeps only a local approximation the the Hessian (Liu & Nocedal, 1989; Zhu et al., 1997). It does this by storing only the most recent updates to the Hessian, dropping those that are over a pre-determined number of steps old. Because results from many prior steps are thrown out, the Hessian approximation is always local to the current best guess solution, an advantage when the system in question is only locally quadratic. Also, because a rank two update is the outer product of two vectors, the vectors can be stored instead of the matrix and the storage requirement is only $O(n)$ instead of the $O(n^2)$ required for a Hessian matrix.[18]

[16] An automated version of this is used in the R Matrix package and it does not sufficiently speed up row reduction/inversion (Bates & Maechler, 2011).

[17] L-BFGS is a modified version of BFGS, so named because it was separately simultaneously discovered by four authors (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970).

[18] See Liu & Nocedal (1989); Zhu et al. (1997) for a description of the method of efficiently storing and using the approximate Hessian matrix.

4.3.1 Convergence

Identifying when an accurate fit has been achieved is difficult because, while it is true that the gradient goes to zero, finding this exact likelihood maximizer is not possible. Knowing that one is sufficiently close to a peak in the likelihood function is very difficult. Because of this I use two methods to confirm the maximization is complete.

Typically numerical methods are considered to have converged when

$$||\beta_{k+1} - \beta_k|| < \epsilon, \tag{4.23}$$

for some small value of ϵ. When used in this thesis, even extremely small values of ϵ were found to not be sufficient for this to be a valid criterion. Refitting the data many times from different starting locations resulted in large differences in β and $\ell(\beta)$ between the fits. A different convergence criteria is needed to assure the likelihood function is maximized.

Instead, the criterion that the gradient be very small is much more useful. I require

$$\left|\left|\frac{\partial \ell(\beta_k)}{\partial \beta}\right|\right|_1 = \sum_m \left|\frac{\partial \ell(\beta_k)}{\partial \beta_m}\right| \quad < \quad \upsilon \tag{4.24}$$

where the value of υ is set to a small value, and a subscript k is used to indicate the iteration number on the estimated value of β. However, the value of $||\frac{\partial \ell(\beta_k)}{\partial \beta}||_1$ does not decrease monotonically, so while I found this works , it does not *guarantee* convergence.

That this criterion was sufficient was checked by solving several of the optimization from five distinct starting points to verify that the results were nearly identical. In every case, it was observed that decreasing the value of υ forced the values of β closer together with a typical mean difference between β estimates being 0.01 for a relatively large value of υ of 100 . That is

$$\frac{1}{n}\sum_i |\beta_i - \beta_i'| < 0.01 \tag{4.25}$$

where $\beta \in \mathrm{I\!R}^n$, β_i is the ith value of β for one estimate and β_i' is the ith value of β for an estimate solved from a separate starting location.

4.4 Discussion and Conclusion

When estimating a decomposition of grades into student and class fixed effects, the best estimator is not *a priori* obvious. The OLS estimator, conditional on its assumptions, should produce accurate estimates. The OMM has more palatable assumptions but may not be an accurate estimator. Despite their shortcomings, simulations showed that the OMM and OLS do a good job of estimating grades decomposition for cases similar to those in the data considered herein.

One additional issue is that consistency of the likelihood ratio tests assumes that the number of fixed effects does not grow with the number of observations. This is not true in the case of student and class fixed effects where a first order bias is induced. The jackknife has the appealing properties of removing first order bias in estimates and being estimable relatively quickly. A second simulation regarding significance testing shows that when the number of observations per student is large, the likelihood ratio tests are not biased towards rejection but that regardless of the number of classes per student, the jackknifed likelihood ratio test is approximately correct but slightly biased towards rejection. The advantage of the jackknife estimate of the likelihood ratio test is that it provides equal levels of protection for large and small numbers of observations per student and is thus robust to small numbers of observations per student in a way that the non-jackknife likelihood ratio test is not.

These models are also difficulty to estimate. Sparse matrix techniques did not provide sufficient simplification but L-BFGS appears to be able to readily estimate these models when there are a large number of fixed effects for students and classes. I also found that a new criterion must be used for convergence–the total gradient, not the length of the last update, must be small. When this criterion is used, convergence is confirmed much more readily.

This class of model is estimable, the correlations can be trusted from OLS or OMM estimators, and likelihood ratio tests can be trusted when using the jackknife (knowing it is consistently somewhat under-conservative) or when there are more than about thirty observations per student.

Table 4.1: Simulation inputs.

Variable	"Uncor"	"Cor"	"High GPA"
avg. student GPA	3.02	3.10	3.82
$E(a) - E(d)$	2.67	2.67	4.29
$\text{Var}(a)$	1.0	1.0	1.0
$\text{Var}(d)$	1.0	1.0	1.0
$\text{Cov}(a, d)$	0	0.7	0
$b_{A/B}$	3	3	3
$b_{B/C}$	2	2	2
$b_{C/D}$	1	1	1
$b_{D/F}$	0	0	0
n_{stud}	32,000	32,000	32,000
classes per student	40	40	40
students per class	32	32	32
observations	1,280,000	1,280,000	1,280,000

Note: a and d are bivariate normally distributed, and the parameters of their distribution is a sufficient statistic for the average student GPA.

Table 4.2: Results of simulations.

	"Uncor"			"Cor"			"High GPA"		
	True	OME	OLS	True	OME	OLS	True	OME	OLS
σ_{a_i}	1.00	1.13	0.66	1.01	1.04	0.80	1.00	1.96	0.33
σ_{d_j}	0.79	0.81	0.51	0.73	0.74	0.57	0.79	0.84	0.23
$\sigma_{\bar{d}_i}$	0.12	0.13	0.08	0.67	0.68	0.52	0.12	0.13	0.04
$\tau - G_{a,G}$	0.51	0.52	0.51	0.24	0.27	0.27	0.59	0.62	0.62
$\tau - G_{d,G}$	0.42	0.42	0.42	−0.14	−0.14	−0.14	0.48	0.48	0.48
$\tau - G_{a,d}$	0.00	−0.01	−0.01	−0.74	−0.72	−0.72	0.00	−0.02	−0.02
ρ_{a_i,GPA_i}	0.95	0.91	0.99	0.87	0.93	0.93	0.85	0.65	0.99
$\rho_{\bar{d}_i,GPA_i}$	0.12	0.12	0.12	−0.83	−0.83	−0.84	0.10	0.10	0.11
$\rho_{a_i,\bar{d}_i}$	0.00	0.01	0.00	−0.99	−0.98	−0.98	0.00	0.04	0.00

Note: Top panel shows standard deviations of student ability (a_i), class difficulty (d_j), and average class difficulty for students ($\bar{d}_j$); the middle panel shows the Goodman-Kruskal correlation coefficients ($\tau - G$) between student ability a_i, grade (G) and student average class difficulty $\bar{d}_i$; the bottom panel shows the Pearson correlation between student GPA (GPA_i), student ability, and student average class difficulty.

Table 4.3: Results of simulations, 4.0s removed.

	"Uncor"			"Cor"			"High GPA"		
	True	OME	OLS	True	OME	OLS	True	OME	OLS
σ_{a_i}	0.98	1.02	0.66	1.00	1.04	0.80	0.86	0.87	0.34
σ_{d_j}	0.79	0.81	0.51	0.73	0.74	0.57	0.79	0.84	0.23
$\sigma_{\bar{d}_i}$	0.12	0.13	0.08	0.67	0.68	0.52	0.12	0.13	0.04
ρ_{a_i,GPA_i}	0.95	0.96	0.99	0.87	0.93	0.93	0.88	0.92	0.99
$\rho_{\bar{d}_i,GPA_i}$	0.12	0.12	0.12	−0.83	−0.83	−0.84	0.10	0.09	0.10
$\rho_{a_i,\bar{d}_i}$	0.00	0.00	0.00	−0.99	−0.98	−0.98	−0.03	−0.04	−0.01

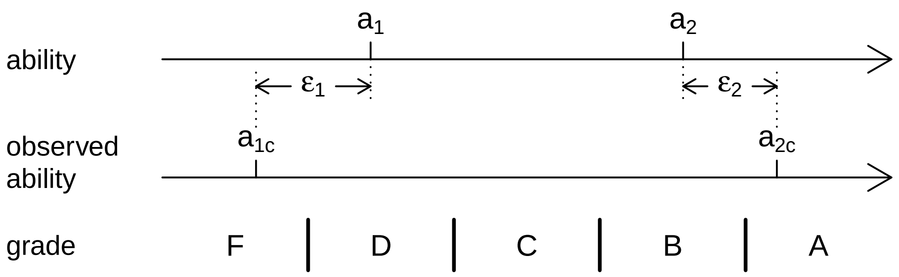
Note: For a description of the statistics, see Table 4.2.

Table 4.4: Fraction of simulations with p-values smaller than 0.10 for a χ^2 test for inclusion of fixed effects.

	LR test			Jackknife LR test		
	Individual (i) FEs					
I	n =1,000	2,000	4,000	1,000	2,000	4,000
8	0.694	0.625	0.586	0.076	0.072	0.028
16	0.150	0.101	0.058	0.105	0.083	0.068
32	0.076	0.021	0.011	0.134	0.095	0.050
	Class (j) FEs					
I	n =1,000	2,000	4,000	1,000	2,000	4,000
8	0.728	0.653	0.564	0.126	0.167	0.155
16	0.242	0.198	0.178	0.175	0.163	0.168
32	0.075	0.097	0.072	0.142	0.194	0.155

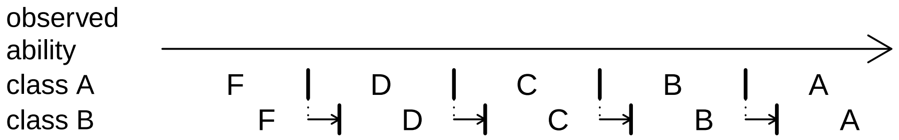
Note: For any cell, an ideal result is 0.10. Results higher than this indicate too many rejections of the null hypothesis while results lower than this indicate too few rejections of the null hypothesis. This is based on a model $Y_{ij} = a_i + d_j + \epsilon_{ij}$ analyzed using the ordered logit. In every case, the null hypothesis is true, that is there is no effect of a_i nor d_j on Y_{ij}. Because of this, p-values smaller than 0.10 should happen about 10% of the time, and an ideal value on this table is 0.10. The top portion shows results for removing fixed effects for n individuals when there are T (left most column) samples taken for each individual. The bottom portion shows results for removing 500 fixed effects for the classes (d_j). Both tests show substantial bias for low T with increases in n only slightly mitigating this effect. However, once T is as large as 32, there is little bias. In these cases the p-value is small so the test is actually conservative. For the jackknife, the test is often slightly positively biased, but the result does not systematically depend on T or n.

Figure 4.1: The relationship between ability, observed ability, and grades.



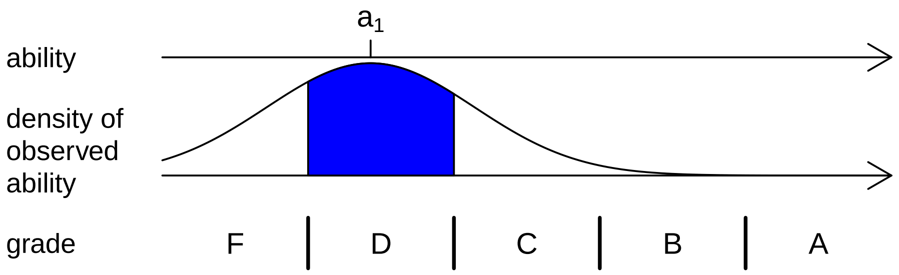
Note: Two students with ability a_1 and a_2 (top scale) have their ability observed plus an error term ($a_{1c} = \epsilon_{1c}$, $a_{2c} = \epsilon_{2c}$; middle scale) and are awarded discrete letter grades based on this observation ('F' and 'A', respectively; bottom scale).

Figure 4.2: Type of variation the estimator allows between classes.



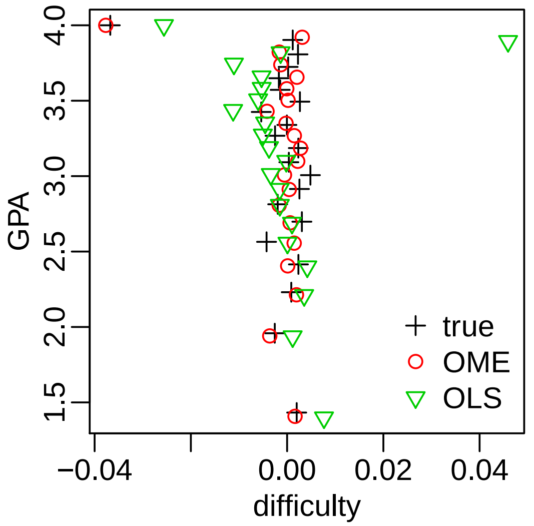
Note: The estimator used allows grades to vary by allowing each class to shift every letter grade up (or down) by the same amount. Here class B is slightly harder and so shifts every grade to the right by the same amount (represented by the arrows).

Figure 4.3: An example of the likelihood of a particular grade given observed ability.



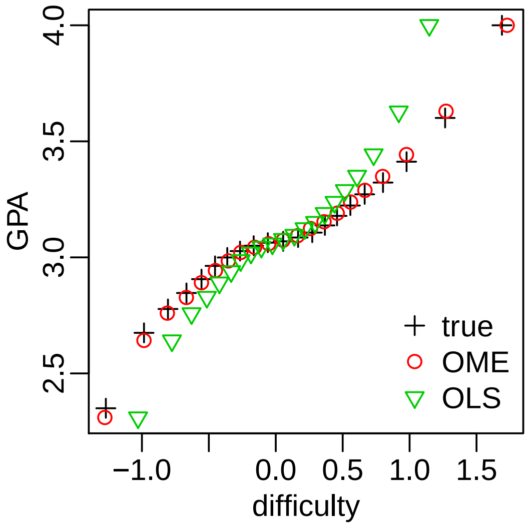
Note: The probability that a student with ability a_1 will receive a 'D' grade is proportional to the shaded area.

Figure 4.4: Simulated performance expansion path for the “uncor” simulation.



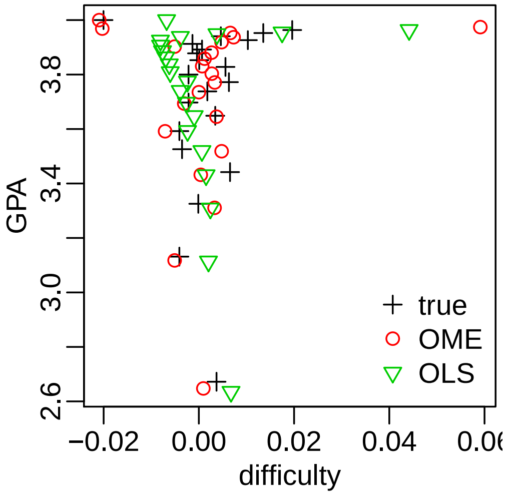
Note: Shows the “true” simulated results (black plusses), OME (red circles), and OLS (green triangles). Students with perfect 4.0 GPAs are separated into their own group (top left most point).

Figure 4.5: Simulated performance expansion path for the “cor” simulation.



Note: Shows the “true” simulated results (black plusses), OME (red circles), and OLS (green triangles). Students with perfect 4.0 GPAs are separated into their own group (top right most point).

Figure 4.6: Simulated performance expansion path for the “High GPA” simulation.



Note: Shows the “true” simulated results (black plusses), OME (red circles), and OLS (green triangles). Students with perfect 4.0 GPAs are separated into their own group (top left most point).

Appendix A

Data

The data for this thesis were provided by the Office of Institutional Research, Planning and Assessment (OIRPA) as two files: a transcript file and a application demographics file. On the transcript file, an observation is a single class that a single student took. So, for example, a student who enrolled in 24 classes would have 24 observations in the transcript file, each of which would contain the student's identifier, the class identifier and the resulting grade as well as several other variables (Table A.1). The file includes all undergraduate transcripts from the University of Maryland from Summer 2003 to Fall 2010. The transcript file is based on data taken from the registrar's office two weeks after the end of the semester and so these are not necessarily the grades that would appear on students' transcript because grades can change after this point.[1]

The demographic data has columns shown on Table A.3. There were some duplicates of `newid` but this was a technical problem because the values on the file were always identical. When duplicate records exist all of the records do not necessarily contain all of the information for the student, for example, an individual might have columns missing or null on some rows regarding that individual that are not missing on other rows regarding that individual. However, when data is present

[1] While I have no method of confirming this, the registrar's office said in an interview that grades do change after this point but it is not the common.

on two rows, that data is always identical. Because of this, all of the data from all rows is copied onto a single row and all other rows are removed.

The transcript and demographics files can be linked by students using the `newid` variable. This is an identifier generated by OIRPA for the purpose of this work, unique to a student and distinct from the social security number and university ID for the purpose of maintaining student anonymity.

The grades are not used exactly as they appear on the file. The letter grades are mapped to grade points per the University of Maryland standard (which does not use the plusses and minuses). Since the OME needs only assume that the values are ordinal, they are included in this specification. In addition, there are several things that can happen besides of award of a letter grade. All possible values of `crs_grade` are shown in Table A.2 along with how they are used.

Construction of degree, enter, and full samples

The data contains three extracts: "degree", "enter", and "full" which represent different levels of filtering the students. Some criteria were applied to all three datasets.

The following describes these criteria and the order in which they were applied, which impacts the final sample. Graduate classes (course number above 500) were removed because they are not typical college classes. Internships (course number 386 and 387) were removed because they are not typical college classes. Transcript observations where there is no letter grade were removed (as described above). Classes that did not award two "interior" grades ('B+' or lower and 'C-' or higher) were also excluded because they were essentially uninformative. Most of these classes awarded only the ‘A’s or ‘F’s but a small number awarded both ‘A’s and ‘F’s but not grades between the two. These classes appear to use a version of pass fail that is inconsistent with the model used for grades. Students who earned fewer than two interior grades were removed from the data. This was done because non-interior grades carry very little information and so these students fixed effects were not well estimated. Because the information content of their grades is low, the effect of removing them on other students or classes is small or zero. Courses that had a total of 5 students or fewer over the eight year sample were also removed because of their unusual nature. In addition, students with fewer than five courses on the data set were removed. Finally, students who were under about 18 or over about 25 at entrance were removed.[2] Students in both of these groups (under 18 and over 25) were observed to take substantially different courses than students within this age range. In particular, the students in the age extremes tended to be much more focused on a particular department than students between 18 and 25. The number of transcript observations removed by each of these criteria is shown in Table A.4. After applying these filters, an observation would be counted as part of the “full” sample.

The “enter” sample is the “full” sample after dropping students who entered after August of 2005 (and thus had fewer than five years to complete their degree) and dropping those who entered in the first observed term (Summer of 2003) with initial credits. Finally, students must have been degree seeking at some point in their career to be in the “enter” sample.

[2] These ages are approximate because the age variable is simply an integer with no “as of” date.

The "degree" sample is the "enter" sample with the additional criterion that student must appear to have graduated. In particular, the student must have 120 total credits (enrolled plus otherwise appearing on the transcript), or have enrolled for at least 8 terms. These criteria probably include a larger set than those students who graduated because students might enroll in 8 terms or 120 credits but still not complete a degree.

Construction of derived variables

Several derived variables are used from these dataset and the following describes their construction.

Each student's GPA is calculated as follows:

$$GPA_i = \frac{\sum_{A_i} GP_{ij} \times cr_{ij}}{\sum_{A_i} cr_{ij}}$$

where GP_{ij} is the grade points (from Table A.2) for student i in class j, and cr_{ij} is the number of credits student i was enrolled in for class j, and A_i is the set of all classes student i received a non-dropped grade in. The set A_i thus excludes classes that did not meet the selection criteria for the "full" sample (see above), even if a valid student enrolled in that class.

The following demographic variables were extracted from the transcript data by extracting the first (chronologically) time the student appears on the transcript data: the first term of enrollment, the number of previous credits at the time of enrollment, the age at entry. From the last term of enrollment, the final major and final term of enrollment are captured and recorded as demographic data.

The following demographic variables were extracted from the transcript data using all of the transcript data for an individual student: the number of terms (the sum of terms where the student is enrolled), the number of classes (the sum of the number of classes the student was enrolled in), if the student was ever listed as degree seeking (if any of the `ever_deg_seek` values were `TRUE`), if the student was ever enrolled (if any of the `ever_enrolled` values were `TRUE`).

Class year is calculated each semester using the total number of credits at the beginning of the semester (last_cum_cr_earn_ug, but called *cr* in the following equation):

$$\text{class year} = \begin{cases} \text{freshman} & \texttt{if} \quad cr < 30 \\ \text{sophomore} & \texttt{if} \quad 30 \leq cr < 60 \\ \text{junior} & \texttt{if} \quad 60 \leq cr < 90 \\ \text{senior} & \texttt{if} \quad 90 \leq cr \end{cases} \tag{A.1}$$

For the regressions, *total credits* is the sum of `crs_credits` on the full dataset by semester. Registered credits is the sum of `crs_credits` where the grading method is "Regular," meaning that the outcome was intended to be a grade. Examples where the outcome is not a grade for a class for which the grading method is"regular" are shown in Table A.2. Every value in that table except "audit" is possible when the grading method is "Regular."

Table A.1: Transcript columns.

Variable name	Description
`newid`	student ID generated by OIRPA
`term`	numeric term (ex: “200508” for Spring of 2005)
`term_transl`	a textual term description (ex: “Spring 2005”)
`course`	UMD course number (ex: “ENGL101”)
`section`	section number
`crs_grade`	the grade awarded in the course
`crs_credit`	number of course credits
`matric_entry_stat_ug`	matriculation status
`crs_grd_meth_cd`	the course grading method code (ex: “R” for regular)
`race_citz_cd`	race / citizen status code
`race_citz`	text description of race /citizen code
`last_cum_cr_earn_ug`	college cumulative credits before the term in question
`stu_campus_code`	the student’s campus code
`ug_gr_lev`	undergraduate level
`official_enrolled_ind`	indicator for if the student is officially enrolled
`major`	the first major of the student when taking this class
`student_type`	an indicator of participation in certain programs
`deg_seeking_ind`	indicator of if the student is degree seeking (matriculated)
`cls_stand_prior`	unknown
`coll_adv`	the college (ex: “BSOS”, the college the Economics department is in.)
`last_cum_gpa_ug`	the cumulative GPA prior to enrollment
`age`	the student’s age
`gender_cd`	the gender of the student

Note: Some of these variables are not used, such as `student_type` and `section`.

Table A.2: Transcript columns.

Grade	GP value	OME value
A+	4	A/A+
A	4	A/A+
A-	4	A-
B+	3	B+
B	3	B
B-	3	B-
C+	2	C+
C	2	C
C-	2	C-
D+	1	D
D	1	D
D-	1	D
F	0	F
withdraw completely	0	F
withdraw	0	F
academic dishonesty	0	F
pass (when taken pass/fail)	observation dropped	
satisfactory (when taken satisfactory/fail)	“	
fail (when taken pass/fail or satisfactory/fail)	“	
missing	“	
incomplete	“	
audit	“	

Note: *Withdraw completely* means that a student withdrew from all his or her classes. *Withdraw* means that a student withdrew from a single class.

Table A.3: Application / demographic columns.

Variable name	Description
newid	a student ID generated by OIRPA
zip5	the zip code from the student's permanent address
sat_high_verbal	SAT verbal score
sat_high_math	SAT math score
act_high_englsih	ACT english score
act_high_math	ACT math score
act_high_reading	ACT reading score
act_high_science	ACT science score
act_high_composite	ACT composite score
sat_recentered	if the SAT score is recentered (all are)
sat_recentered_cd	numeric code for previous
hs_acad_gpa	high school academic GPA
weighted_gpa_ind	unknown
high_school	name of the high school attended
high_school_cd	numeric code for previous
hs_class_rank_pct	high school class rank as a percentage
transfer_gpa	GPA at transfer institution
last_trans_inst	name of the last institution
last_trans_inst_cd	numeric code for previous

Note: All variables are as of application.

Table A.4: Observations removed by sample selection criteria.

Criteria	n remaining	n removed
no filters	1,840,212	—
graduate classes	1,836,364	3,848
internships	1,831,622	4,742
no grade	1,789,063	42,559
course awarded interior grades	1,776,799	12,264
student earned interior grades	1,685,971	90,828
course ever had more than five enrollees	1,685,345	626
student ever enrolled in five classes	1,675,859	9,486
age 18 to 25	1,621,707	54,152
“full” sample	1,621,707	—
“enter” sample	655,570	966,137
“degree” sample	523,151	132,419

Note: An observation is an individual transcript entry (ex: Sam takes organic chemistry and gets a ‘B’). The bottom section shows the number of observations in each of the samples. The number removed column is the difference between the n remaining column from that line and the one above it.

Bibliography

Abowd, J. M., Creecy, R. H., & Kramarz, F. (2002). Computing person and firm effects using linked longitudinal employer-employee data. Availble on the author's website.

Angrist, J. D., Lang, D., & Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, *1*(1), 136–63.

Arcidiacono, P., Foster, G., Goodpaster, N., & Kinsler, J. (2011). Estimating spillovers using panel data, with an application to the classroom. Available on the authors' website.

Babcock, P. S., & Marks, M. (2010). The falling time cost of college: Evidence from half a century of time use data. Working Paper 15954, National Bureau of Economic Research.

Bates, D., & Maechler, M. (2011). *Matrix: Sparse and Dense Matrix Classes and Methods*.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *The Journal of Political Economy*, *75*, 352–365.

Bettinger, E. P., Evans, B. J., & Pope, D. G. (2011). Improving college performance and retention the easy way: Unpacking the ACT exam. Working Paper 17119, National Bureau of Economic Research.

Betts, J. R., & Morell, D. (1998). The determinants of uncergraduate grade point average. *Journal of Human Resources*, *34*(2), 268–288.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, *22*(1), 134–167.

Broyden, C. G. (1970). The convergence of a class of double-rank minimizaiton algorithms. *Journal of the Institute of Mathematics and Its Applications*, *6*(1), 76–90.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Econmic Studies*, *47*(1), 225–238.

Cohn, E., Cohn, S., Balch, D. C., & Bradley Jr., J. (2004). Determinants of undergraduate GPAs: SAT scores, high-school GPA and high-school rankscores, high-school GPA and high-school rank. *Economics of Education Review*, *23*, 577–586.

College Board (2009). Total group profile report: Total group. Tech. rep., College Board.

Dale, S. B., & Krueger, A. B. (2002). Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics*, *117*(4), 1491–1527.

DeSimone, J. S. (2008). The impact of employment during school on college student academic performance. Working Paper 14006, National Bureau of Economic Research.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *journal of Educational Measurement*, *39*(1), 59–84.

Drew, C. (2011). Why science majors change their mind (it's just so darn hard). *The New York Times*, *November 4*.

Duncan, G. T. (1978). An empirical study of jackknife-constructed confidence regions in nonlinear regression. *Technometrics*, *20*(2), 123–129.

Eaton, B. C., & Eswaran, M. (2008). Differential grading standards and student incentives. *Canadian Public Policy*, *34*(2), 215–236.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philedelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Fernandez-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, *150*, 71–85.

Ferrer-i-Carbonell, A., & Frijters, P. (2004). How important is methodology for the estimates of the determinants of happiness? *The Economic Journal*, *114*(497), 641–659.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal*, *13*(3), 317–322.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. Los Angeles, CA: Sage, 2nd edition ed.

Freeman, D. G. (1999). Grade divergence as a market outcome. *Journal of Economic Education*, *30*(44), 344–351.

Geiser, S., & Studley, R. (2001). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the university of california. Tech. rep., University of California Office of the President.

Ghent, A. W. (1984). Examination of five tau variants suited to ordered contingency tables, from the viewpoint of biological research. *American Midland Naturalist*, *112*(2), 332–268.

Goldfarb, D. (1970). A family of variable metric updates derived by variational means. *Mathematics of Computation*, *24*(109), 23–26.

Green, W. (2002). The bias of the fixed effects estimator in nonlinear models. Available on the author's website.
URL http://pages.stern.nyu.edu/∼wgreene/

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica*, *45*(1), 1–22.

Grove, W. A., & Wasserman, T. (2003). The life-cycle pattern of collegiate GPA: Longitudinal cohort analysis and grade inflation. *Journal of Economic Education*, *35*(2), 162–174.

Hahn, J., & Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, *72*(4), 1295–1319.

Hamilton, B. W. (1975). Zoning and property taxation in a system of local governments. *Urban Studies*, *12*, 205–211.

Hanushek, E. A. (2006). School resources. In E. A. Hanushek, & F. Welch (Eds.) *Handbook of Economics of Education*, vol. 2, chap. 14, (pp. 865–908). Elsevier.

Heckman, J. J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process. In C. Manski, & M. D. (Eds.) *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press.

Jing, B.-Y., Yuan, J., & Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association*, *104*(487), 1224–1232.

Jones, E. B., & Jackson, J. D. (1990). College grades and labor market rewards. *Journal of Human Resources*, *25*, 253–266.

Klopfenstein, K., & Thomas, M. K. (2009). The link between Advanced Placement experience and early college success. *Southern Economic Journal*, *75*(3), 873–891.

Leonhardt, D. (2011). Top colleges, largely for the elite. *New York Times*.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large sample optimization. *Mathematical Programing*, *45*, 503–528.

Loury, L. D., & Garman, D. (1995). College selectivity and earnings. *Journal of Labor Economics*, *13*, 289–308.

Mankiw, G. (2011). A regression I would like to see. Part of Greg Mankiw's blog: Random Observations for Students of Economics.
URL http://gregmankiw.blogspot.com/2011/05/regression-i-would-like-to-see.html

Matloff, N. S. (1980). Algorithm as 148: The jackknife. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *29*(1), 115–117.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(2), 370–384.

Oates, W. E. (2005). The many faces of the tiebout model. In W. A. Fischel (Ed.) *The Tiebout Model at fifty: Essays in public Economics in Honor of Wallace Oates*, (pp. 21–45). Lincoln Institute of Land Policy.

Owen, A. B. (2001). *Empirical Likelihood*. No. 92 in Monographs in Statistics and Applied Probability. Chapman & Hall/CRC.

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

Sabot, R., & Wakeman-Linn, J. (1991). Grade inflation and course choice. *Journal of Economic Perspectives*, *5*(1), 159–170.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, *24*(111), 647–656.

Stinebrickner, R., & Stinebrickner, T. R. (2003). Understanding educational outcomes of students from low-income families: Evidence from a liberal arts college with a full tuition subsidy program. *Journal of Human Resources*, *38*(3), 591–617.

U.S. Census Bureau (2001). Census 2000 summary file 3. Tech. rep., U.S. Census Bureau.

Wainer, H. (1986). Five pitfalls encountered when trying to compare states on their SAT scores. *journal of Educational Measurement*, *23*(1), 69–81.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer Series in Statistics. New York: Springer.

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778. L-BFGS-B: FORTRAN subroutines for large-scale bound constrained optimization. *ACM Transactions on Mathematical Software*, *23*(4), 550–560.