University of California
Santa Barbara

Three Papers in Applied Microeconomics and Econometrics

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Economics

by

Valerie K. Bostwick

Committee in charge:

Professor Kelly Bedard, Co-Chair
Professor Douglas Steigerwald, Co-Chair
Professor Richard Startz

June 2016

The Dissertation of Valerie K. Bostwick is approved.

Professor Richard Startz

Professor Kelly Bedard, Co-Chair

Professor Douglas Steigerwald, Co-Chair

June 2016

Three Papers in Applied Microeconomics and Econometrics

by

Valerie K. Bostwick

Acknowledgements

I would like to give sincere thanks first and foremost to each of my committee members for all of their feedback, advise, and support. To Kelly Bedard, thank you for imparting all of your signaling model wisdom and especially for your guidance and reassurance during the stress of the job market. To Doug Steigerwald, thank you for your constant encouragement over the years and for helping me to realize my potential, not as just a student, but also as a research colleague. And to Dick Startz, thank you for reading all of my drafts at least twice as often as anyone else and still managing to provide always useful and (sometimes) witty feedback.

I would also like to thank the many graduate students who have participated in the Econometrics Research Group and Human Capital Working Group during my time at UCSB. Without their quarterly feedback, encouragement, and criticism this dissertation would almost surely be an unpolished disaster.

Valerie K. Bostwick

Department of Economics
University of California, Santa Barbara
2127 North Hall, Santa Barbara, CA 93106

vkbostwick@gmail.com
http://vkbostwick.weebly.com

EDUCATION

Ph.D. Economics, University of California, Santa Barbara, 2016
M.A. Economics, University of California, Santa Barbara, 2011
B.A. Economics, Princeton University, 2007, *Magna cum laude, minor in Political Economy*

RESEARCH INTERESTS

Economics of Education, Applied Econometrics, Labor Economics

PUBLICATIONS

Obtaining Critical Values for Test of Markov Regime Switching (with Douglas Steigerwald)
Stata Journal, 2014, vol. 14, issue 3, pages 481-498

Signaling in Higher Education: The Effect of Access to Elite Colleges on Choice of Major
Economic Inquiry, 2016, vol. 54, issue 3, pages 1382-1401

WORKING PAPERS

Saved By the Morning Bell: School Start Time and Teen Car Accidents

RESEARCH EXPERIENCE

Graduate Student Fellow, Broom Center for Demography, 2013 - current
Invited Member, Econometrics Research Group (with Douglas Steigerwald & Dick Startz), 2012-current
Invited Member, Human Capital Research Group (with Kelly Bedard & Dick Startz), 2012-current
Research Assistant for Douglas Steigerwald, 2012-2013
Research Assistant for Paulina Oliva, 2012
Research Assistant for Peter Rupert, Economic Forecasting Project, 2011-2012

TEACHING EXPERIENCE

Teaching Assistant:
- Econometrics (1st Year Ph.D. Sequence), 2014-2015
- Introduction to Probability & Statistics for Econometrics (Master's level), F2012 & F2013
- Principles of Microeconomics, 2011-2012, 2015-2016
- Principles of Macroeconomics, S2012, S2013, W2014, S2014, S2016

Laboratory Assistant, Broom Social Demography Lab, 2014-2015

PROFESSIONAL EXPERIENCE

Economic Research Analyst, Federal Trade Commission, Bureau of Economics, 2009-2010
Research Associate, Economists Incorporated, 2007-2009

SPECIAL PROGRAMS & CONFERENCES

Western Economic Association International 91st Annual Conference, 2016
Association for Education Finance and Policy 41st Annual Conference, 2016
All-California Labor Economics Conference, Graduate Poster Session, 2015
Labor Economics Lunch, University of California, Santa Barbara, 2013, 2014, 2015
Center for the History of Political Economy, Summer Institute on the History of Economics, Duke University, 2011

REFEREE SERVICE

Industrial and Labor Relations Review

FELLOWSHIPS & GRANTS

Travel Grant, Economics Department, University of California, Santa Barbara, 2016
Travel Grant, Broom Center for Demography, 2015
Andron Fellowship, University of California, Santa Barbara, 2010
Raymond K. Myerson Family Trust Graduate Fellowship, University of California, Santa Barbara, 2010

SPECIAL SKILLS

Computer: C, C++, R, Stata, Mata, SAS, MATLAB, ArcGIS
Language: Proficient in Spanish

Abstract

Three Papers in Applied Microeconomics and Econometrics

by

Valerie K. Bostwick

This dissertation is comprised of three distinct papers covering topics in applied microeconomics and applied econometrics. The first paper addresses a common problem faced by empirical researchers wishing to estimate Markov regime-switching models. For these models, testing for the possible presence of more than one regime requires the use of a non-standard test statistic. The analytic steps needed to implement the test of Markov regime-switching proposed by Cho & White (2007) are derived in detail in Carter & Steigerwald (2013). We summarize those implementation steps and address the computational issues that arise. A new Stata command to compute the regime-switching critical values, `rscv`, is introduced and presented in the context of empirical economic research. This paper is joint work with Douglas Steigerwald, and has previously appeared in the Stata Journal (Bostwick and Steigerwald, 2014).

In the second paper, I address a question in the field of economics of education: that is, whether college students use their choice of major as a signal of unobserved productivity in the labor market. I propose a model of postsecondary education in which major field of study can be used by individuals to signal productivity to employers. Under this signaling model, I show that geographic areas with high access to elite universities result in fewer science, technology, engineering, and mathematics (STEM) majors among lower ability students at non-elite colleges. Using data from the National Center for Education Statistics' Baccalaureate and Beyond survey, I find evidence that is consistent with the signaling model prediction, specifically a 2.3-3.7 percentage point (or 16-25%) decrease

in the probability of choosing a STEM major among lower ability students in areas with greater access to elite colleges. This paper has previously appeared in Economic Inquiry (Bostwick, 2016).

In the third paper, I analyze an unexpected consequence of a highly debated education policy. Many school districts are now considering delaying high school start times to accommodate the sleep schedules of teens. This paper explores whether such policy changes can have an impact on teen car accident rates. This impact could function both through a direct effect on teen sleep deprivation and indirectly through changes to the driving environment, i.e. shifting teen commute times into the high volume, "rush hour" of the morning. I find that, during the morning commute hours, any potential effect stemming from avoided sleep deprivation is offset by the effect of shifting teen driving into rush hour, so that a 15 minute delay in high school start times leads to a 21% increase in morning teen accidents. However, by focusing on late-night accidents, I also find evidence of a persistent sleep effect. By decreasing teen sleep deprivation, a 15 minute delay in school start times leads to a 26% decrease in late-night teen accidents.

Contents

Chapter 1

Obtaining Critical Values for Test of Markov Regime Switching

1.1 Introduction

Markov regime-switching models are frequently used in economic analysis and are prevalent in a variety of fields including finance, industrial organization, and business cycle theory. Unfortunately, conducting proper inference with these models can be exceptionally challenging. In particular, testing for the possible presence of more than one regime requires the use of a non-standard test statistic and critical values that may differ across model specifications.

Cho and White (2007) demonstrate that, due to the unusually complicated nature of the null space, the appropriate measure for a test of more than one regime in the Markov regime-switching framework is a quasi-likelihood ratio (QLR) statistic. They provide an asymptotic null distribution for this test statistic from which critical values should be drawn. Because this distribution is a function of a Gaussian process, the critical values are not easily obtained from a simple closed-form distribution. Moreover, the elements of the Gaussian process underlying the asymptotic null distribution are dependent upon one another. For this reason the critical values depend on the covariance of the Gaussian

process and, due to the complex nature of this covariance structure, are best calculated using numerical approximation. In this article we summarize the steps necessary for such an approximation and introduce the new Stata command, `rscv`, which implements the methodology to produce the desired regime-switching critical values for a QLR test of only one regime.

We focus on the case of a simple linear model with Gaussian errors, but the QLR test and the `rscv` command are generalizable to a much broader class of models. This methodology can be applied to models with mulitple covariates and non-Gaussian errors. It is also applicable to regime-switching models where the dependent variable is vector valued, although the difference between distributions must be in only one mean parameter. Although most regime-switching models are thought of in the context of time-series data, we provide an example in Section 1.5 of how the QLR test can be used in cross-section models. However, there is one notable restriction on the allowable class of regime-switching models. Carter and Steigerwald (2012) establish that the quasi-maximum likelihood estimator created by the use of the quasi-log-likelihood is inconsistent if the covariates include lagged values of the dependent variable. For this reason, the QLR test should be used with extreme caution on autoregressive models.

The article is organized as follows. In Section 1.2 we describe the unusual null space that corresponds to a test of only one regime versus the alternative of regime-switching. In Section 1.3 we present the QLR test statistic, as derived by Cho and White (2007), and the corresponding asymptotic null distribution. We also summarize the detailed analysis in Carter and Steigerwald (2013) describing the covariance structure of the relevant Gaussian process. In Section 1.4 we describe the methodology used by the `rscv` command to numerically approximate the relevant critical values. We also present the syntax and options of the `rscv` command and provide sample output. We illustrate use of the `rscv` command with an application from the economics literature in Section

1.5. Finally, we conclude with some remarks on the general applicability of this command and the underlying methods.

1.2 Null Hypothesis

Specification of a Markov regime-switching model requires a test to confirm the presence of multiple regimes. The first step is to test the null hypothesis of a single regime against the alternative hypothesis of Markov switching between two regimes. If this null hypothesis can be rejected, then the researcher can progress to estimation of Markov regime-switching models with two, or more, regimes. The key to conducting valid inference is then a test of the null hypothesis of a single regime, which yields an asymptotic size equal to or less than the nominal test size.

To understand how to conduct valid inference for the null hypothesis of only a single regime, consider a basic regime-switching model

$$y_t = \theta_0 + \delta s_t + u_t, \tag{1.1}$$

where $u_t \sim i.i.d.N\left(0, \sigma^2\right)$. The unobserved state variable $s_t \in \{0, 1\}$ indicates regimes: in state 0, y_t has mean θ_0, while in state 1, y_t has mean $\theta_1 = \theta_0 + \delta$. The sequence $\{s_t\}_{t=1}^n$ is generated by a first-order Markov process with $\mathbb{P}\left(s_t = 1 | s_{t-1} = 0\right) = p_0$ and $\mathbb{P}\left(s_t = 0 | s_{t-1} = 1\right) = p_1$.

The key is to understand the parameter space that corresponds to the null hypothesis. Under the null hypothesis there exists a single regime, with mean θ_*. Hence the null parameter space must capture all the possible regions that correspond to a single regime. The first region corresponds to the assumption that $\theta_0 = \theta_1 = \theta_*$, which carries with it the implicit assumption that each of the two regimes is observed with positive probability:

$p_0 > 0$ and $p_1 > 0$. The non-standard feature of the null space is the inclusion of two additional regions, each of which also correspond to a single regime, with mean θ_*. The second region corresponds to the assumption that only regime 0 occurs with positive probability, $p_0 = 0$, and that $\theta_0 = \theta_*$. Note that in this second region, the mean of regime 1, θ_1 is not identified, so that this region in the null hypothesis does not impose any value on $\theta_1 - \theta_0$. The third region is a mirror image of the second region, where now the assumption is that regime 1 occurs with probability 1: $p_1 = 0$ and $\theta_1 = \theta_*$ The three regions are depicted in Figure 1.1. The vertical distance measures the value of p_0 and of p_1 and the horizontal distance measures the value of $\theta_1 - \theta_0$. Thus the vertical line at $\theta_1 = \theta_0$ captures the region of the null parameter space that corresponds to the assumption that $\theta_0 = \theta_1 = \theta_*$ together with $p_0, p_1 \in (0, 1)$. The lower horizontal line captures the region of the null parameter space where $p_0 = 0$ and $\theta_1 - \theta_0$ is unrestricted. Similarly, the upper horizontal line captures the region of the null parameter space where $p_1 = 0$ and $\theta_1 - \theta_0$ is unrestricted.

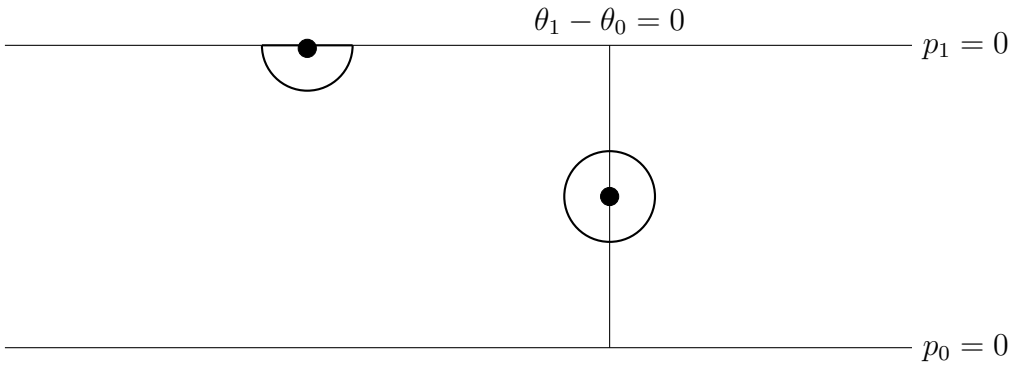


Figure 1.1: All three regions of the null hypothesis $H_0 : p_0 = 0$ and $\theta_0 = \theta_*; p_1 = 0$ and $\theta_1 = \theta_*$; or $\theta_0 = \theta_1 = \theta_*$ together with local neighborhoods of $p_1 = 0$ and $\theta_0 = \theta_1 = \theta_*$

The additional curves that correspond to the values $p_0 = 0$ and $p_1 = 0$ play a crucial role in guarding against the misclassification of a small group of extremal values as a second regime. In Figure 1.1 we depict the null space together with local neighborhoods

for two points in this space. These two neighborhoods illustrate the different roles of the three curves in the null space. Points in the circular neighborhood of the point on $\theta_1 - \theta_0 = 0$ correspond to processes with two regimes that have only slightly separated means. On the other hand, points in the semicircular neighborhood around the point on $p_1 = 0$ correspond to processes in which there are two regimes with widely separated means, one of which occurs infrequently. As one is often concerned that rejection of the null hypothesis of a single regime is due to a small group of outliers, rather than multiple regimes, including these boundary values reduces precisely this type of false rejection. Consequently, a valid test of the null hypothesis of a single regime must account for the entire null region and include all three curves.

1.3 Quasi-Likelihood Ratio Test Statistic

To implement a valid test of the null hypothesis of a single regime, a likelihood ratio statistic is needed. When considering the likelihood ratio statistic for a Markov regime-switching process, Cho and White (2007) find that the necessary inclusion of $p_0 = 0$ and $p_1 = 0$ in the parameter space creates significant difficulties in the asymptotic analysis. These difficulties lead them to consider a quasi-likelihood ratio (QLR) statistic for which the Markov structure of the state variable is ignored and $\{s_t\}$ is instead a sequence of i.i.d. random variables.

This i.i.d. restriction allows Cho and White to consider only the stationary probability, $\mathbb{P}(s_t = 1) = \pi$, where $\pi = p_0/(p_0 + p_1)$. Because $\pi = 1$ if and only if $p_1 = 0$ (and $\pi = 0$ if and only if $p_0 = 0$), the null hypothesis for a test of one regime based on the QLR statistic is expressed with three curves. The null hypothesis is, $H_0 : \theta_0 = \theta_1 = \theta_*$ (curve 1), $\pi = 0$ and $\theta_0 = \theta_*$ (curve 2), $\pi = 1$ and $\theta_1 = \theta_*$ (curve 3). The alternative hypothesis is $H_1 : \pi \in (0, 1)$ and $\theta_0 \neq \theta_1$.

For our basic model in (1.1), the quasi-log-likelihood analyzed by Cho and White is

$$L_n\left(\pi,\sigma^2,\theta_0,\theta_1\right)=\frac{1}{n}\sum_{t=1}^{n}l_t\left(\pi,\sigma^2,\theta_0,\theta_1\right), \tag{1.2}$$

where $l_t\left(\pi,\sigma^2,\theta_0,\theta_1\right):=\log\left(\left(1-\pi\right)f\left(y_t|\sigma^2,\theta_0\right)+\pi f\left(y_t|\sigma^2,\theta_1\right)\right)$ and $f\left(y_t|\sigma^2,\theta_j\right)$ is the conditional density with $j=0,1$. Define $\left(\hat{\pi},\hat{\sigma}^2,\hat{\theta}_0,\hat{\theta}_1\right)$ to be the parameter values that maximize the quasi-log-likelihood function. Let $\left(1,\tilde{\sigma}^2,\cdot,\tilde{\theta}_1\right)$ be the parameter values that maximize L_n under the null hypothesis that $\pi=1$. The QLR statistic is then

$$QLR_n=2n\left(L_n\left(\hat{\pi},\hat{\sigma}^2,\hat{\theta}_0,\hat{\theta}_1\right)-L_n\left(1,\tilde{\sigma}^2,\cdot,\tilde{\theta}_1\right)\right). \tag{1.3}$$

The asymptotic null distribution of QLR_n is (Cho and White, 2007, Theorem 6(b), p. 1692),

$$QLR_n\Rightarrow\max\left[\left[\max\left(0,G\right)\right]^2,\sup_{\Theta}\left[\mathcal{G}\left(\theta_0\right)_-\right]^2\right], \tag{1.4}$$

where $\mathcal{G}\left(\theta_0\right)$ is a Gaussian process, $\mathcal{G}\left(\theta_0\right)_-:=\min\left[0,\mathcal{G}\left(\theta_0\right)\right]$, and G is a standard Gaussian random variable that is correlated with $\mathcal{G}\left(\theta_0\right)$. (For a more complete description of (1.4) see Bostwick and Steigerwald (2012)).

The critical value for a test based on the statistic QLR_n thus corresponds to a quantile for the largest value over $\max\left(0,G\right)^2$ and $\sup_{\Theta}\left[\mathcal{G}\left(\theta_0\right)_-\right]^2$. In order to determine this quantity one must account for the covariance among the elements of $\mathcal{G}\left(\theta_0\right)$ as well as their covariance with G. The structure of this covariance, which is described in detail in Bostwick and Steigerwald (2012), is

$$\mathbb{E}\left[\mathcal{G}\left(\theta_0\right)\mathcal{G}\left(\theta_0'\right)\right]=\frac{e^{\eta\eta'}-1-\eta\eta'-\frac{\left(\eta\eta'\right)^2}{2}}{\left(e^{\eta^2}-1-\eta^2-\frac{\eta^4}{2}\right)^{\frac{1}{2}}\left(e^{\left(\eta'\right)^2}-1-\left(\eta'\right)^2-\frac{\left(\eta'\right)^4}{2}\right)^{\frac{1}{2}}}, \tag{1.5}$$

where $\eta = \frac{\theta_0 - \theta_*}{\sigma}$ and $\eta' = \frac{\theta_0' - \theta_*}{\sigma}$. The quantity $\sup_\Theta \left[\mathcal{G}(\theta_0)_-\right]^2$ that appears in the asymptotic null distribution is determined by this covariance. Since the regime-specific parameters enter (1.5) only through η, a researcher does not need to specify the parameter space Θ to calculate $\sup_\Theta \left[\mathcal{G}(\theta_0)_-\right]^2$. All that is required is the set H that contains the number of standard deviations that separate the regime means. Finally, in order to fully capture the behavior of the asymptotic null distribution of QLR_n, we must also account for the covariance between G and $\mathcal{G}(\theta_0)$. Cho and White show that $\text{Cov}(G, \mathcal{G}(\theta_0)) = \left(e^{\eta^2} - 1 - \eta^2 - \frac{\eta^4}{2}\right)^{-\frac{1}{2}} \eta^4$.

1.4 The rscv Command

1.4.1 Syntax

`rscv [, ll(`*value*`) ul(`*value*`) r(`*value*`) q(`*value*`)]`

1.4.2 Description

`rscv` simulates the asymptotic null distribution of QLR_n and returns the corresponding critical value. If no options are specified, `rscv` returns the critical value for a size 5 percent QLR test with a regime separation of ± 1 standard deviation calculated over 100,000 replications.

1.4.3 Options

`ll`(*value*) specifies a lower bound on the interval H containing the number of standard deviations separating regime means, where $\eta \in H$. The default value is -1.

`ul`(*value*) specifies an upper bound on the interval H containing the number of standard deviations separating regime means. The default value is 1.

`r`(*value*) specifies the number of simulation replications. The default value is 100,000.

`q`(*value*) specifies the quantile for which a critical value should be calculated. The default value is 0.95, which corresponds to a nominal test size of 5 percent.

1.4.4 Simulation Process

For a QLR test with size 5 percent, the critical value corresponds to the 0.95 quantile of the limit distribution given on the right side of (1.4). Because the dependence in the process $\mathcal{G}(\theta_0)$ renders numeric integration infeasible, we construct the quantile by simulating independent replications of the process. In this section, we describe the simulation process used to obtain these critical values and how each of the `rscv` command options affect those simulations.

As the covariance of $\mathcal{G}(\theta_0)$ depends only on an index η, we do not need to simulate $\mathcal{G}(\theta_0)$ directly. Instead we simulate $\mathcal{G}^A(\eta)$, which we will construct to have the same covariance structure as $\mathcal{G}(\theta_0)$. The process $\mathcal{G}^A(\eta)$ will therefore provide us with the correct quantile, while relying solely on the index, η.

To construct $\mathcal{G}^A(\eta)$ for the covariance structure in (1.5) recall that, by a Taylor-series expansion, $e^\eta = 1 + \eta + \frac{\eta^2}{2!} + \cdots$. Hence, for $\{\epsilon_k\}_{k=0}^{\infty} \sim i.i.d.N(0,1)$:

$$\sum_{k=3}^{\infty} \frac{\eta^k}{\sqrt{k!}}\epsilon_k \sim N\left(0, e^{\eta^2} - 1 - \eta^2 - \frac{\eta^4}{2}\right).$$

Using this fact, our simulated process is constructed as

$$\mathcal{G}^A(\eta) = \left(e^{\eta^2} - 1 - \eta^2 - \frac{\eta^4}{2}\right)^{-\frac{1}{2}} \sum_{k=3}^{K-1} \frac{\eta^k}{\sqrt{k!}}\epsilon_k,$$

where K determines the accuracy of the Taylor-series approximation. Note that the covariance of this simulated process, $\mathbb{E}\left[\mathcal{G}^A(\eta)\,\mathcal{G}^A(\eta')\right]$, is identical to the covariance

structure of $\mathcal{G}(\theta_0)$ in (1.5).

We must also account for the covariance between G and $\mathcal{G}(\theta_0)$. Cho and White (2007) establish that this covariance corresponds to the term in the Taylor-series expansion for $k = 4$. For this reason we set $G = \epsilon_4$ so that $\text{Cov}(G, \mathcal{G}(\theta_0)) = \text{Cov}\left(G, \mathcal{G}^A(\eta)\right)$. The critical value that corresponds to (1.4) for a test size of 5 percent is therefore the 0.95 quantile of the simulated value

$$\max\left\{\left[\max(0, \epsilon_4)\right]^2, \max_{\eta \in H}\left[\min\left(0, \mathcal{G}^A(\eta)\right)\right]^2\right\}. \tag{1.6}$$

The `rscv` command executes the numerical simulation of (1.6) by first generating the series $\{\epsilon_k\}_{k=0}^{K} \sim i.i.d.N(0,1)$. For each value in a discrete set of $\eta \in H$, it then constructs $\mathcal{G}^A(\eta) = \left(e^{\eta^2} - 1 - \eta^2 - \frac{\eta^4}{2}\right)^{-\frac{1}{2}} \sum_{k=3}^{K-1} \frac{\eta^k}{\sqrt{k!}}\epsilon_k$. The command then obtains the value $m_i = \max\left\{\left[\max(0, \epsilon_4)\right]^2, \max_\eta\left[\min\left(0, \mathcal{G}^A(\eta)\right)\right]^2\right\}$ corresponding to the right side of (1.4) for each replication (indexed by i). Let $\left\{m_{[i]}\right\}_{i=1}^{r}$ be the vector of ordered values of m_i calculated in each replication. The command `rscv` returns the critical value for a test with size q from $m_{[(1-q)r]}$.

For each replication, `rscv` calculates $\mathcal{G}^A(\eta)$ at a fine grid of values over the interval H. To do so requires three quantities: the interval H (which must encompass the true value of η), the grid of values over H (given by the grid mesh), and the number of desired terms in the Taylor-series approximation, K. The user specifies the interval H using the `ll` and `ul` options. If θ_0 is thought to lie within 3 standard deviations of θ_1, the interval is $H = [-3.0, 3.0]$. Because the process is calculated at only a finite number of values the accuracy of the calculated maximum increases as the grid mesh shrinks. For this reason the command `rscv` implements a grid mesh of 0.01, as recommended in Cho and White (2007, p. 1693). For the interval $H = [-3.0, 3.0]$, and with a grid mesh of 0.01, the process is calculated at the points $(-3.00, -2.99, \ldots, 3.00)$.

Given the grid mesh of 0.01 and the user-specified interval H, we must determine the appropriate value of K. To do so, consider the approximation error, $\xi_{K,\eta} = \left(e^{\eta^2} - 1 - \eta^2 - \frac{\eta^4}{2}\right)^{-\frac{1}{2}} \sum_{k=K}^{\infty} \frac{\eta^k}{\sqrt{k!}} \epsilon_k$. We want to ensure that, as K increases, the variance of $\xi_{K,\eta}$ is decreasing towards zero. Carter and Steigerwald (2013) show that, for large K, $\text{Var}(\xi_{K,\eta}) \leq e^{2J \log \eta - K \log K}$. The command `rscv` therefore implements a value of K such that, for the user-specified interval H, $\left(\max_H |\eta|\right)^2 / K \leq 1/2$.

The `rscv` command also allows the user to specify the number of simulation replications and the desired quantile. Note that for large values of H and the default number of replications ($r = 100,000$), the `rscv` command may require more memory than a 32-bit operating system can provide. In this case, the user may need to specify a smaller number of replications in order to calculate the critical values for the desired interval, H. Critical values derived using fewer simulation replications may be stable only to a single significant digit. Table 1.1 depicts the results of `rscv` for a size 5 percent test over varying values of `ll`, `ul`, and `r`.

Table 1.1: Critical values for linear models with Gaussian errors

	H	$[-1,1]$	$[-2,2]$	$[-3,3]$	$[-4,4]$	$[-5,5]$
Replications	100,000	4.9	5.6	6.2	6.7	7.0
	10,000	4.9	5.6	6.2	6.6	7.1

Nominal level 5 percent; grid mesh of 0.01.

1.5 Example

We demonstrate how to test for the presence of multiple regimes through an example that captures many features of empirical interest. Importantly, the example we study generalizes (1.1) in several important ways: both the intercept and a slope coefficient differ over regimes; the error variance differs across regimes; and the regime probability

depends on the covariates. For this very general model we first detail how to construct a QLR test statistic in Stata and then describe how to use the new Stata command `rscv` to obtain an appropriate critical value.

Our example is derived from Bloom et al. (2003), who test whether the large differences in income levels across countries are better explained by differences in intrinsic geography or by a regime-switching model of multiple equilibria with poverty traps. To this end, the authors use cross-sectional data to analyze the distribution of per capita income levels for countries with similar exogenous characteristics and test for the presence of multiple regimes.

Unlike the simple model, (1.1), that we have considered up until now, Bloom et al. present a model that includes several added complexities that are commonly used in regime-switching applications. These additions include covariates with coefficients that vary across regimes, as well as error variances that are regime-specific. The authors also allow the regime probabilities to depend on the included covariates.

Bloom et al. propose a model of regime-switching between two equilibria. Regime 1 occurs with probability $p(x)$ and corresponds to countries that are in a poverty trap equilibrium:

$$y = \mu_1 + \beta_1 x + \epsilon_1 \ , \ Var(\epsilon_1) = \sigma_1^2. \tag{1.7}$$

Regime 2 occurs with probability $1 - p(x)$ and corresponds to countries in a wealthy equilibrium:

$$y = \mu_2 + \beta_2 x + \epsilon_2 \ , \ Var(\epsilon_2) = \sigma_2^2. \tag{1.8}$$

In both regimes, y is log Gross Domestic Product (GDP) per capita and x is absolute latitude, which functions as a catchall for a variety of exogenous geographic characteristics.

This model is slightly different from a Markov regime-switching model in that the

authors are looking at different regimes in a cross-section, rather than over time. For this reason, the probability of being in either regime is stationary and the unobserved regime indicator is an i.i.d. random variable. These modifications correspond exactly to those made by Cho and White (2007) to create the quasi-log-likelihood, so that in this model the log-likelihood ratio and the QLR are one and the same.

To construct a QLR test statistic we must estimate the model under the null hypothesis of only a single regime and under the alternative hypothesis of two regimes. The most important aspect of constructing the test statistic is to understand what model should be estimated under the alternative hypothesis. As Carter and Steigerwald (2013) discuss, the asymptotic null distribution (1.4) is derived under the assumption that the difference between regimes be in only the intercept, μ_j. Thus, to form the test statistic the two regime model that is estimated is: regime 1 occurs with probability p and corresponds to

$$y = \mu_1 + \beta x + \epsilon, \tag{1.7'}$$

while regime 2, which occurs with probability $(1 - p)$ corresponds to

$$y = \mu_2 + \beta x + \epsilon, \tag{1.8'}$$

where $Var(\epsilon) = \sigma^2$. We have simplified (1.7) and (1.8) in three ways: the slope coefficient is constant across regimes; the variance of the error terms is constant across regimes; and the regime probability does not depend on the exogenous characteristics, x.

Simplifying the model in this way does not diminish the validity of the QLR as a test of a single regime for the model in (1.7) and (1.8). Note that under the null hypothesis of one regime there is necessarily only one error variance, only one coefficient for each covariate, and a regime probability equal to 1. Thus, under the null hypothesis, the QLR

test will necessarily have the correct size even if the data is accurately modeled by a more complex system. Following a rejection of the null hypothesis using this restricted quasi-log-likelihood, the researcher can then confidently proceed to estimate a model with regime-specific variances and coefficients, if desired.[1]

For (1.7′) and (1.8′) the quasi-log-likelihood is

$$L_n\left(p,\sigma^2,\sigma,\mu_1,\mu_2\right)=\frac{1}{n}\sum_{t=1}^{n}l_t\left(p,\sigma^2,\beta,\mu_1,\mu_2\right),$$

where $l_t\left(p,\sigma^2,\beta,\mu_1,\mu_2\right):=\log\left(pf\left(y_t|\sigma^2,\beta,\mu_1\right)+\left(1-p\right)f\left(y_t|\sigma^2,\beta,\mu_2\right)\right)$ and $f\left(y_t|\sigma^2,\beta,\mu_j\right)$ is the conditional density for $j=1,2$. It is common to assume as Bloom et al. do, that ϵ is a normal random variable,[2] so that $f\left(y_t|x_t;\sigma^2,\beta,\mu_j\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-1}{2\sigma^2}\left(y_t-\mu_j-\beta x_t\right)^2}$. Let $\left(\hat{p},\hat{\sigma}^2,\hat{\beta},\hat{\mu}_1,\hat{\mu}_2\right)$ be the values that maximize L_n and let $\left(1,\tilde{\sigma}^2,\tilde{\beta},\tilde{\mu}_1,\cdot\right)$ be the values that make L_n as large as possible under the null hypothesis. The QLR statistic is then

$$QLR_n=2n\left(L_n\left(\hat{p},\hat{\sigma}^2,\hat{\beta},\hat{\mu}_1,\hat{\mu}_2\right)-L_n\left(1,\tilde{\sigma}^2,\tilde{\beta},\tilde{\mu}_1,\cdot\right)\right). \tag{1.9}$$

To obtain $L_n\left(1,\tilde{\sigma}^2,\tilde{\beta},\tilde{\mu}_1,\cdot\right)$ we simply estimate a linear regression of y on x, which corresponds to maximizing

$$\frac{1}{n}\sum_{t=1}^{n}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-1}{2\sigma^2}\left(y_t-\mu_1-\beta x_t\right)^2}\right).$$

While the parameter estimates can be obtained with a simple OLS command, we need the value of the log-likelihood, so we detail how to use Stata commands to obtain this value.

[1] With a more complex data generating process these restrictions may however lead to an increase in the probability of failing to reject a false null hypothesis and, hence, a decrease in the power of the QLR test.

[2] Bloom et al. (2003) assume normally distributed errors but the QLR test also allows for any error distribution within the exponential family.

In what follows, we use the same Penn World Table and CIA World Factbook data as in Bloom et al. to test for the presence of multiple equilibria.[3] To find $\left(1, \tilde{\sigma}^2, \tilde{\beta}, \tilde{\mu}_1, \cdot\right)$, simply use the following code, which relies on the Stata command `ml`.

```
. program define llfsingle
  1. version 10.1
  2. args lnf mu beta sigma
  3. quietly replace ‘lnf’= (1/_N)*ln(((2*_pi*‘sigma’^2)^(-1/2))*exp((-1/
> (2*‘sigma’^2))*(lgdp-‘mu’-‘beta’*latitude)^2))
  4. end

. ml model lf llfsingle /mu /beta /sigma

. ml max

initial:       log likelihood =     -<inf>  (could not be evaluated)
feasible:      log likelihood =  -127.9261
rescale:       log likelihood = -31.297788
rescale eq:    log likelihood = -2.3397622
Iteration 0:   log likelihood = -2.3397622  (not concave)
Iteration 1:   log likelihood = -1.5887217  (not concave)
Iteration 2:   log likelihood = -1.2837809
Iteration 3:   log likelihood = -1.2491574
Iteration 4:   log likelihood = -1.1988511
Iteration 5:   log likelihood = -1.1982504
Iteration 6:   log likelihood = -1.1982487
Iteration 7:   log likelihood = -1.1982487

                                              Number of obs   =        152
                                              Wald chi2(0)    =          .
Log likelihood = -1.1982487                   Prob > chi2     =          .
```

[3]Latitude data for countries appearing in the 1985 Penn World Tables and missing from the CIA World Factbook comes from https://www.google.com/.

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mu	6.927805	1.420095	4.88	0.000	4.144469	9.711141
beta	.0408554	.049703	0.82	0.411	-.0565607	.1382714
sigma	.8019654	.5670751	1.41	0.157	-.3094815	1.913412

```
. mat gammasingle=e(b)
```

Using these estimates, we evaluate L_n at its maximum to find $L_n\left(1, \tilde{\sigma}^2, \tilde{\delta}, \cdot, \tilde{\mu}_2\right)$.

```
. gen llf1regime=ln(((2*_pi*gammasingle[1,3]^2)^(-1/2))*exp((-1/(2*gamma
> single[1,3]^2))*(lgdp-gammasingle[1,1]-gammasingle[1,2]*latitude)^2))

. quietly summ llf1regime

. quietly replace llf1regime=r(sum)

. disp "Final estimated quasi-log-likelihood for 1 regime: " llf1reg
Final estimated quasi-log-likelihood for 1 regime: -182.1338
```

Thus we have $n \cdot L_n\left(1, \tilde{\sigma}^2, \tilde{\beta}, \tilde{\mu}_1, \cdot\right) = -182.1388$.

Under the alternative hypothesis of two regimes, direct maximization is more difficult, as the quasi-log-likelihood involves the log of the sum of two terms:

$$L_n\left(p, \sigma^2, \beta, \mu_1, \mu_2\right) = \frac{1}{n}\sum_{t=1}^{n} \log\left(pf\left(y_t|\sigma^2, \beta, \mu_1\right) + (1-p)\, f\left(y_t|\sigma^2, \beta, \mu_2\right)\right).$$

The expectations-maximization (EM) algorithm provides a method for circumventing this difficulty. This algorithm requires iterative estimation of the latent regime probabilities, p, and maximization of the resultant log-likelihood function until parameter estimates converge. The EM algorithm proceeds as follows:

1. Choose a starting guess for the parameter values: $p^{(0)}, \sigma^{2(0)}, \beta^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}$

2. For each observation, calculate $\eta_t = \mathbb{P}(s_t = 1|y_t, x_t)$ such that

$$\hat{\eta_t} = p^{(0)} \frac{f\left(y_t|\sigma^{2(0)}, \beta^{(0)}, \mu_1^{(0)}\right)}{p^{(0)} f\left(y_t|\sigma^{2(0)}, \beta^{(0)}, \mu_1^{(0)}\right) + (1-p^{(0)}) f\left(y_t|\sigma^{2(0)}, \beta^{(0)}, \mu_2^{(0)}\right)}$$

3. Using Stata's `ml` command, find parameter values $p^{(1)}, \sigma^{2(1)}, \beta^{(1)}, \mu_1^{(1)}, \mu_2^{(1)}$ that maximize the complete log-likelihood:

$$L_n^C\left(p, \sigma^2, \beta, \mu_1, \mu_2\right) = \frac{1}{n}\sum_{t=1}^{n}(\hat{\eta_t} \log f\left(y_t|\sigma^2, \beta, \mu_1\right) + (1-\hat{\eta_t}) \log f\left(y_t|\sigma^2, \beta, \mu_2\right)$$

$$+(1-\hat{\eta_t})\log(1-p) + \hat{\eta_t}\log p)$$

4. To test for convergence, calculate

   (a) $\max((p^{(1)}, \sigma^{2(1)}, \beta^{(1)}, \mu_1^{(1)}, \mu_2^{(1)}) - (p^{(0)}, \sigma^{2(0)}, \beta^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}))$,

   (b) $|L_n^C\left(p^{(1)}, \sigma^{2(1)}, \beta^{(1)}, \mu_1^{(1)}, \mu_2^{(1)}\right) - L_n^C\left(p^{(0)}, \sigma^{2(0)}, \beta^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}\right)|$,

   (c) and (using numeric derivatives) $\max(L_n^{C\prime})$.

5. If all three convergence criteria are less than some tolerance level (we use $\frac{1}{n}$) then quit and use $p^{(1)}, \sigma^{2(1)}, \beta^{(1)}, \mu_1^{(1)}, \mu_2^{(1)}$ as final parameter estimates. Else, repeat Steps 2-5 with $p^{(1)}, \sigma^{2(1)}, \beta^{(1)}, \mu_1^{(1)}, \mu_2^{(1)}$ as the new starting guess.

The following code illustrates the implementation of these steps for the model at hand.

```
. program define llfmulti
  1. version 10.1
  2. args lnf mu1 mu2 beta sigma p
  3. quietly replace ‘lnf’= (1/_N)*((1-etahat)*(ln((2*_pi*‘sigma’^2)^(-1
```

```
> /2))+((-1/(2*‘sigma’^2))*(lgdp-‘mu2’-‘beta’*latitude)^2)+ln(1-‘p’))+et
> ahat*(ln((2*_pi*‘sigma’^2)^(-1/2))+((-1/(2*‘sigma’^2))*(lgdp-‘mu1’-‘be
> ta’*latitude)^2)+ln(‘p’)))
  4. end

. gen error=10

. gen tol=1/_N

. while error>tol {
  2. quietly replace f1=((2*_pi*gammahat[1,4]^2)^(-1/2))*exp((-1/(2*gamm
> ahat[1,4]^2))*(lgdp-gammahat[1,1]-gammahat[1,3]*latitude)^2)
  3. quietly replace f2=((2*_pi*gammahat[1,4]^2)^(-1/2))*exp((-1/(2*gamm
> ahat[1,4]^2))*(lgdp-gammahat[1,2]-gammahat[1,3]*latitude)^2)
  4. quietly replace fboth=gammahat[1,5]*f1+(1-gammahat[1,5])*f2
  5. quietly replace etahat=gammahat[1,5]*f1/fboth
  6. ml model lf llfmulti /mu1 /mu2 /beta /sigma /p
  7. ml init gammahat, copy
  8. quietly ml max
  9. mat gammanew=e(b)
 10. *Check for convergence using user-defined program nds
. nds
 11. quietly replace error=max(nd1,nd2,nd3,nd4,nd5)
 12. matrix gammahat=gammanew
 13. }

. ml display

                                                Number of obs   =        152
                                                Wald chi2(0)    =          .
Log likelihood = -1.4441013                     Prob > chi2     =          .
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mu1	6.532847	1.148891	5.69	0.000	4.281062	8.784632
mu2	7.813265	1.45266	5.38	0.000	4.966102	10.66043
beta	.0451607	.0374139	1.21	0.227	-.0281691	.1184905
sigma	.5986278	.4232938	1.41	0.157	-.2310127	1.428268

Using these estimates, we evaluate L_n at its maximum to find $L_n\left(\hat{p}, \hat{\sigma}^2, \hat{\delta}, \hat{\mu}_1, \hat{\mu}_2\right)$ and then calculate QLR_n.

```
. quietly replace f1=((2*_pi*gammanew[1,4]^2)^(-1/2))*exp((-1/(2*gammane
> w[1,4]^2))*(lgdp-gammanew[1,1]-gammanew[1,3]*latitude)^2)

. quietly replace f2=((2*_pi*gammanew[1,4]^2)^(-1/2))*exp((-1/(2*gammane
> w[1,4]^2))*(lgdp-gammanew[1,2]-gammanew[1,3]*latitude)^2)

. gen lf2reg=gammanew[1,5]*f1+(1-gammanew[1,5])*f2

. gen llf2regime=ln(lf2reg)

. quietly summ llf2regime

. quietly replace llf2regime=r(sum)

. disp "Final estimated quasi-log-likelihood for 2 regimes: " llf2regime
Final estimated quasi-log-likelihood for 2 regimes: -179.9662
```

Thus we have $n \cdot L_n\left(\hat{p}, \hat{\sigma}^2, \hat{\beta}, \hat{\mu}_1, \hat{\mu}_2\right) = -179.9662$. To calculate the test statistic,

```
. gen QLR=2*(llf2reg-llf1reg)
```

```
. disp "Quasi-Likelihood Ratio test statistic of 1 regime: " QLR
Quasi-Likelihood Ratio test statistic of 1 regime: 4.3352051
```

These estimates and the resulting QLR test statistic are summarized in Table 1.2. For the complete Stata code used to create Table 1.2, see Appendix A.

Table 1.2: QLR Test of Two Regimes vs. One Regime

	Single Regime	Two Regimes	
		Regime I	Regime II
Constant (μ_1, μ_2)	6.928	6.533	7.813
Latitude (β)	0.041	0.045	
SD of error (σ)	0.802	0.599	
Probability of Regime I (p)		0.771	
Log likelihood (L_n)	-182.1	-180.0	
QLR_n		4.3	

Finally, we use the `rscv` command to calculate the critical value for the QLR test of size 5 percent. We allow for the possibility that the two regimes are widely separated and set $H = [-5.0, 5.0]$. The command and output are shown below.

```
. rscv ,ll(-5) ul(5) r(100000) q(0.95)
  7.051934397
```

Given that this critical value of 7.05 exceeds the QLR statistic of 4.3, we cannot reject the null hypothesis of a single regime.

This result is consistent with the findings of Bloom et al., although they use a different method to obtain the necessary critical values. They report a likelihood ratio and the corresponding critical values for a restricted version of their model where the regime probabilities are fixed (p does not depend on x). Using this restricted model, the authors do not reject the null hypothesis of a single regime. At the time that Bloom et al. (2003) was published, researchers had yet to successfully derive the asymptotic null distribution

for a likelihood ratio test of regime-switching. For this reason, the authors employ Monte Carlo methods to generate their critical values using random data generated from the estimated relationship given by the model in (1.7) and (1.8). The primary disadvantage of this approach is that the derived critical values are then dependent upon the authors' assumptions concerning the underlying data generating process.

Bloom et al. go on to report a likelihood ratio test of a single regime model against the unrestricted model with latitude-dependent regime probabilities. Using the unrestricted model, the likelihood ratio and simulated critical values allow the authors to reject the null hypothesis in favor of the alternative of two regimes. Because the null distribution derived by Cho and White (2007) applies only to the restricted QLR presented in (1.9), we are unable to use the QLR test, and hence the `rscv` command, to obtain the critical values necessary to evaluate this unrestricted test statistic.

1.6 Discussion

For the case of a simple linear model with Gaussian errors, we provide a methodology and a new Stata command, `rscv`, to construct critical values for a test of regime-switching. Despite the complexity of the underlying methodology, the execution of `rscv` is relatively simple and merely requires the researcher to provide a range for the standardized distance between regime means. We demonstrate in Section 1.5 both how these methods can be generalized to a very broad class of models and the restrictions necessary to properly estimate the QLR statistic and utilize the `rscv` critical values.

Chapter 2

Signaling in Higher Education: The Effect of Access to Elite Colleges on Choice of Major

2.1 Introduction

It is a well-established fact that more educated individuals earn higher wages. There are two primary theories for the source of this education wage premium. The first is the human capital accumulation, or full information, model in which education directly enhances productivity and thereby increases an individual's wages. The second is the signaling model of the labor market first proposed by Spence (1973), wherein the monetary returns to schooling are partially explained by information asymmetries between individuals and employers. These two models lead to strikingly similar labor market predictions, which has made testing between them difficult. I provide a new test of the signaling model against the full information model using individuals' choices of college major. This test is novel in that it focuses on the quality of an individual's education as a potential signal, rather than the quantity of education. Furthermore, while other tests have focused on high school students or other specific groups (e.g. GED takers in Tyler et al. (2000) and MBA students in Hussey (2012)), this test sheds light onto the

decision-making process of college-going students.

In a full information model, productivity is costlessly observed by both the individual and the employer. Each individual chooses her optimal level of education to improve productivity and wages, given her ability and the marginal cost of schooling. In the signaling framework, information is asymmetric such that employers cannot perfectly observe individual ability or true productivity. Employers use education levels to infer expected productivity while each individual chooses her level of schooling to signal a higher innate ability and potential productivity to employers. The equilibrium result in both models is that higher ability individuals obtain more education and accordingly earn higher wages.

While the equilibria in both models appear similar, the extent to which individuals sort on ability depends on the amount of information asymmetry in the labor market. Under the signaling model, the individual's social and private returns to education do not necessarily coincide. The private return is a combined effect of the social return from increased productivity and the signaling effect of being identified (perhaps falsely) as a high ability individual. This disparity between private and social returns could lead to an inefficient allocation of schooling in the competitive labor market. The implication that investment in schooling might lead to a deadweight loss for society has prompted many attempts to distinguish between these two theories (Wolpin, 1977; Riley, 1979; Lang and Kropp, 1986; Bedard, 2001; Hussey, 2012) and to measure what portion of the returns to schooling are attributable to signaling effects (Altonji and Pierret, 1997; Tyler et al., 2000; Altonji and Pierret, 2001; Fang, 2006; Clark and Martorell, 2014). However, the empirical evidence differentiating the two models has been fairly limited and testing between the full information and signaling models of education has proven difficult.

I propose a new model, applying the signaling framework of Spence (1973) to the postsecondary environment. In this model, major field of study can be used by individuals

to signal unobserved ability and productivity to employers. I show that this leads to a prediction that geographic areas with high access to elite colleges result in fewer science, technology, engineering, and mathematics (STEM) majors among lower ability students at non-elite colleges. This is distinct from the prediction of a full information model where the level of access to elite schools should only affect those high ability individuals who are eligible to attend an elite school. Using data from the National Center for Education Statistics' Baccalaureate and Beyond survey and the geographic variation in access to elite schools across the U.S., I find evidence that is consistent with the signaling model prediction.

This paper adds to the existing literature on testing between the signaling and full information frameworks and specifically builds upon the the comparative statics approach used in both Lang and Kropp (1986) and Bedard (2001) to successfully identify signaling effects among high school students. Furthermore, this paper contributes new evidence that information asymmetries and signaling considerations continue to play a significant role in college-level decision-making and that the quality of education, in addition to the quantity, can successfully be used by students to signal unobserved ability. This evidence that students are using college major as an ability signal indicates that some portion of the substantial wage returns to college major choice (Daymonti and Andrisani, 1984; James et al., 1989; Grogger and Eide, 1995; Arcidiacono, 2004; Bettinger, 2010) can be attributed to private returns to the individual above and beyond the socially optimal value of the degree.

The remainder of this paper is organized as follows. Section 2.2 presents a signaling model of postsecondary education. Section 2.3 sets forth an empirical approach for testing the signaling and full information predictions. Section 2.4 discusses the data and Section 2.5 presents the results. Finally, Section 2.6 offers some concluding remarks.

2.2 Theoretical Framework

2.2.1 Asymmetric Information Model

Consider an environment in which every individual has an ability level, a_i, drawn from a continuous distribution, $f(a)$. Employers cannot directly observe an individual's ability level, but instead receive two potential signals: college quality (Q) and major choice (M). For ease of exposition, I focus here on the simple case of only 2 college types (Q_H = elite and Q_L = non-elite) and 2 major choices (M_H = STEM and M_L = non-STEM).[1]

In this context, it is somewhat simpler and more intuitive to model the college enrollment decision and the major choice decision sequentially. However, a simultaneous modeling of these two choices ultimately yields the same theoretical predictions, provided the following 2 conditions hold: (1) the equilibrium ability sorting is restricted to the case where the highest ability students choose (Q_H, M_H), followed by the next highest ability group choosing (Q_H, M_L), then (Q_L, M_H), and finally (Q_L, M_L); and (2) there is an excess supply of applicants to elite colleges. Given the symmetry between these two models, I address the simultaneous case fully in Appendix Section B.1 and proceed here with the sequential model.

In the first stage, each student applies to a range of schools and, ideally, attends the highest quality college that she is eligible for. Symmetrically, college admissions are based on ability alone, creating a strict cutoff point in the distribution of ability, a^{Q_H}. Of course, there will also be those high ability individuals, $a_i > a^{Q_H}$, who do not attend the best possible school that they qualify for. This could be due to the high financial cost or for other reasons such as a desire to stay close to home, etc. To account for this, I follow

[1]Given previous findings showing that quantitative human capital is a scarce resource that yields higher returns in the labor market (Paglin and Rufolo, 1990), the value M can be thought of as representing the quantitative component of each major.

Bedard (2001) and allow for a uniformly distributed constraint so that some fraction, $1-p$, of all eligible students are "directly constrained" from attending an elite college.

Once enrolled in school, the student must next decide on a field of study. Without loss of generality I assume that there is no human capital accumulation due to college quality and major choice so that, in equilibrium, firms set wages equal to expected ability. The individual's major choice problem can then be written as,

$$\max_{M_i} \mathbb{E}[a|Q_i, M_i] - C_{M_H}(a_i, Q_i). \tag{2.1}$$

The function $C_{M_H}(a_i, Q_i)$ represents the effort cost of choosing a STEM major, which depends on both ability and college quality. This cost is decreasing in ability, $\partial C_{M_H}(a_i, Q_i)/\partial a_i < 0$, as in the traditional Spence model (Spence, 1973).[2] As Spence points out, it is this decreasing cost assumption that is critical to ensuring that major choice serves as a distinguishing signal and leads to a separating equilibrium. Within each college quality level, students will choose to major in STEM provided that the added benefit, $\mathbb{E}[a|Q_i, M_H] - \mathbb{E}[a|Q_i, M_L]$, is greater than the added cost, $C_{M_H}(a_i, Q_i)$. This leads to two cutoff points in the ability distribution, $a_{M_H}^{Q_L}$ and $a_{M_H}^{Q_H}$, that represent the marginal individuals for whom the added benefit and cost of majoring in a STEM field are exactly equal. The two conditions forming the cutoff points that define the separating equilibrium can be written formally as:

$$\mathbb{E}[a|a^{Q_H} \leq a < a_{M_H}^{Q_H}] = \mathbb{E}[a|a \geq a_{M_H}^{Q_H}] - C_{M_H}(a_{M_H}^{Q_H}, Q_H), \tag{2.2}$$

$$\mathbb{E}[a|a < a_{M_H}^{Q_L}] = \psi(a) - C_{M_H}(a_{M_H}^{Q_L}, Q_L), \tag{2.3}$$

[2]The sign of the effect of college quality on cost is ambiguous from a theoretical perspective and does not have any effect on the model predictions of interest here.

where

$$\psi(a) = \frac{[F(a^{Q_H}) - F(a_{M_H}^{Q_L})]\mathbb{E}[a|a_{M_H}^{Q_L} \leq a < a^{Q_H}] + (1-p)[1 - F(a^{Q_H})]\mathbb{E}[a|a \geq a^{Q_H}]}{F(a^{Q_H}) - F(a_{M_H}^{Q_L}) + (1-p)[1 - F(a^{Q_H})]}. \tag{2.4}$$

This separating equilibrium is shown for a uniform ability distribution in the top panel of Figure 2.1. The function $\psi(a)$ is the expected ability/wages of individuals at non-elite colleges who choose STEM majors (individuals in the area labeled (Q_L, M_H) in Figure 2.1). This is an average of the abilities of the individuals who are not eligible to attend an elite school and then choose STEM, $a_{M_H}^{Q_L} < a_i < a^{Q_H}$, and the abilities of the directly constrained individuals, $a_i \geq a^{Q_H}$, weighted by their relative prevalence in the population at non-elite schools. It is implicit in (2.4) that the equilibrium cutoff for choosing a STEM major at non-elite schools is below the elite college admissions cutoff point, $a_{M_H}^{Q_L} \leq a^{Q_H}$. This requires an assumption on the upper bound of the cost of choosing STEM at non-elite schools; $C_{M_H}(a^{Q_H}, Q_L) \leq a^{Q_H} - \mathbb{E}[a|a < a^{Q_H}]$, such that it is optimal to choose a STEM major for at least the most able student who is not eligible to attend an elite college.[3] The separating equilibrium defined by conditions (2.2)-(2.4) is somewhat non-standard due to the constraint, p, but is similar to that in Bedard (2001). This equilibrium differs in the fact that the two major cutoff points are independently determined. This follows directly from the sequential nature of the decision-making process but is also true in a simultaneous model when there exists an excess supply of applicants to elite colleges (see Appendix Section B.1).

Next consider a scenario in which there is decreased access to elite colleges and therefore a larger fraction of eligible students who are directly constrained, $1-\tilde{p}$, where $\tilde{p} < p$. The bottom panel of Figure 2.1 shows the new separating equilibrium resulting from this change in the constraint. Identification of the signaling effect of this difference in the

[3] I address the implications of a violation of this assumption in Appendix Section B.3.

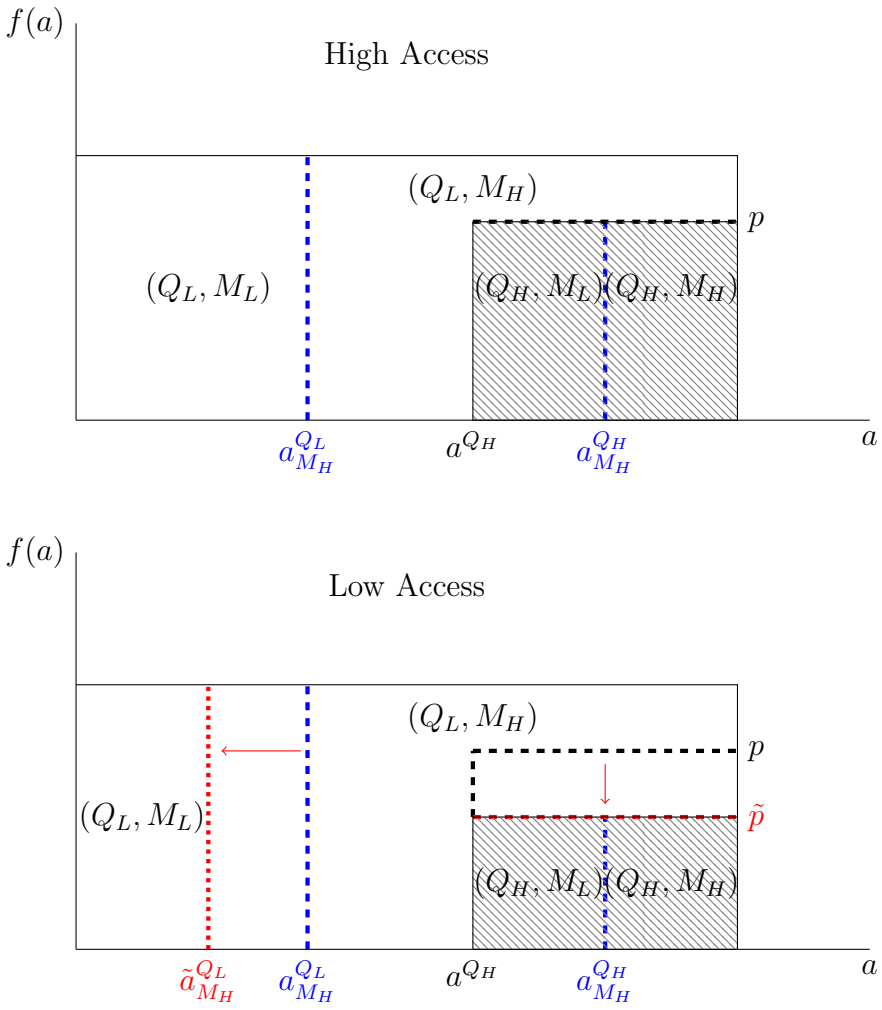


Figure 2.1: Separating equilibrium for uniform ability under 2 levels of uniform constraint

constraint relies on two primary assumptions:

Assumption 1 *The distribution of ability, $f(a)$, does not depend on the constraint level, p;*

Assumption 2 *The cost of majoring in a STEM field at a non-elite college, $C_{M_H}(a_i, Q_L)$, does not depend on the constraint level, p.*

Under these two assumptions, I show that an increase in the fraction of students who

are constrained leads to a negative shift in the non-elite major cutoff point, $a_{M_H}^{Q_L}$,[4] only in an asymmetric information framework.[5]

At elite colleges, the distribution of ability is unchanged; therefore the maximization problem in (2.1) and major cutoff point in (2.2) are also unchanged. For this reason, the major choice decisions of elite college students will not provide a testable implication of the signaling framework. Therefore, **I will focus exclusively on the implications for students at non-elite colleges for the remainder of this paper**. At non-elite schools the shift in the constraint causes an increase in the enrollment of individuals from the top end of the ability distribution. Given the existing major cutoff point, these high ability individuals will clearly choose STEM majors and will thus drive up the expected ability for all STEM majors at non-elite colleges, $\psi(a)$. Now consider the individual who is at the margin, $a_i = a_{M_H}^{Q_L}$, such that the added benefit of choosing STEM is exactly equal to the marginal cost when the constraint is equal to p. It is clear that the increase in the expected ability of STEM majors caused by the stricter constraint, $\tilde{p}$, will drive up the associated marginal benefit (and have no effect on the cost) thereby inducing the formerly indifferent individual to switch into a STEM field. This results in a negative shift in the major cutoff point in the non-elite schools from $a_{M_H}^{Q_L}$ to $\tilde{a}_{M_H}^{Q_L}$.

This shift in the non-elite major cutoff point can be shown mathematically by taking the total derivative of equations (2.3) and (2.4) and solving for $\frac{da_{M_H}^{Q_L}}{dp}$:

$$\frac{da_{M_H}^{Q_L}}{dp} = \frac{\partial\psi(a)/\partial p}{\partial\mathbb{E}[a|a < a_{M_H}^{Q_L}]/\partial a_{M_H}^{Q_L} - \partial\psi(a)/\partial a_{M_H}^{Q_L} + \partial C_{M_H}(a_{M_H}^{Q_L}, Q_L)/\partial a_{M_H}^{Q_L}}. \tag{2.5}$$

[4]Or conversely, a decrease in the fraction of constrained students will lead to a positive shift of the non-elite major cutoff point.

[5]I make a third, non-essential assumption that the admissions cutoff, a^{Q_H}, does not shift in response to the change in the constraint. Relaxing this assumption may lead to an even larger shift in the non-elite major cutoff, or it may have a mitigating effect on the cutoff shift. For details, see Appendix Section B.4.

The quantity in the denominator, $\partial \mathbb{E}[a|a < a_{M_H}^{Q_L}]/\partial a_{M_H}^{Q_L} - \partial \psi(a)/\partial a_{M_H}^{Q_L}$ $+\partial C_{M_H}(a_{M_H}^{Q_L}, Q_L)/\partial a_{M_H}^{Q_L}$, is the first derivative of condition (2.3), which must be negative under local stability of the equilibrium. The numerator, which is the direct effect of *decreasing* the fraction constrained on the expected ability of STEM majors at non-elite schools, can also be shown to be negative:

$$\frac{\partial \psi(a)}{\partial p} = \frac{[1 - F(a^{Q_H})][F(a^{Q_H}) - F(a_{M_H}^{Q_L})]}{\gamma}(\mathbb{E}[a|a_{M_H}^{Q_L} \leq a < a^{Q_H}] - \mathbb{E}[a|a > a^{Q_H}]), \quad (2.6)$$

$$\text{where } \gamma = F(a^{Q_H}) - F(a_{M_H}^{Q_L}) + (1-p)[1 - F(a^{Q_H})].$$

The leading fraction in (2.6) is a ratio of populations, which is clearly positive: $1 - F(a^{Q_H}) > 0$; $F(a^{Q_H}) - F(a_{M_H}^{Q_L}) > 0$; $\gamma > 0$. The quantity $(\mathbb{E}[a|a_{M_H}^{Q_L} \leq a < a^{Q_H}] - \mathbb{E}[a|a > a^{Q_H}])$ is the difference between the expected ability of STEM majors at non-elite schools who are below the admissions cutoff and the expected ability of all students who are above the admissions cutoff. It is clear that $\mathbb{E}[a|a_{M_H}^{Q_L} \leq a < a^{Q_H}] < \mathbb{E}[a|a > a^{Q_H}]$ so that $\mathbb{E}[a|a_{M_H}^{Q_L} \leq a < a^{Q_H}] - \mathbb{E}[a|a > a^{Q_H}] < 0$ and $\frac{\partial \psi(a)}{\partial p} < 0$.

Thus, (2.5) is the product of two negative values and $\frac{da_{M_H}^{Q_L}}{dp} > 0$, so that **the effect of increased access to elite colleges and a smaller fraction of students who are directly constrained is an increase in the non-elite major cutoff point.** Conversely, decreased access and a larger fraction of directly constrained students (as in the bottom panel of Figure 2.1) will have a negative effect on the non-elite cutoff point.

This same effect holds for a non-uniform constraint when the constraint is an increasing function of ability, $p(a)$. Given the availability of merit-based scholarships and grants, it seems likely that the most able students are also the least likely to face a constraint to entering their college of choice. The above calculations are somewhat complicated by this addition, but the results will hold under two conditions: (1) $p(a)$ must be increasing in ability (so that the fraction constrained, $1 - p(a)$, is smallest for the most able);

and (2) $p(a) < 1$ for all ability levels both before and after the change in the fraction constrained.[6] Note that this more flexible constraint, when applied to the simultaneous decision-making model described in Appendix Section B.1, will allow for the possibility that some students who would otherwise sort into non-STEM majors at elite colleges may decide to opt-out of attending the elite college in favor of majoring in a STEM field at a non-elite school (as long as the probability of making this choice is decreasing in ability).

2.2.2 Full Information Model

The predictions from a full information model are more straightforward. Under full information, employers can observe individual ability so that wages perfectly reflect marginal product (which is now a function of individual ability as well as the human capital accumulated from college quality and major choice). For this reason, differences in access to elite colleges will only affect the major choice for students who are directly constrained. Those individuals who are directly constrained from entering an elite college will instead attend a non-elite school and consequently choose a STEM major, as in the asymmetric information model. The key difference between the two models is that in the full information model, the decision of the marginal student who is indifferent between a STEM and non-STEM major at a non-elite college is unaffected by the proportion of high ability students in that school. For those students at non-elite colleges who are not eligible to attend the elite schools, regardless of the constraint, both wages and the effort cost of choosing a STEM major are unaffected (because wages equal marginal product rather than expected marginal product) and thus the major choice problem is also unaffected.[7] The key prediction of the full information model is that altering the level of

[6] See Appendix section B.2 for more detail.

[7] The cost of choosing a STEM major is unchanged by assumption (see Section 2.2.1, Assumption 2). If this cost is lower when the fraction of constrained students is higher then the full information

constraint does not shift the cutoff points in any way.

The shift in the STEM cutoff point at non-elite colleges that is unique to the signaling framework allows for a test between the full information and asymmetric information models. If there exist separate geographical areas such that the level of access to elite colleges (and therefore the level of the constraint) differs across these areas, then a positive (negative) shift in the STEM cutoff point at non-elite schools in the high (low) access areas is consistent only under the asymmetric information model.

An important caveat to the full information model prediction is that it does not account for potential peer effects. If lower ability students prefer to be in classrooms with high ability students (for reasons unrelated to signaling) then having additional high ability peers in the STEM fields at non-elite colleges in low access areas would make those majors more attractive to the marginal student. This could lead to a negative shift in the STEM cutoff point in the low access areas, even under full information, and make the two models indistinguishable. I cannot directly test for the presence of peer effects in this paper. However, existing research on peer effects and STEM major persistence indicate that this may not be a concern.

Recent research shows that having higher ability STEM peers actually *decreases* the probability of graduating with a STEM major (Fischer, 2015; Luppino and Sander, 2015). Luppino and Sander (2015) analyze a rich dataset covering all students at the 8 University of California campuses and find that students who attend campuses with higher ability peers in the STEM fields are less likely to graduate with a STEM degree. Similarly, Fischer (2015) finds that, for women, a higher percentage of high ability peers in an introductory STEM course decreases the probability of persistence in a STEM field (she finds no discernable peer effect on men). This is not entirely surprising given the

model will predict a negative shift in the major cutoff point, as in the asymmetric model. I address this possibility in Section 2.5 and show that the data does not suggest a difference in the cost of choosing a STEM field.

prevalence of rigid grading curves and highly competitive environments in STEM undergraduate courses. Other research by Ost (2010) finds that STEM persistence is highly influenced by relative grades. Students who receive lower relative grades in STEM vs. non-STEM courses are more likely to switch out of STEM majors. If an increase in the number of high ability peers in STEM courses leads to lower relative grades for the marginal students, then this research implies that many of those students would consequently leave the STEM majors. All of this evidence indicates that having additional high ability peers in the STEM fields at non-elite colleges in low access areas would make lower ability students less likely to major in a STEM field, thereby shifting the major cutoff point to the right. This is the opposite effect of that predicted by the signaling model in Section 2.2.1. Therefore, if there are significant peer effects at work, they will mitigate the signaling effect and I will observe no shift in the major cutoff, or perhaps even a shift in the opposite direction of the signaling model prediction. In this case I will not be able to reject the null hypothesis of a full information model.

2.3 Empirical Approach

Recall that the testable prediction provided by the two models above involves only the major choices of students at non-elite colleges, so the following empirical analysis excludes students at elite colleges and all equations are implicitly conditional on $Q_i = Q_L$. In order to conduct the test, I implement a simple probit model of the probability that a student at a non-elite college chooses a STEM major wherein an individual's latent ability is modeled by

$$a_i = X_i'\beta + \epsilon_i. \tag{2.7}$$

The vector X_i represents individual characteristics (such as age, sex, parents' education), which might influence the individual's aptitude for a STEM major and ϵ_i is a normally distributed error term. The observable outcome, y_i, is an indicator for whether individual i (attending a non-elite college) chooses a STEM major:

$$y_i = \begin{cases} 0 & \text{if } a_i \leq a_{M_H}^{Q_L} \\ 1 & \text{if } a_i > a_{M_H}^{Q_L}. \end{cases} \tag{2.8}$$

Under the asymmetric information model, the cutoff point, $a_{M_H}^{Q_L}$, is a function of being in a high or low access area so that,

$$a_{M_H}^{Q_L} = \alpha + \gamma H_i, \tag{2.9}$$

where H_i is an indicator for being in a high access area with a small fraction of directly constrained individuals. The signaling model predicts that γ will be positive (to reflect a positive shift in the cutoff point).

The probability of observing a student at a non-elite college in a STEM field is then,

$$\begin{aligned} \mathbb{P}\left(y_i = 1 | X_i, H_i\right) &= \mathbb{P}\left(a_i > a_{M_H}^{Q_L} | X_i, H_i\right) \\ &= \mathbb{P}\left(X_i'\beta + \epsilon_i > \alpha + \gamma H_i | X_i, H_i\right) \\ &= \mathbb{P}\left(\epsilon_i < -\alpha - \gamma H_i + X_i'\beta | X_i, H_i\right) \\ &= \Phi\left(-\alpha - \gamma H_i + X_i'\beta\right). \end{aligned} \tag{2.10}$$

Note the reversed signs on the parameters, α and γ, that enter into the equation for the

cutoff point. This indicates that estimation of a regression model of the form,

$$\mathbb{P}\left(y_i = 1 | X_i, H_i\right) = \Phi\left(\theta_0 + \theta_1 H_i + X_i'\beta\right), \tag{2.11}$$

will produce estimates such that $\theta_0 = -\alpha$ and $\theta_1 = -\gamma$. Therefore, in the results in Section 2.5, a negative estimate of θ_1 will indicate a positive shift in the cutoff point, $a_{M_H}^{Q_L}$, in high access areas, as is consistent with the asymmetric information model.

For identification of γ it is important that the high access indicator enters into the cutoff point alone and not also into the vector X_i. For this reason, I must restrict the estimation sample to individuals who are not directly affected by the constraint, i.e. $a_i < a^{Q_H}$. Otherwise the high access indicator, H_i, will enter directly into the vector X_i by removing mass from the upper end of the ability distribution.[8]

The primary empirical challenge is then identifying low access areas (those with a high fraction of constrained students) and high access areas (those with a smaller fraction of constrained students). It is reasonable to assume that proximity to a college decreases both the financial and psychic costs of attendance. The cost of attending a given college might be substantially reduced simply due to the lower financial costs of moving and holiday/summer travel. Furthermore, growing up in proximity to many elite colleges may lead to additional awareness of the benefits of attending such a school and perhaps additional information on how to best prepare for and apply to those colleges. A student is also more likely to attend a college close to his home, family, and existing social network in order to avoid the additional social and psychic costs associated with moving to a distant region. Therefore, I assume that students living near a large number of elite

[8]It is also important that the high access and low access regions satisfy Assumption 1 (see Section 2.2.1). If the distribution of ability differs between region types, then the variable H_i will enter directly into the vector X_i (even after the sample is limited to lower ability individuals) because it will add/subtract mass from the bottom end of the ability distribution. I address this possibility in Section 2.5 and show that the data does not suggest a difference in ability distributions across the two region types.

colleges are less likely to be constrained from entering an elite school and use geographic variation in proximity to elite colleges across the U.S. to identify high and low access areas. Existing research (Do, 2004; Griffith and Rothstein, 2009) and empirical data support this assumption; conditional on attending college, a freshman student is 4.6 times more likely to enroll at Yale University if they attended high school in Connecticut. Even compared to students from nearby Massachusetts, the Connecticut high schooler is 2.2 times more likely to enroll at Yale.[9]

Consider for example student i living in Massachusetts and student j living in Kansas who are both high ability students with $a_i = a_j > a^{Q_H}$. Living within driving distance of Boston and its myriad elite colleges, student i is more likely to consider applying to an elite school and will face less of a financial and emotional burden if she attends one of these schools than will student j. For this reason, it is more likely that i ends up attending an elite school while it is more likely that j attends a non-elite school in Kansas. Now consider student k and student l who also live in Massachusetts and Kansas, respectively, but who have ability $a_k = a_l < a^{Q_H}$. Neither student will be able to gain admission to an elite college. When these two students later enter the labor market employers will observe their major choice and both the quality and location of the colleges they each attended. Because student l attended a non-elite school in a low access region (Kansas), employers will determine his expected ability (and corresponding wage) based on the ability distribution in the bottom panel of Figure 2.1. As discussed in Section 2.2.1, the presence of additional high ability students (like student j) at non-elite schools in low access regions changes the incentive structure for their fellow classmates. The equilibrium result is that student l, who would have chosen a non-STEM major in a

[9]Data is from the Integrated Postsecondary Education Data System and references freshman enrollment from the fall of 2004. For all elite colleges (as defined in Section 2.4) freshmen are at least 1.9 times more likely to enroll if they attended high school in the same state and for some elite schools this probability ratio is as high as 230.

high access region (like student k), will instead be more likely to be motivated to choose a quantitative major in order to "blend in" with his constrained high ability classmates.

Note that this identification strategy assumes a national market for college graduates. However, to the extent that there are local labor markets, there may be general equilibrium effects on the wages of college graduates. If the increased supply of STEM graduates resulting from a higher fraction of constrained students has a general equilibrium effect, it would be to drive down the wages for STEM majors in the low access markets. This would make STEM a less attractive option and shift the non-elite major cutoff to the right, which would mitigate the negative shift predicted by the signaling model. Therefore, general equilibrium effects will only make it less likely that I will be able to detect any signaling effects in the asymmetric information model.

2.4 Data

I implement this empirical approach using data from the National Center for Education Statistics' Baccalaureate & Beyond Longitudinal Study (B&B), the Integrated Postsecondary Education Data System (IPEDS), and the U.S. News Best Colleges 2013 rankings (USN). I utilize the USN rankings data to create a measure of college quality.[10] The USN data include the 75th percentile SAT/ACT scores for each school's 2012 freshman class. I convert ACT scores into SAT scores using the correspondences suggested in Dorans (1999) and then sort schools based on the derived 75th percentile SAT score. Colleges in the top 5% of all 1,525 USN colleges are classified as elite schools and the remaining colleges are classified as non-elite. The list of all elite colleges under this

[10]The U.S. News rankings are divided into 10 separate categories (by Carnegie classification). Absent a method of comparing the #3 ranked National University with the #3 ranked Liberal Arts College, I cannot use the actual rankings as a metric for college quality.

definition is given in Appendix Table B.1.[11]

The B&B study surveys 3 cohorts of individuals who completed bachelor's degrees in 1993, 2000, and 2008. The data contain extensive information on the undergraduate experience, including institution attended and field of study. The data also contain demographic and background characteristics including: age at graduation, gender, race/ethnicity, parents' highest level of education, and SAT and ACT exam scores. Using the B&B data, I define STEM majors to include individuals in the following fields: mathematics, natural sciences, engineering, and computer science. Non-STEM majors include: all social sciences, liberal arts and humanities, education, business, and vocational fields.[12]

The combined 3 cohorts of the B&B data provide 32,490 observations[13] with complete institution, major, and demographic information and of this total, 28,100 attend non-elite colleges.[14] I use individual-level SAT scores from the B&B data to limit the sample to students who are unlikely to be directly constrained ($a_i < a^{Q_H}$). This derived score is also a combination of SAT scores and ACT scores that have been converted using the correspondence tables in Dorans (1999). For each of the 3 B&B cohorts and using the USN college quality measure, I find the derived SAT score, S^*, for which approximately 80% of students' scores at elite colleges are above S^*. This cutoff falls at 1,100 for the 1993 cohort, 1,160 for the 2000 cohort, and 1,200 for the 2008 cohort. I use these scores as

[11] In order to address any concern that the quality of these schools may be changing over time, I use IPEDS data on 75th percentile SAT score in 2001 (the earliest year available) to show the 2001 rankings of each school in columns 3 and 4 of Table B.1.

[12] There is some evidence that economics graduates have earnings and SAT scores much more in line with STEM graduates than with other social sciences graduates (Black et al., 2003). I define economics to be a non-STEM field in this analysis, however the main results are robust to including economics in the STEM category.

[13] All observation counts are rounded to the nearest 10, in compliance with NCES security requirements.

[14] 1,380 observations are dropped because they did not match with any institution within the USN data and an additional 1,050 observations are dropped because the institution did not report the 75th percentile SAT/ACT score for incoming freshmen. The remaining 1,970 students attended elite colleges.

the estimate for a^{Q_H} and limit the sample to students who attend a non-elite college and who have a derived SAT score less than the a^{Q_H} cutoff corresponding to their cohort.[15] This provides a final sample size of 22,330 non-elite college students who are below the a^{Q_H} cutoff.

I employ the IPEDS data on each institution's latitude and longitude coordinates as well as each school's total freshman enrollment in Fall 2007 to identify which non-elite colleges are in high or low access areas. I first calculate both the number of elite colleges within a 100-mile radius of each non-elite school and the % of total freshman enrollment ("open seats") within that same radius that is at elite colleges. Each non-elite institution is then categorized as being in either a high or low access area based on various combined threshold values of these 2 measures of local elite college access. Using both of these measures ensures that I capture the effect of many small elite liberal arts colleges in an area as well as the effect of a single elite institution that is relative large compared to the local non-elite alternatives. One advantage of this identification strategy is that it allows for the inclusion of state fixed effects in the regression model so that instead of comparing students in Alabama to students in California, I compare students in Bakersfield, CA to students in San Francisco, CA. The choice of a 100 mile radius is meant to encompass the credible set of options for the average student attending each non-elite college and is consistent with the raw data. In the B&B sample of students who attend non-elite colleges, the median distance between a student's college and his hometown is 45 miles and over 75% of these students attend a school within 150 miles of their parents' residence.[16] As a robustness check, Section 2.5 provides two additional identification strategies: one that classifies each U.S. state as high or low access and

[15]Students who did not report an SAT/ACT score are presumed to be below the a^{Q_H} cutoff and are included in the main sample. I show that the main results are robust to excluding these individuals and to changes in the cutoff threshold in Section 2.5.1

[16]This excludes all student whose distance to home is greater than 3,500 miles to rule out international students.

another that identifies groups of states (regions) as high or low access.

The model to be estimated via Maximum Likelihood is:

$$\mathbb{P}\left(STEM_{icst}=1\right)=\Phi\left(\alpha+\beta HighAccess_{c}+\delta'X_{i}+\theta'Z_{c}+\gamma_{s}+\eta_{t}\right), \tag{2.12}$$

where the variable $STEM_{icst}$ is an indicator that equals 1 if individual i at college c located in state s from cohort t chooses a STEM major. The variable $HighAccess_{c}$ is a measure of how many elite colleges exist in the 100 mile radius surrounding the college attended and/or the percent of local college enrollment that belongs to elite insitutions. The vector X_i captures individual characteristics including: age, gender, race/ethnicity, and parents' education. The vector Z_c captures school-level characteristics, namely indicators based on the school's carnegie classification and indicators for the level of urbanization in the area surrounding the college. The vector γ_s captures all time-invariant state fixed effects. The standard errors are estimated using the cluster-robust estimator with clustering at the institution-level.[17] Recall from Section 2.3 that if the estimate of β is negative, then the data supports the prediction of the signaling model.

[17]Standard errors estimated with clustering at the state-level are somewhat smaller than those reported in the tables below. Thus, the reported estimates can be considered a conservative bound on variance of the estimator.

Table 2.1: Summary Statistics

	Institutions[1]		States		Regions	
	High Access	Low Access	High Access	Low Access	High Access	Low Access
$\mathbb{P}(STEM = 1)$	0.12	0.16	0.14	0.17	0.14	0.16
Age	25.2	25.6	25.4	25.7	25.3	25.8
Female	0.62	0.59	0.6	0.58	0.6	0.58
SAT Score	693.8	642.5	676.8	608.5	694.9	586.7
Missing SAT Score	0.28	0.34	0.3	0.37	0.28	0.40
Race/Ethnicity:						
White	0.71	0.79	0.76	0.8	0.76	0.80
Black	0.084	0.091	0.092	0.083	0.096	0.077
Hispanic	0.11	0.061	0.073	0.072	0.074	0.072
Asian	0.07	0.031	0.05	0.023	0.049	0.028
Parents' Highest Education:						
High School or Less	0.30	0.28	0.29	0.29	0.29	0.29
Some College	0.23	0.23	0.24	0.22	0.24	0.22
Bachelor's Degree	0.24	0.25	0.24	0.25	0.24	0.25
Master's Degree or Higher	0.23	0.24	0.24	0.23	0.24	0.23
Institution Type:[2]						
National University	0.36	0.50	0.44	0.51	0.44	0.51
Liberal Arts College	0.046	0.041	0.048	0.031	0.049	0.032
Regional College	0.046	0.071	0.054	0.087	0.064	0.065
Regional University	0.55	0.39	0.46	0.37	0.45	0.40
Urbanization Level:						
Large City	0.30	0.18				
Mid-size City	0.27	0.40				
Urban Fringe of Large City	0.25	0.085				
Urban Fringe of Mid-size City	0.095	0.071				
Large Town	0.009	0.075				
Small Town	0.032	0.15				
Rural	0.034	0.026				
Elementary & Secondary Education:						
Expenditures per Pupil			10,564	8,315	10,773	8,281
Pupil-to-Teacher Ratio			16.9	16.4	16.8	16.7
Math NAEP Score			277.6	275.0	277.3	276.0
Reading NAEP Score			262.9	260.4	262.4	261.5
Observations	5,570	16,760	15,540	6,790	13,960	8,370

[1]Institutions are defined as high access if there are ≥ 4 elite schools within a 100 mile radius or $\geq 15\%$ of seats within 100 miles are at elite schools. States are defined as high access if they have $> 1,000$ elite seats or $> 5\%$ of the state's freshman attend an elite school anywhere in the country. Regions are defined using the classification shown in Figure 2.3 and are high access if they have at least 10,000 elite seats.

[2]Institution types are defined by U.S. News and World Report

2.5 Results

Table 2.2 shows the marginal effects results of estimating the specification described by (2.12) using both number of nearby elite colleges and/or the percent of freshman enrollemnt that is at elite schools within a 100 mile radius of each non-elite school to identify high access areas. The first two columns show the results of using continuous measures for the level of access. These results indicate that 1 additional elite school within a 100 mile radius decreases the probability of choosing a STEM major for lower ability students at non-elite colleges by 0.4 percentage points. Given that the average probability of choosing a STEM major in this sample is approximately 14.7%, this effect is equivalent to a 3% decrease in STEM prevalence. Similarly, increasing the percent of available freshman seats within a 100 mile radius that are at elite schools decreases the probability of choosing STEM by 0.2 percentage points or 2%. However, the best measure

Table 2.2: Effect of Elite Presence Within 100 Mile Radius on $\mathbb{P}$(STEM) at Non-Elite Colleges

	# Schools	% Seats	$\geq$ 2 Schools or $\geq$ 10% Seats	$\geq$ 3 Schools or $\geq$ 10% Seats	$\geq$ 3 Schools or $\geq$ 15% Seats	$\geq$ 4 Schools or $\geq$ 15% Seats
	(1)	(2)	(3)	(4)	(5)	(6)
High Access	-0.0040**	-0.0024*	-0.0230*	-0.0285**	-0.0365***	-0.0329**
	(0.0018)	(0.0014)	(0.0122)	(0.0128)	(0.0115)	(0.0127)
High Access %			0.42	0.35	0.31	0.26
Observations	22,330	22,330	22,330	22,330	22,330	22,330

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

All specifications include controls for demographic variables, institution type, urbanization, cohort fixed effects, and state fixed effects. Standard errors in parentheses are clustered by institution.

of elite college access should capture both the number of schools and their relative sizes, so the remaining columns use indicator variables for various combined thresholds of the two access measures. Column (3) is the most liberal definition of high access, requiring only 2 nearby elite colleges or 10% of local freshman enrollment to be at elite schools in order for a non-elite school to be classified as high access. The thresholds become stricter in each column to the right in Table 2.2, with the strictest definition in column

(6) being the preferred specification. The results of these threshold specifications indicate that the effect of having a higher elite college presence within a 100 mile radius is a 2.3-3.7 percentage point decrease in the probability of choosing a STEM major at non-elite colleges, which is equivalent to a 16-25% decline. This represents a considerable positive shift of the major cutoff point at non-elite colleges in high access areas and is consistent with the signaling model prediction.

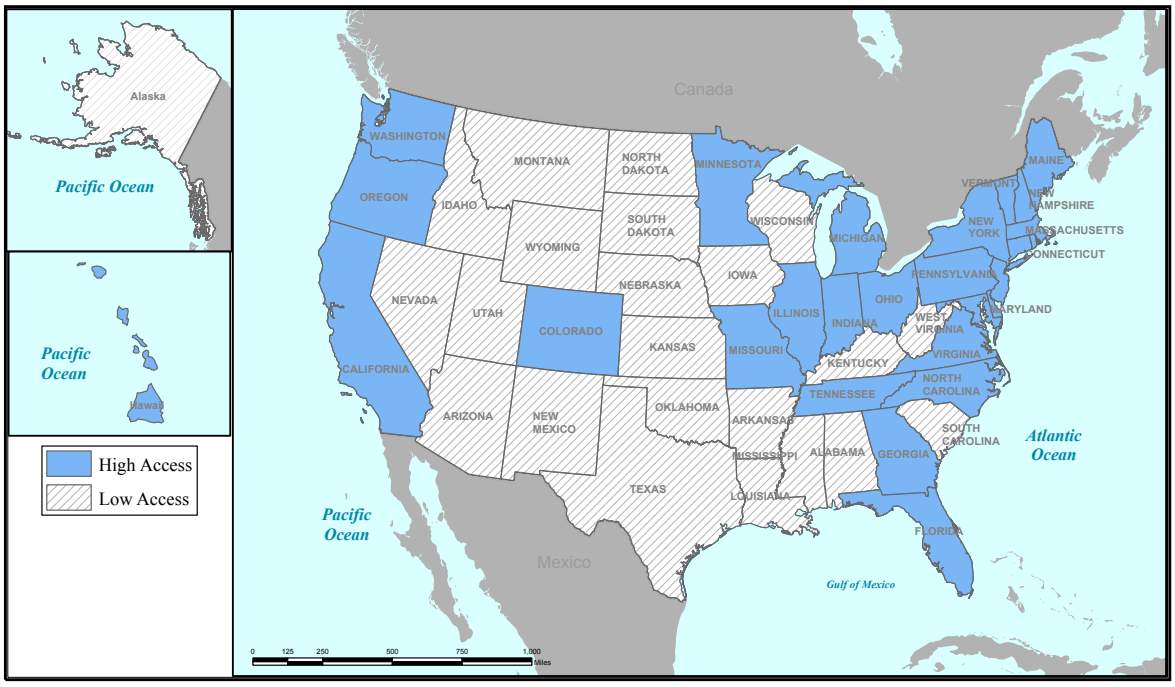
To check that these results are not specific to this 100-mile radius strategy of defining high access areas, I also estimate the model using two alternative identification strategies. In the first approach, I classify each U.S. state as a high or low access area so that all students attending a non-elite college in a high access state are assumed to have faced a lower probability of being constrained from entering an elite college than those students attending non-elite schools in low access states. Of course, the boundaries between states are largely arbitrary and, especially in the East where states are smaller, the barriers to moving across state lines are low. Thus, classifying students in Delaware as having less access to the elite colleges of New York City than the students in Albany, NY is somewhat misleading given that Delaware is actually located at a closer distance to New York City. To minimize this type of misclassification, in the second alternative identification strategy, I define groups of states (regions) as high or low access.

The IPEDS data provide each institution's freshman enrollment by state of high school residence for the years 1988, 1996, and 2004 (the years that each B&B cohort were likely freshmen).[18] In the state-level identification strategy, I use this data to determine which U.S. states have a high concentration of open seats at elite colleges and which states send a large fraction of their students to elite schools located anywhere within the U.S. (averaged over the years 1988, 1996, 2004). I use both of these measures in order to

[18]The 1993 cohort would most likely have been freshmen in the Fall of 1989, but high school residence data is not available in IPEDS for that year.

mitigate the problem of arbitrary state boundaries mentioned above. For example, Table 2.3 shows that while Connecticut only has 2 very small elite schools, 13.4% of freshmen from Connecticut attend elite schools, which is one of the highest percentages in the country. These students may be attending elite colleges in neighboring states such as Massachusetts. Using only the number of elite schools or elite seats within each state to measure accessibility would lead to the misclassification of states like Connecticut into the low access group. I define high access states to be those that satisfy either of the following criteria: (1) more than 1,000 seats at local elite schools, or (2) send more than 5% of their students to elite colleges anywhere in the U.S. (shown in Figure 2.2).[19]

Figure 2.2: Access to Elite Colleges by State



[19] As a robustness check, I drop all students from the sample who attend an institution within 50 miles of any state border. This removes more than half of the sample, but the resulting estimates are actually larger and remain statistically significant.

Table 2.3: Prevalence of Elite Colleges by State

State	# of Elite Colleges	# of Freshman Seats at Elite Colleges in State	% of Freshmen from State at Any Elite College	High Access
CA	9	19,520	16.0	1
NY	11	17,488	12.5	1
MA	11	14,282	17.7	1
MI	1	6,871	8.0	1
VA	3	6,089	12.7	1
PA	5	5,808	6.3	1
GA	2	4,959	8.5	1
IL	3	4,008	5.5	1
CO	2	3,272	5.0	1
TN	2	2,202	4.6	1
MO	1	2,185	3.3	1
NC	2	2,138	2.9	1
IN	1	2,122	2.0	1
OH	3	2,082	4.3	1
CT	2	2,056	13.4	1
MD	1	1,875	10.6	1
MN	3	1,819	5.7	1
DC	1	1,748	22.1	1
ME	3	1,488	12.9	1
RI	1	1,429	11.1	1
NJ	1	1,172	11.5	1
NH	1	1,081	10.2	1
TX	1	726	3.4	0
VT	1	580	10.5	1
IA	1	434	2.6	0
WA	1	417	5.9	1
OR	1	338	5.6	1
AR	1	320	2.8	0
FL	1	232	6.3	1
HI	0	0	7.7	1
DE	0	0	5.6	1
AZ	0	0	4.7	0
AK	0	0	4.6	0
NM	0	0	3.8	0
NV	0	0	3.4	0
WY	0	0	2.9	0
MT	0	0	2.9	0
AL	0	0	2.9	0
KS	0	0	2.7	0
SC	0	0	2.6	0
WI	0	0	2.4	0
MS	0	0	2.4	0
ID	0	0	2.4	0
KY	0	0	2.2	0
OK	0	0	2.0	0
UT	0	0	1.9	0
NE	0	0	1.8	0
WV	0	0	1.8	0
SD	0	0	1.7	0
ND	0	0	1.6	0
LA	0	0	1.6	0

The regression to be estimated for the state-level identification approach is:

$$\mathbb{P}\left(STEM_{icst}=1\right)=\Phi\left(\alpha+\beta HighAccess_{s}+\delta'X_{i}+\theta'Z_{c}+\gamma'E_{st}+\eta_{t}\right). \tag{2.13}$$

This differs from the main specification (2.12) in the variable of interest, $HighAccess_s$, which now indicates whether the college attended is in a state with high access to elite colleges. Clearly, the state fixed effects can no longer be included and are replaced by the vector E_{st}, which captures state-level characteristics that might influence major choice for each cohort: expenditure per pupil in public elementary and secondary schools, pupil-to-teacher ratio in public elementary and secondary schools, and National Assessment of Educational Progress (NAEP) 8th grade math and reading scores.[20] Ideally, I would also control for state economic indicators such as median household income, however the number of time-varying controls is limited by the fact that I only have 3 years of data.

Table 2.4: Effect of State-Level Access to Elite Colleges on $\mathbb{P}$(STEM) at Non-Elite Colleges

	High Access = $>1,000$ seats or $>5\%$ students (1)	High Access = $>1,000$ seats only (2)	High Access = $>5\%$ students only (3)	High Access = $>4,000$ seats or $>10\%$ students (4)
High Access	-0.0258***	-0.0172**	-0.0204**	-0.0145*
	(0.0083)	(0.0071)	(0.0107)	(0.0087)
High Access %	69.3	59.8	51.4	39.9
Observations	22,330	22,330	22,330	22,330

* $p<0.10$, ** $p<0.05$, *** $p<0.01$

All columns include controls for age, gender, parents' education, race/ethnicity, institution type, state-level expenditures-per-pupil and pupil-to-teacher ratio, state NAEP scores for math and reading, and cohort fixed effects. Standard errors in parentheses are clustered by state.

Table 2.4 displays the marginal effects results of estimating equation (2.13) using a Maximum Likelihood estimator. Column (1) shows the results using the preferred def-

[20]Education expenditure and pupil-to-teacher ratio data are from the NCES Common Core of Data surveys from the 1988-1989, 1995-1996, and 2003-2004 school years. NAEP scores from 2003 are used for all cohorts as this is the first year that data for all states is available.

inition of high access states (more than 1,000 elite seats or more than 5% of students attending elite schools) while columns (2)-(4) show the results using increasingly strict cutoff points for inclusion in the high access group. These estimates are consistent across definitions at approximately a 1.5-2.6 percentage point (or 10-18%) decrease in the probability of choosing a STEM major for low ability students at non-elite colleges in high access states. These results are also consistent with the main findings in Table 2.2, although somewhat smaller, and support the predictions of the asymmetric information model.

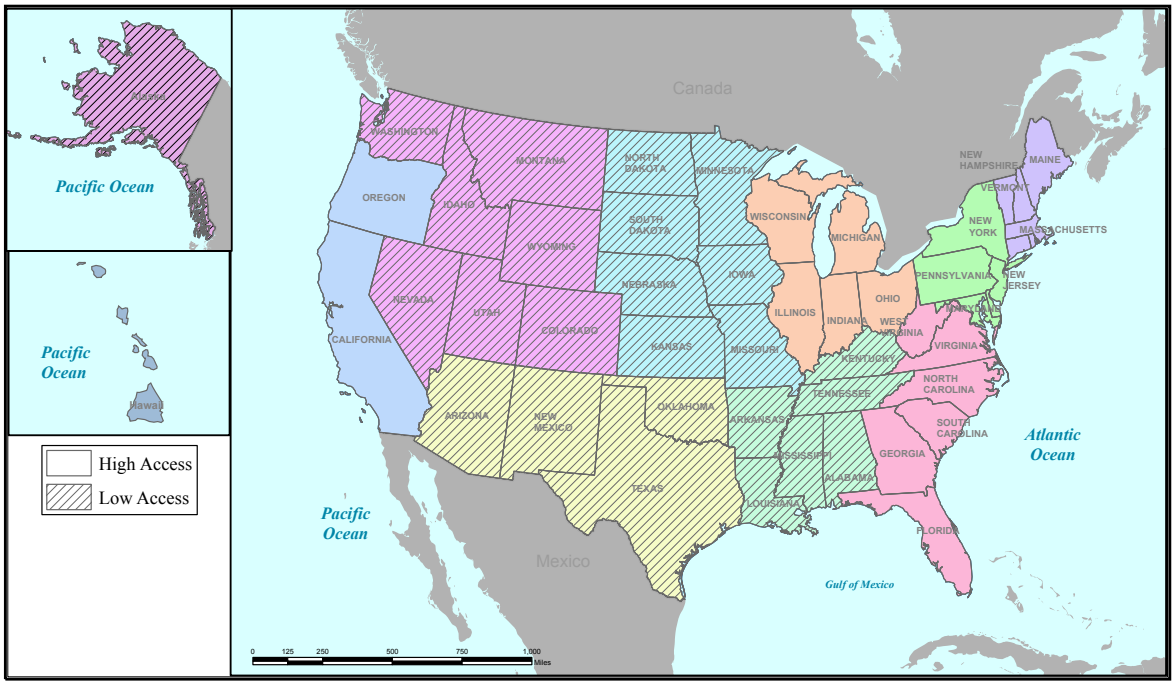
To further minimize the possibility of misclassification, I next aggregate up to the region level in the second alternative identification strategy. Using the IPEDS data, I divide the country into 9 regions, keeping states with high inter-mobility of students grouped together. I define high access regions to be those at least 10,000 seats at local elite colleges (shown in Figure 2.3).[21]

Table 2.5: Prevalence of Elite Colleges by Region (9 Region Classification)

Region	States	# of Freshman Seats at Elite Colleges in Region	# of Elite Colleges in Region	High Access
1	DE, DC, MD, NJ, NY, PA	28,091	19	1
2	CT, ME, MA, NH, RI, VT	20,916	19	1
3	CA, HI, OR	19,858	10	1
4	IL, IN, MI, OH, WI	15,083	8	1
5	FL, GA, NC, SC, VA, WV	13,418	8	1
6	IA, KS, MN, MO, NE, ND, SD	4,438	5	0
7	AK, CO, ID, MT, NV, UT, WA, WY	3,689	3	0
8	AL, AR, KY, LA, MS, PR, TN	2,522	3	0
9	AZ, NM, OK, TX	726	1	0

[21]This cutoff forms a natural break point in the data (shown in Table 2.5) and the results shown below are not sensitive to decreasing this requirement (and thereby including the Plains region in the high access group) or to increasing the requirement (and thereby excluding the Southeast region from the high access group).

Figure 2.3: Access to Elite Colleges by Region



The regression model corresponding to this region-level strategy is:

$$\mathbb{P}\left(STEM_{icrt} = 1\right) = \Phi\left(\alpha + \beta HighAccess_r + \delta' X_i + \theta' Z_c + \eta_t\right), \tag{2.14}$$

where the variable $STEM_{icrt}$ is now an indicator that equals 1 if individual i at college c located in region r from cohort t chooses a STEM major and the variable $HighAccess_r$ indicates whether the college attended is in a high access region.

Table 2.6 shows the results of estimating the region-level model given by (2.14). Column (1) shows the marginal effects estimates using the region definitions shown in Figure 2.3. The magnitude of this estimate is consistent with the institution and state-level findings at approximately 2.5 percentage points. However, because of the small number of clusters in this model (9 regions), conducting proper inference is problematic. To address this issue, I apply the two-step approach of Donald and Lang (2007), which

Table 2.6: Effect of Region-Level Access to Elite Colleges on ℙ(STEM) at Non-Elite Colleges

	9 Regions		17 Regions	
	Probit	D&L	Probit	D&L
	(1)	(2)	(3)	(4)
High Access	-0.0249***	-0.0307**	-0.0249***	-0.0286**
	(0.0078)	(0.0112)	(0.0079)	(0.0098)
High Access %	63.9	63.9	54.1	54.1
Observations	22,330	9	22,330	17

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

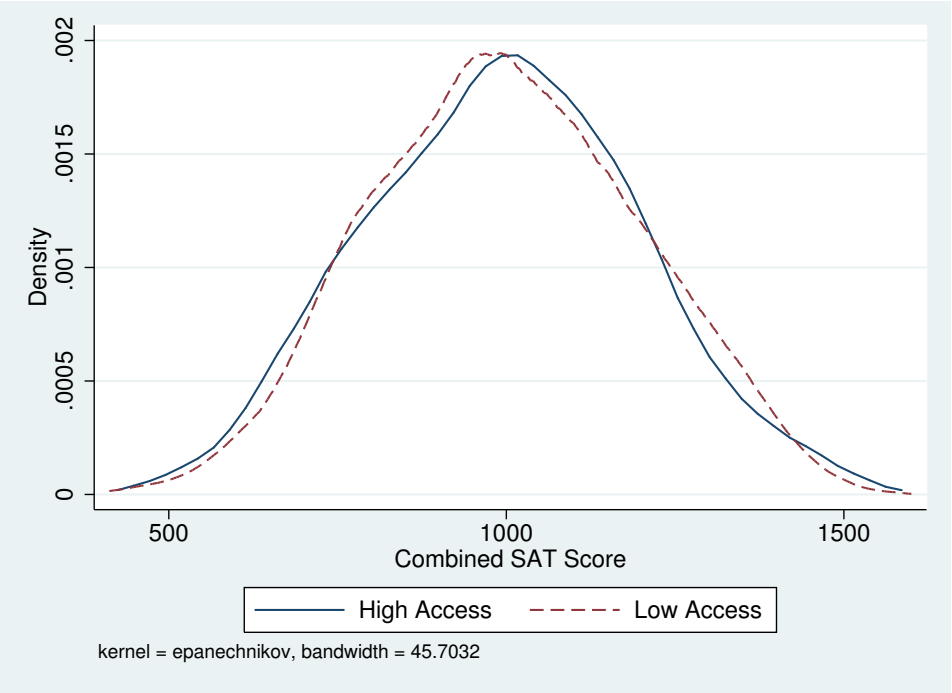
All columns include controls for demographic variables, institution type, and cohort fixed effects. Standard errors in parentheses are clustered by region for columns (1) and (3). Estimates and standard errors in columns (2) and (4) are obtained from the two-stage D&L method.

can provide more conservative standard errors. Under the two-stage Donald and Lang (D&L) approach using a probit model, it is unclear how one might obtain marginal effects. For this reason I estimate the first stage using a linear probability model. Column (2) shows the results of this estimation method, which are actually somewhat larger (a 3.1 percentage point increase) and statistically significant at the 5% level.[22] To ensure that these results are not specific to this particular region classification, I also create a 17 region classification of the states that is shown in Appendix Figure B.1. The results of estimating the model using this region definition (shown in columns (3) and (4) of Table 2.6) are very similar to all previous specifications. The effect of being in a region with high access to elite schools is a 2.5-3.1 percentage point, or 17-21%, decrease in the probability of choosing a STEM major.

The two main concerns with the above specifications stem from potential failures of Assumptions 1 and 2 (as stated in Section 2.2.1). Specifically, if the distribution of ability is shifted to the left in high access areas or if the cost of choosing a STEM major

[22]The point estimates differ between the probit model in column (1) and the D&L method in column (2) for two reasons: first, the use of the linear probability model in the first stage, and second, the unequal sizes of the region clusters.

Figure 2.4: Distribution of SAT Scores by Parents' State of Residence



is greater in high access areas, then the decrease in the probability of choosing a STEM major could be explained by the full information model. To address the first concern, I use data from the 1993 cohort, which includes parents' state of residence, to test whether the SAT score distribution of students from high access states is lower than the SAT score distribution of students from low access states. Figure 2.5 shows the density of combined SAT scores for these two groups. The two density functions certainly do not appear to be shifted in either direction and a Kolmogorov-Smirnov test for equality of the two distributions fails to reject the null hypothesis of identical distributions with a p-value of 0.838.

The second concern is more difficult to address empirically. I provide 4 pieces of evidence that shed light on the possibility of differential STEM costs across access-types. The first is a test of differential college quality across high and low access areas. If the quality of non-elite schools in high access areas is higher, then it might be possible that

it is more costly to enter into a STEM major at those schools. This would cause a shift in the major cutoff point even in the full information framework. Table 2.7 shows that the distribution of non-elite college quality is similar across low and high access areas, regardless of whether those areas are defined by region, state, or the 100 mile radius around each institution. A chi-squared test fails to reject the null hypothesis that the distributions are the same for both the low and high access areas.

Table 2.7: Quality Distribution of Non-Elite Schools

	Institutions		States		Regions	
Tier	High Access	Low Access	High Access	Low Access	High Access	Low Access
0: Missing USN data	8.2%	6.7%	7.3%	6.7%	6.9%	7.4%
1: Bottom 25%	12.8%	10.4%	10.5%	12.3%	11.8%	9.6%
2: Top 51-75%	29.9%	37.6%	33.2%	41.0%	34.1%	38.2%
3: Top 26-50%	26.7%	23.7%	25.9%	21.3%	25.2%	23.2%
4: Top 11-25%	16.7%	17.7%	18.2%	15.7%	17.1%	18.1%
5: Top 6-10%	5.7%	3.9%	4.9%	3.0%	4.9%	3.4%
	$\chi^2_{(4)} = 7.57$, $p = 0.181$		$\chi^2_{(5)} = 8.69$, $p = 0.122$		$\chi^2_{(5)} = 3.85$, $p = 0.572$	

Rows/tiers defined by the 75th percentile freshman SAT score given by the USN data.

Institutions are defined as high access if there are ≥ 4 elite schools within a 100 mile radius or $\geq 15\%$ of seats within 100 miles are at elite schools. States are defined as high access if they have $> 1,000$ elite seats or $> 5\%$ of the state's freshman attend an elite school anywhere in the country. Regions are defined using the classification shown in Figure 2.3 and are high access if they have at least 10,000 elite seats.

A second strategy is to look for systematic mobility among students at non-elite colleges. Given that the quality of non-elite institutions is similar across high and low access areas, if the effort cost of choosing STEM is greater in higher access areas, then the full information model would predict a flow of STEM students from high access areas into low access areas. Using the B&B 1993 cohort data, I find that the fraction of non-elite college students who move from a high access state to a low access state is quite low at 8.5% (movement in the opposite directions is slightly more prevalent at 11.5%). Table 2.8 shows that these students are no more likely to major in a STEM field than the students who stay behind in the high access states. They are also no more likely to major

in STEM than the non-elite college students that they join in the low access states. The same is true of the students who move in the other direction from a low access state to a high access state. This finding suggests that the cost of a STEM major is equal across high and low access areas.

Table 2.8: Probability of Choosing a STEM Major At Non-Elite Colleges

	State of College Attended			
Parents' State of Residence	High Access	Low Access	Diff. High-Low	p-value
High Access	18.6%	19.5%	-0.86%	0.724
Low Access	19.6%	19.5%	0.02%	0.995
Diff. High-Low	-0.97%	-0.09%		
p-value	0.739	0.972		

States are defined as high access if they have $> 1,000$ elite seats or $> 5\%$ of the state's freshman attend an elite school anywhere in the country.

Thirdly, I estimate the specification in column (1) of Table 2.4 excluding all students whose high school and college are not in the same state. The resulting point estimate is actually much larger at 5.3 percentage points and remains significant at the 1% level. This indicates that the results are not being driven by student mobility across state lines.

Finally, and most importantly, the inclusion of state fixed effects in the main specification (shown in Table 2.2) means that the ability distribution and/or the cost of a STEM major would have to vary systematically within each state in order for pure human capital accumulation to have caused the observed differences in major choice. Given all of these findings, it seems unlikely that any observed decrease in the probability of choosing STEM at non-elite schools is due to differences in human capital accumulation rather than a signaling effect.

2.5.1 Robustness Checks

The shift in the major cutoff estimated above is robust to a number of alternative specifications. Tables 2.9 - 2.12 display the results of several of these robustness checks. In each table, the "Inst" columns show estimates of equation (2.12), including state fixed effects, using the preferred definition of high access non-elite colleges: at least 4 elite schools within a 100 mile radius or at least 15% of local freshman enrollment at elite colleges (as in column (6) of Table 2.2). The "State" columns show estimates of equation (2.13) using the preferred definition of high access states: more than 1,000 elite seats or more than 5% of the state's freshmen attend an elite school (as in column (1) of Table 2.4). The "Region" columns show estimates of equation (2.14) using the 9 region classification shown in Figure 2.3 and employing the two-stage D&L method of estimation (as in column (2) of Table 2.6).

One variable that has thus far been excluded from estimation is intelligence/academic ability. The closest measure of academic ability in the B&B data is SAT/ACT score; unfortunately the sample includes more than 5,000 individuals who did not report an SAT or ACT score. I offer three methods for dealing with this missing data. Columns (1)-(3) of Table 2.9 show the results from assigning an SAT/ACT score of zero to those students who are missing a score and including both the score and an indicator for individuals with missing scores in the vector of covariates. Alternatively, columns (4)-(6) use a linear regression of observed SAT/ACT scores on age bins, gender, race/ethnicity, and parents' education to impute scores for those individuals with missing data. Finally, since it is unclear whether an individual's combined SAT/ACT score is the best measure of ability in this model, given that the focus is on the choice of a quantitative field of study, the estimates in columns (7)-(9) show the results of using only the math portion

of the imputed score to measure ability.[23] These alternative specifications show that the main results are robust to including measures of individual academic ability in the model. These estimates are approximately equivalent to an 11-22% decrease in the probability of choosing a STEM field.

Table 2.9: Robustness Checks - Controlling for Ability

	Include SAT Control			Impute Missing SAT Scores			Impute Missing SAT I Math		
	Inst	State	Region	Inst	State	Region	Inst	State	Region
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
High Access	-0.0316**	-0.0230***	-0.0278**	-0.0297**	-0.0192**	-0.0284**	-0.0228*	-0.0166**	-0.0263*
	(0.0126)	(0.0083)	(0.0117)	(0.0125)	(0.0083)	(0.0115)	(0.0114)	(0.0082)	(0.0116)
SAT Combined Score	0.0272***	0.0274***	0.0259***	0.0250***	0.0253***	0.0249***			
	(0.0032)	(0.0040)	(0.0030)	(0.0032)	(0.0037)	(0.0032)			
SAT Missing Indicator	0.351***	0.348***	0.267***	0.0083	0.0047	0.0048	0.0251***	0.0225**	0.0263***
	(0.0451)	(0.0601)	(0.0289)	(0.0090)	(0.0094)	(0.0093)	(0.0090)	(0.0102)	(0.0092)
SAT Math Score							0.0416***	0.0427***	0.0410***
							(0.0060)	(0.0068)	(0.0058)
State Fixed Effects	X			X			X		
State Education Variables		X			X			X	
Observations	22,330	22,330	9	21,070	21,070	9	20,300	20,300	9

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Estimates and standard errors in columns (3), (6), and (9) are obtained from the two-stage D&L method. Standard errors in parentheses are clustered by state for columns (2), (5), and (8) and by institution for columns (1), (4), and (7). All specifications include controls for demographic variables, institution type, and cohort fixed effects.

Institutions are defined as high access if there are ≥ 4 elite schools within a 100 mile radius or $\geq 15\%$ of seats within 100 miles are at elite schools. States are defined as high access if they have $> 1,000$ elite seats or $> 5\%$ of the state's freshman attend an elite school anywhere in the country. Regions are defined using the classification shown in Figure 2.3 and are high access if they have at least 10,000 elite seats.

Table 2.10 shows that the results are not sensitive to the cutoff point, a^{Q_H}, that determines which students are unlikely to be directly constrained ($a_i < a^{Q_H}$) and therefore included in the sample. In columns (1)-(3), I find the SAT score, S^*, for which 95% of students at elite colleges score above S^*. This value is 930 for the 1993 cohort, 980 for the 2000 cohort, and 1,030 for the 2008 cohort. All students at non-elite colleges who score above S^* are then excluded from the sample. This lower cutoff point ensures that the results are not driven by the inclusion of directly constrained students choosing STEM majors at non-elite colleges in low access areas. If, however, it is possible to buy a spot at an elite college regardless of having lower academic qualifications, then students

[23]Note that I also recalculate the high ability cutoff point ($a_i < a^{Q_H}$) using the imputed SAT/ACT combined scores in columns (4)-(6) and using the imputed SAT/ACT math scores in columns (7)-(9).

from very high income families should also be excluded from the sample. I use the entire B&B sample to find the 90th percentile family income, which is $158,000 in 2008 dollars. Columns (4)-(6) drop all students with a family income above this amount from the sample. Finally, Columns (7)-(9), combine these requirements and drop all students who either have an SAT score above the 95% S^* or who have a family income above $158,000. The results in each of these specifications vary between an 11% and 22% decrease in the probability of choosing a STEM field, demonstrating that the major choices of directly constrained students are likely not driving the estimated shift in the major cutoff point.

Table 2.10: Robustness Checks - Cutoff for Directly Constrained Students

	5th Percentile Elite SAT Score			Drop Incomes Above $158,000			5th Pctl Elite SAT & Drop High Incomes		
	Inst	State	Region	Inst	State	Region	Inst	State	Region
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
High Access	-0.0248*	-0.0157*	-0.0303*	-0.0309**	-0.0264***	-0.0320**	-0.0253*	-0.0169*	-0.0304*
	(0.0122)	(0.0097)	(0.0138)	(0.0127)	(0.0085)	(0.0108)	(0.0128)	(0.0103)	(0.0133)
State Fixed Effects	X			X			X		
State Education Variables		X			X			X	
Observations	15,040	15,040	9	20,860	20,860	9	14,300	14,300	9

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Estimates and standard errors in columns (3), (6), and (9) are obtained from the two-stage D&L method. Standard errors in parentheses are clustered by state for columns (2), (5), and (8) and by institution for columns (1), (4), and (7). All specifications include controls for demographic variables, institution type, and cohort fixed effects.

Institutions are defined as high access if there are ≥ 4 elite schools within a 100 mile radius or $\geq 15\%$ of seats within 100 miles are at elite schools. States are defined as high access if they have $> 1,000$ elite seats or $> 5\%$ of the state's freshman attend an elite school anywhere in the country. Regions are defined using the classification shown in Figure 2.3 and are high access if they have at least 10,000 elite seats.

Table 2.11 shows that the results are not sensitive to the definition of elite schools. Columns (1) - (3) include all students at schools that were not matched with the USN rankings data (and were therefore missing college quality data) as non-elite institutions. Columns (4) - (6) show the results of decreasing the requirement that determines which schools are elite to include the top 10% of schools as ranked by their 75th percentile SAT score. Finally, columns (7) - (9) use data from IPEDS on the 25th percentile SAT math score of applicants to each school and average over the years 2005-2008 to rank each institution. Using this alternative ranking, the top 5% of schools are classified as

elite and all others as non-elite. The results in each of these specifications indicate a 15-25% decrease in the probability of choosing a STEM field, demonstrating that the estimated shift in the major cutoff point is robust to varying definitions of elite and non-elite schools.

Table 2.11: Robustness Checks - Elite Definitions

	Add Missing			Elite = Top 10%			IPEDS Tiers		
	Inst	State	Region	Inst	State	Region	Inst	State	Region
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
High Access	-0.0259*	-0.0220***	-0.0280**	-0.0204	-0.0215**	-0.0310**	-0.0307**	-0.0338***	-0.0370**
	(0.0131)	(0.0084)	(0.0115)	(0.0125)	(0.0090)	(0.0118)	(0.0151)	(0.0093)	(0.0145)
State Fixed Effects	X			X			X		
State Education Variables		X			X			X	
Institution Type Variables	X	X	X	X	X	X			
Observations	23,300	23,300	9	19,090	19,090	9	21,230	21,230	9

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Estimates and standard errors in columns (2), (4), and (6) are obtained from the two-stage D&L method. Standard errors in parentheses are clustered by state for columns (1), (3), and (5). All specifications include controls for demographic variables and cohort fixed effects.

Institutions are defined as high access if there are ≥ 4 elite schools within a 100 mile radius or $\geq 15\%$ of seats within 100 miles are at elite schools. States are defined as high access if they have $> 1,000$ elite seats or $> 5\%$ of the state's freshman attend an elite school anywhere in the country. Regions are defined using the classification shown in Figure 2.3 and are high access if they have at least 10,000 elite seats.

Finally, because engineering departments often require a separate application in the admissions process and may have different admissions criteria, it is unclear that engineering students face the same decision problem as other college applicants. Table 2.12 shows that the main results are robust to dropping all engineering majors from the analysis.

The results in Tables 2.2 - 2.12 provide empirical evidence of a positive shift of the major cutoff point at non-elite colleges in high access areas. These results are consistent across 3 separate identification strategies and are robust to many possible variations of each specification. These estimates range between a 1.5 and 3.7 percentage point (or 10-25%) decrease in the probability of choosing a STEM major for students at non-elite colleges in areas with a large elite college presence. This shift is consistent with the predictions of a signaling model of college major choice.

Table 2.12: Robustness Checks - Drop Engineering Majors

	Institution	State	Region
	(1)	(2)	(3)
High Access	-0.0156	-0.0169**	-0.0296**
	(0.0106)	(0.0077)	(0.0104)
State Fixed Effects	X		
State Education Variables		X	
Observations	21,310	21,310	9

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Estimates and standard errors in column (3) are obtained from the two-stage D&L method. Standard errors in parentheses are clustered by institution for column (1) and by state for column (2). All specifications include controls for demographic variables, institution type, and cohort fixed effects.

Institutions are defined as high access if there are ≥ 4 elite schools within a 100 mile radius or $\geq 15\%$ of seats within 100 miles are at elite schools. States are defined as high access if they have $> 1,000$ elite seats or $> 5\%$ of the state's freshman attend an elite school anywhere in the country. Regions are defined using the classification shown in Figure 2.3 and are high access if they have at least 10,000 elite seats.

2.6 Conclusion

In this paper, I have developed an asymmetric information framework for college education in which college major field of study is used by individuals to signal productivity to employers. I show that this signaling model predicts that increased access to elite colleges leads to a decrease in the proportion of STEM majors, specifically among lower ability students at non-elite colleges. This is distinct from the prediction of a full information model where increased access to elite colleges only affects those high ability individuals who are directly constrained from attending elite schools. Using geographic variation in the concentration of elite colleges in the U.S., I provide empirical evidence that supports the signaling model prediction. I find that in areas with substantial access to elite schools the probability of choosing a STEM major for students at non-elite schools who are not at the top of the ability distribution is approximately 10-25% lower. This result indicates

that there is scope for signaling behavior within the context of postsecondary education and that qualitative measures of education, such as the choice of major, can be used as labor market productivity signals.

Chapter 3

Saved By the Morning Bell: School Start Time and Teen Car Accidents

3.1 Introduction

Recent medical research has found that adolescents experience changes to the biological clock near the onset of puberty that induce "night owl" sleep patterns. This effect, combined with early high school start times has led to widespread, chronic sleep deprivation among teens. Surveys show that less than one in ten high school students get the recommended amount of sleep on school nights (NSF, 2006). This type of chronic sleep loss can have significant health and economic consequences such as increased feelings of depression, impaired memory and focus, and increased risk of adolescent obesity (Ped, 2014). In susceptible young people, this pattern may lead to academic and behavioral problems as well as increased risk for accidents and injuries.

One oft-proposed solution to this problem of sleep deprivation is to allow teenagers additional sleep time in the mornings by delaying high school start times. This paper explores an unexpected health consequence of changes to school start times: the impact on teen car accidents. Changes to school start times may alter teen car accident risk both through a direct effect on sleep deprivation and indirectly through changes to the

driving environment, making the direction of the overall effect theoretically ambiguous. Using data from the state of Kansas, I exploit within-school variation in start times over a 9 year period to identify the effect of high school start time on the average number of teen car accidents. Implementing a Fixed-Effects Poisson Quasi-Maximum Likelihood approach, I find that a 15 minute delay in high school start times leads to a 21% increase in teen car accidents during the morning commuting hours. At the average, this effect is equivalent to approximately 124 additional morning teen car accidents per year across all of rural Kansas. This suggests that any effect stemming from avoided sleep deprivation is being offset by the effect of shifting teen driving into the high volume "rush hour" of the morning. However, by focusing on late-night accidents when there is little to no possibility of traffic congestion, I am able to disentangle the two mechanisms and find evidence of a persistent sleep effect. The avoided sleep deprivation caused by a 15 minute delay in school start time leads to a 26% decrease in late-night teen accidents. At the average, this effect is equivalent to approximately 68 fewer accidents of this type per year for teen drivers across all of rural Kansas.

The remainder of this paper is organized as follows. Section 3.2 provides background information on teenage sleep patterns and their potential effects. This section also summarizes the existing literature showing significant effects of school start time on academic outcomes and survey findings showing a correlation between school start times and teen car accident rates. Section 3.3 describes the data and the empirical context of rural Kansas. Section 3.4 lays out the empirical strategy and the advantages of using the Fixed-Effects Poisson Quasi-Maximum Likelihood estimator. Section 3.5 presents the results and Section 3.6 offers some concluding remarks.

3.2 Teens & Sleep

Until recently, it has been the general understanding that delayed bed times of teenagers were the result of peer culture and other psychosocial factors. However, recent medical research has found that there are biological explanations for the later sleep schedules of teenagers. After the onset of puberty, adolescents experience delayed secretion of nocturnal melatonin, a lengthening of the period of the circadian clock, and a slower build-up of homeostatic sleep pressure (Carskadon et al., 1998; Jenni et al., 2005). The combined effect of these changes is that teens experience a sleep cycle delay of approximately 2 hours relative to their pre-adolescent baseline (Ped, 2014). In practice, this means that the average teenager has trouble falling asleep before 11pm and imposition of early school start times may require unrealistic or even unattainable bedtimes to provide adequate time for sleeping (Wahlstrom, 2002). Research has shown that early school start times are associated with significant sleep deprivation to the point where students can fall into REM sleep in only 3.4 minutes – a level that is consistent with the sleep patterns of patients with narcolepsy (Carskadon et al., 1998).

The public reaction to these findings has been widespread in the US. At the national level, bill H.R. 1306: ZZZ's to A's Act has been introduced in the House of Representatives and proposes "to conduct a study to determine the relationship between school start times and adolescent health, well-being, and performance." At the local level, school districts in at least 43 states have made policy changes aimed at improving adolescent sleep levels by delaying high school and/or middle school start times.[1] Most recently, Fairfax County in Virginia implemented a new policy starting in 2015-2016 requiring all high schools to start between 8-8:10am. Unfortunately, such policies can be very costly. One of the main reasons for early high school start times is to maintain a tiered busing

[1] http://www.startschoollater.net/success-stories.html

system. In many districts, school buses run on a loop schedule wherein high school students are dropped off first, then middle school students, and pick-up of elementary school students is last. In the case of Fairfax County, the new policy required the purchase of 27 new buses and cost a total of $4.9 million.[2] Furthermore, there are concerns that moving school start times might add to rush hour traffic congestion or that lower-income students might be adversely affected by a decreased ability to work after-school jobs or to care for younger siblings.

Despite these costs, there may be room for substantial gains from decreasing teenage sleep deprivation through improved academic, behavioral, and health outcomes. Several papers have attempted to establish a link between later school start times and improved cognitive function through test scores and/or grades among adolescent students (Carrell et al., 2011; Hinrichs, 2011; Edwards, 2012). Carrell et al. (2011) find that freshmen at the US Air Force Academy benefit greatly from a delay in school start time. They show that a 50 minute delay has the same effect on academic achievement as increasing teacher quality by 1 standard deviation and that the effect persists throughout the day (rather than being driven by first period performance alone). Edwards (2012) finds a similar positive effect of school start time on standardized test scores in the middle-school age group. Hinrichs (2011) conducts two similar analyses and finds no effect of high school start time on individual-level SAT/ACT scores or on school-level standardized test scores. However, this may be due to the dependent variable being a noisier measure of in-school learning in the first case and of being limited to school-level data in the second case.

A second line of survey work has explored the possible correlation between later school start times and decreased teen car accidents (Danner and Phillips, 2008; Vorona et al., 2011; Martiniuk et al., 2013; Wahlstrom et al., 2014). This is a potentially important health outcome as medical studies show that sleep deprivation can produce psychomotor

[2]http://www.fcps.edu/news/starttimes.shtml

impairments equivalent to those induced by alcohol consumption at or above the legal limit. Furthermore, it is known that young drivers are at a much higher risk for drowsy driving and sleep-related crashes (Durmer and Dinges, 2005). Danner and Phillips (2008) conducted a questionnaire investigating the effects of a 1-hour delay in school start times in a single large school district in Kentucky. They find that average hours of nightly sleep increased after the policy change and that average car accident rates for teen drivers in the study dropped 16.5% compared with the 2 years prior, whereas teen accident rates for the rest of the state increased 7.8% over the same time period. In another survey of over 9,000 students in eight public high schools, Wahlstrom et al. (2014) find a negative correlation between school start times and the number of car accidents involving surveyed students. The average number of car accidents reported by teens in the survey was reduced by 70% when the surveyed schools shifted start times from 7:35 AM to 8:55 AM.[3]

While the average changes reported in these papers provide only descriptive evidence of a school start time effect on teen car accidents, they are also quite large. The avoided financial and health/mortality costs corresponding to causal effects of such magnitudes could easily make start time policy changes worthwhile. This paper contributes to this line of exisiting descriptive research by providing regression-based analysis to control for other factors that might influence teen accident rates and by distinguishing between the separate (and potentially opposing) effects of avoided sleep deprivation and increased traffic congestion.

[3]There is related evidence showing that sleep-deprivation can also increase adult car accident risk in the short-run. Smith (2015) finds that the transition to Daylight Savings Time (DST) increases adult fatal car accidents specifically through an effect on sleep-deprivation rather than through changes to the driving environment (through ambient light). However, this effect persists for less than a week in the adult population, whereas in the teenage population early start times can cause a persistent increase in car accident risk throughout the school-year.

3.3 Data & The Kansas Context

The data on high school start time includes each public high school in the state of Kansas over the school-years 2004-2005 to 2012-2013.[4] School-level covariates including enrollment by grade, enrollment by race/ethnicity, and urbanization codes were obtained from the National Center for Education Statistics' Common Core of Data (CCD) for school-years 2004-2005 to 2012-2013. Over this time period, there were a substantial number of small changes to high school start times across the state. These shifts were primarily driven by budgetary concerns. Many Kansas high schools chose to lengthen the school day by small amounts in order to reduce the total number of days in the school year. Table 3.1 displays the amount of within-school variation in start time. In each year, approximately 10% of schools changed their start times by an average of 9 minutes. This average time change encompasses substantial variation with some schools shifting by as much as 55 minutes.

Table 3.1: Variation in High School Start Time (Within School)

Year	# Schools	# Start Changers	% Changed	Avg Mins Changed	Std. Dev. Mins Changed	Max Mins Changed
2005	323					
2006	321	25	7.79	9.68	12.07	55
2007	315	29	9.21	9.17	8.59	35
2008	310	26	8.39	8.35	6.14	35
2009	312	29	9.29	8.69	4.58	20
2010	308	57	18.51	9.09	4.77	25
2011	305	49	16.07	9.43	5.82	30
2012	304	36	11.84	7.08	5.22	30
2013	294	23	7.82	6.83	3.52	15
Total	2,792	274	9.81	8.65	6.48	55

One advantage to focusing on the state of Kansas for this analysis is that it has one of the highest rates of teen driving in the country. The 2013 Youth Risk Behavior Survey

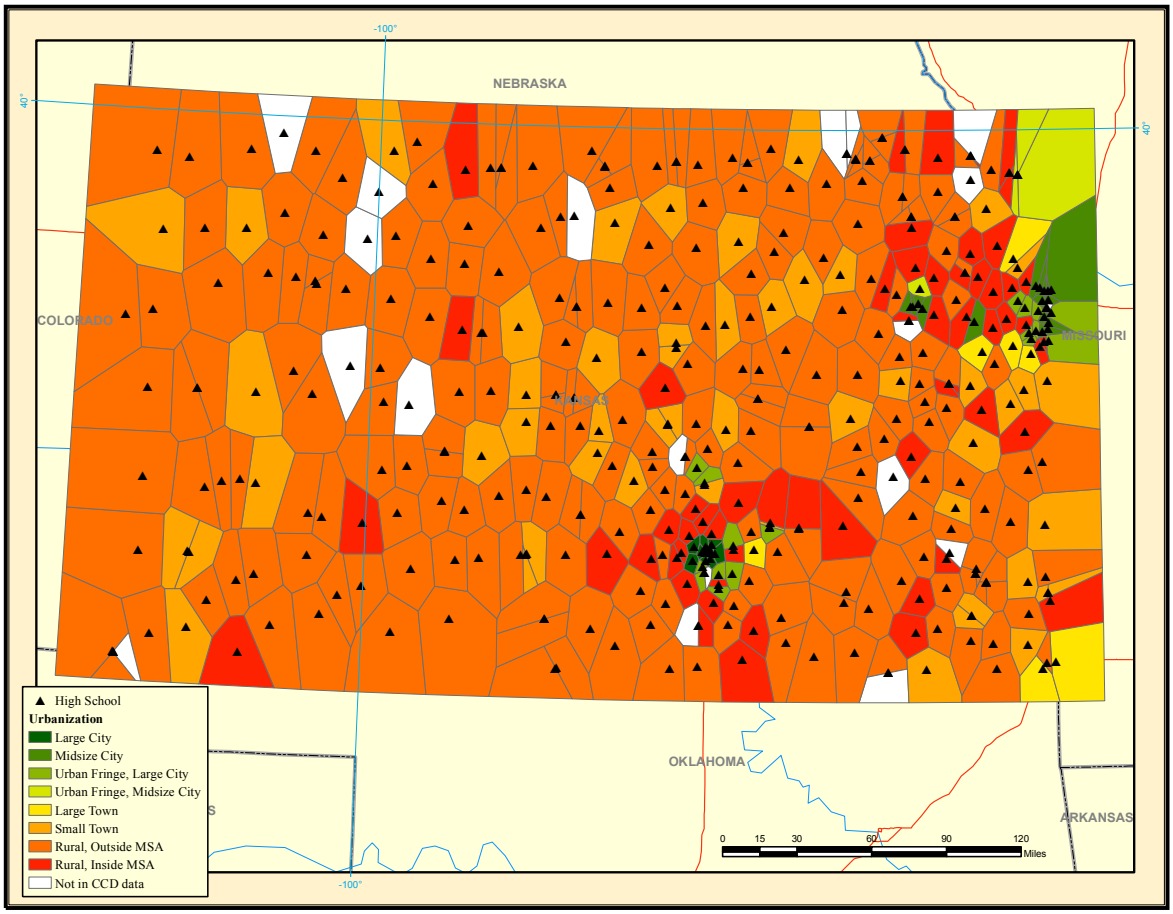
[4]Data for the years 2006-2007 through 2012-2013 was provided by the Kansas Department of Education. Additional years of data were provided thanks to Dr. Peter Hinrichs (2004-2005 to 2005-2006).

found that over 86% of Kansas teens aged 16-19 reported driving in the month prior to the survey (compared to 76% nationally) (Shults et al., 2015). An earlier survey, the 2006 National Young Driver Survey, indicates that a large fraction of this teen driving includes trips to and from school. This nationally-representative survey finds that 57% of 9th-11th graders report driving to school (Winston, 2007). In studying teen driving outcomes, it is important to note that teen drivers have a much higher accident rate than adult drivers. Drivers aged 16-19 are 3 times more likely to be in a fatal car accident than adults according to the US Department of Transportation's Fatality Analysis Reporting System.[5] However, improved car safety and stricter driving laws have made substantial progress in mitigating teen riskiness such that fatal teen car accidents have actually declined by 55% in the years 2004-2013 (Shults et al., 2015). One such law targeted at teen driving behavior was implemented in Kansas during this period. Prior to 2010, Kansas teens were eligible to receive a learner's permit at the age of 14 and a full driver's license at the age of 16. Starting in 2010, the state implemented a graduated driver's license program which introduced an intermediate step called a conditional driver's license between receiving the learner's permit and the full license. The conditional permit is granted starting at age 16 and restricts underage passengers to siblings only and driving times to the hours of 5am-9pm. The full driver's license is then granted at age 17 or 6 months after receiving the conditional license, whichever is first.

Car accident data was provided by the Kansas Department of Transportation for the years 2004 to 2014. These data encompass every accident involving a driver aged 14-18 and include a unique accident identification number, date and time of accident, number of cars involved, and latitude/longitude coordinates for the location of the accident. I map each high school (excluding online schools and 24-hour schools) using latitude/longitude coordinates from the CCD and use Thiessen polygons to create a zone for each school

[5]http://www.nhtsa.gov/FARS

Figure 3.1: 2013 Urbanization of Kansas High Schools & Thiessen Polygons



such that all space within school A's zone is closer to school A than to the next nearest school. I then use the latitude/longitude coordinates from the car accident data to map each accident to a high school zone. This results in a count of car accidents involving a teen driver for each school in each month of each year (excluding weekends and summer months). In order to minimize misclassification and spill-over effects across closely clustered schools, I drop all schools that fall into in the Large City and Midsize City urbanization categories according to the CCD. Figure 3.1 maps each Kansas high school along with the corresponding Thiessen polygon and urbanization code for the most recent year of the data. Table 3.2 shows that, as Kansas is a predominantly rural state,

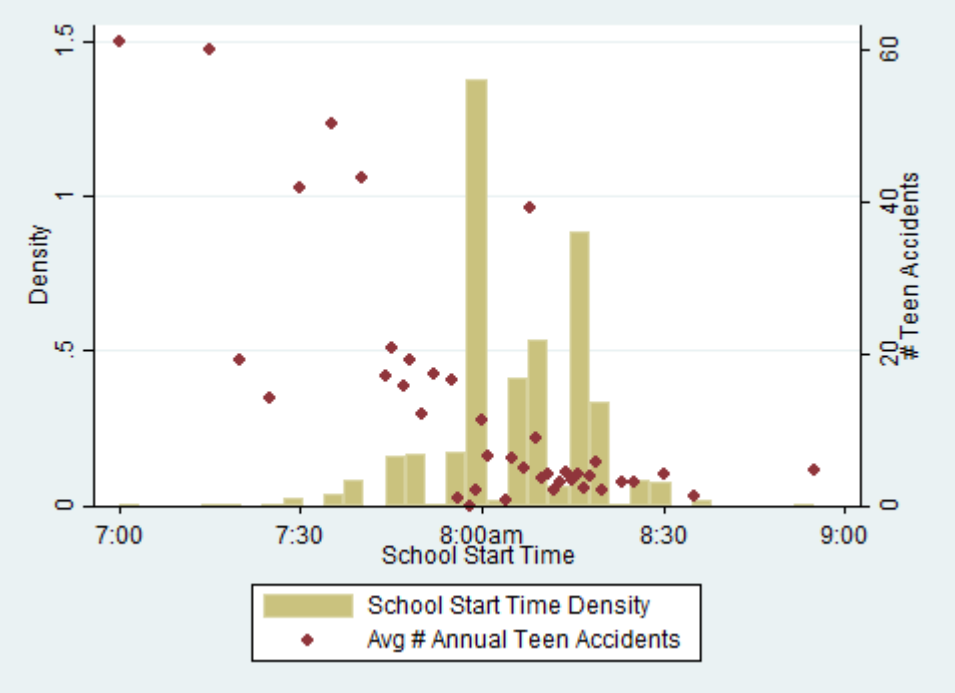
dropping these 2 urban categories eliminates only 8% of schools (although those schools do account for 39% of teen car accidents within the state).

Table 3.2: Urbanization Summary Statistics

	% of Obs	% of Teen Crashes
Large City	3.09	16.3
Midsize City	5.00	23.1
Urban Fringe of Large City	4.67	16.4
Urban Fringe of Midsize City	0.76	0.8
Large Town	1.94	5.3
Small Town	14.25	17.0
Rural, Outside MSA	58.06	8.8
Rural, Inside MSA	12.21	12.4

All data is aggregated to the school-year observation level and excludes the summer months (June-August) and weekends. Categories to be dropped: Large City & Midsize City.

Figure 3.2: Histogram of High School Start Times



The final sample is an unbalanced panel of 369 schools spanning 9 school-years (2004-

2005 to 2012-2013), resulting in 25,128 school-year-month observations.[6] A summary of the data is shown in Table 3.3. Additionally, Figure 3.2 shows the distribution of school start times along with the raw negative correlation between start time and average number of teen car accidents.

Table 3.3: Summary Statistics

	mean	sd	min	max
School Start Time	8:06am	11.26mins	7:00am	8:55am
Enrollment Grades 9-12 (10's)	33.53	41.94	0.500	216.5
% Students Black	1.977	3.837	0	35.44
% Students Hispanic	7.125	11.34	0	85.19
Car Accidents - Drivers Aged 14-18:				
Total	0.978	2.347	0	30
2-Car Only	0.658	1.784	0	22
Single-Car Only	0.241	0.592	0	8
Morning Commute (6am-10am)	0.194	0.603	0	9
2-Car Only	0.135	0.476	0	6
Single-Car Only	0.042	0.215	0	3
Late-Night (9pm-4am)	0.063	0.280	0	4
2-Car Only	0.023	0.168	0	3
Single-Car Only	0.039	0.206	0	3

All data is aggregated to the school-year-month observation level and excludes the summer months (June-August), weekends, and urban areas. N =25,128.

3.4 Empirical Strategy

To accomodate the count nature of the car accident data, I model the following log-linear relationship,

$$\log \mathbb{E}[A_{iym}] = \beta S_{iy} + \eta X_{iy} + DST_{ym} + GDL_{ym} + D_i + D_y + D_m, \tag{3.1}$$

[6] All results reported in Section 3.5 are robust to limiting the data to a balanced panel of 259 schools.

where the dependent variable, A_{iym}, is the count of accidents involving teens aged 14-18 near high school i, in year y and month m. The variable S_{iy} measures the start time for each high school as the number of minutes after midnight and is then divided by 15 (e.g. 8am=8*60/15=32) so that the coefficient, β, can be interpreted as the effect of a 15 minute delay in school start time. The vector X_{iy} includes total enrollment for grades 9-12, the percent of students who are black, percent of students who are hispanic, and urbanization indicators for each school and year. The vector DST_{ym} is a set of indicators meant to capture the 2007 change in the span of Daylight Savings Time. The Energy Policy Act of 2005 moved the start of Daylight Savings Time from April to March and the end from October to November. Therefore, DST_{ym} includes 4 indicators for March, April, October, and November each interacted with a post-change indicator ($\mathbb{1}[y \geq 2007]$). The variable GDL_{ym} is an indicator variable for all months after the introduction of the graduated driver's license in January of 2010. The model also includes fixed effects for each school, year, and month.

This model is not well-estimated by Least Squares because of the high incidence of zero-count observations on monthly teen car accidents. I instead estimate the model using Fixed-Effects Poisson Quasi-Maximum Likelihood (QML).[7] This estimator has the useful property of being robust to misspecification of the density function as Poisson and instead requires only that the conditional mean be correctly specified: $\mathbb{E}[A_{iym}|Z_{iym}] = D_i \exp(Z'_{iym}\gamma)$ where $Z_{iym} = (S_{iy}, X'_{iy}, DST_{ym}, GDL_{ym}, D_y, D_m)$ and γ is the corresponding vector of coefficients in (3.1). The QML estimator also corrects for the common problem of excess zeros in count data. This issue arises when the count variable includes a high incidence of zero-count observations (as in the teen accident data) – much higher than a Poisson distribution would predict. A benefit of the QML estimator is that once condi-

[7] To account for the fact that each school has a different student population, and therefore a different potential for teen car accidents, I include total enrollment for grades 9-12 as the exposure variable in the Poisson regressions.

tioned on the total number of accidents within a school, the conditional density becomes a multinomial distribution (Hausman et al., 1984). The multinomial density will easily fit a large number of zero-count observations to a school that has a low total count of accidents over all time periods.[8] Thus, the Fixed-Effects Poisson Quasi-Maximum Likelihood approach is particularly well-suited to the context of teen car accident data. To address the possibility that overdispersion of the data may lead to understated standard errors, I implement cluster-robust standard errors which account for both overdispersion and within-school correlation in the dependent variable (Wooldridge, 1999).[9]

3.5 Results

The results of estimating equation (3.1) via Poisson Fixed-Effects QML are shown in Table 3.4. Coefficients can be interpreted as the percent change in teen car accidents due to a 15 minute change in school start time. Columns (1) and (2) include accidents at all times of day and show that there is no overall effect of school start time on either 2-car or single-car teen accidents. However, it is likely that these estimates incorporate a significant amount of noise if the effect is concentrated during students' morning commute. Columns (3) and (4) include only accidents occurring between 6am and 10am (this spans from 1 hour before the earliest start time to 1 hour after the latest). Here we see that the estimated effect of a 15 minute delay in school start time is a 21% increase in 2-car

[8]Note that this model cannot be estimated for schools where the total count of accidents over all time periods is zero. Thus any schools that do not experience a teen car accident at any point over the 9 years of the data will be dropped from the estimation sample. (This will be especially salient in the specifications where I limit the dependent variable to teen accidents occuring during narrow time windows within the school day.) However, the QML estimator for a Poisson model is analytically identical to the unconditional Maximum Likelihood (ML) estimator in a model that includes dummy variables for each individual school (Lancaster, 2000). The primary disadvantage to the ML approach is that estimation with the inclusion of so many indicator variables is often computationally infeasible.

[9]The QML estimator only restricts the within-school mean and variance to be equal, so that the vast majority of the overdispersion in the teen car accident data is accounted for even without robust standard errors.

teen accidents. Given that the average number of 2-car morning teen car accidents per school per month is 0.135, this effect is equivalent to approximately 124 additional teen accidents per year across all of rural Kansas. This result indicates that any potential effect of a delay in school bell time due to decreased sleep deprivation is completely offset by the congestion effect of moving student commute times into high-volume traffic hours. This finding is, to some extent, specific to the context of this data in that the majority of school start times in Kansas are between 7:30-8:30am so that the effect of delays to school start times are identified off of movement in and around peak worker commuting hours.

Table 3.4: Effect of High School Start Time on Teen Car Accidents

	All Day		Morning Commute (6-10am)		Late-Night (9pm - 4am)	
	2-Car	Single-Car	2-Car	Single-Car	2-Car	Single-Car
	(1)	(2)	(3)	(4)	(5)	(6)
Start Time	**0.037**	**-0.086**	**0.207****	**0.080**	**-0.241**	**-0.292****
	(0.071)	**(0.059)**	**(0.104)**	**(0.094)**	**(0.214)**	**(0.133)**
School FEs	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X
Month FEs	X	X	X	X	X	X
N	23,373	24,066	18,954	18,099	10,332	17,649

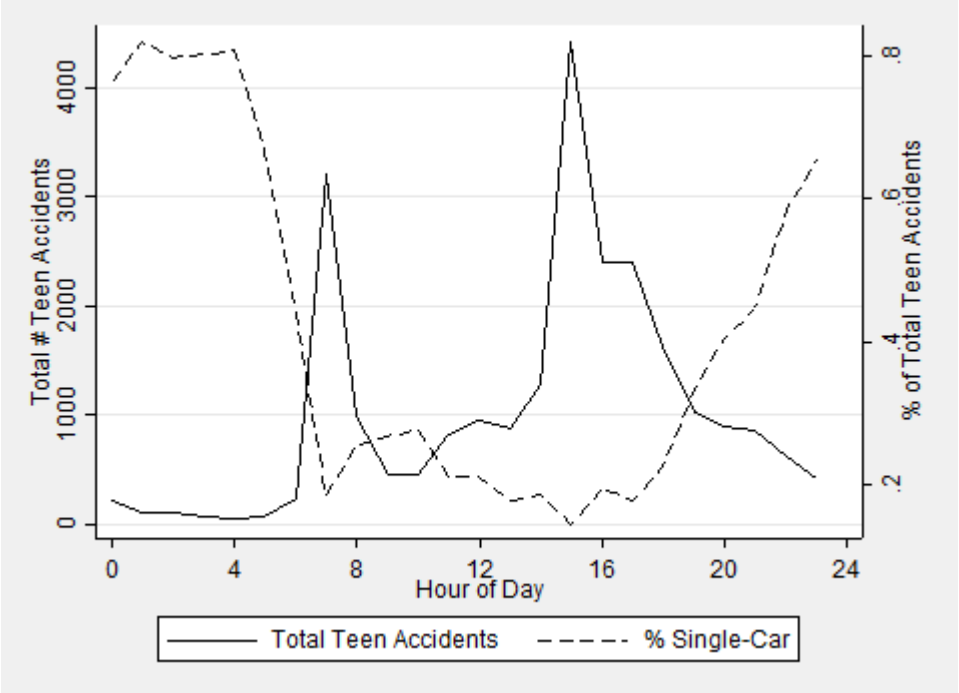
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Standard errors in parentheses are clustered by school. All columns include: total enrollment, % students black, % students hispanic, urbanization indicators, indicators for the 2007 change to Daylight Savings Time and the 2010 change to teen driver's license requirements. Each column also includes total enrollment as the exposure variable with coefficient constrained to be one.

In order to isolate the sleep effect of delayed start time from this congestion effect, I next focus on the late-night time period when there is little scope for congestion effects on teen accident risk. If there is a sleep effect, then it should still be detectable in the evening as medical research indicates that the physical effect of sleep deprivation persists – and in fact increases – throughout the day (Durmer and Dinges, 2005). Column (5) of Table 3.4 shows the results of estimating the model in (3.1) with the dependent variable restricted

to 2-car teen accidents occuring between the hours of 9pm and 4am on weeknights (this excludes both Friday night and the early-morning hours of Monday) during the school year. The estimate of the effect of a 15 minute delay in school start time on 2-car teen accidents is now negative and very large, although not statistically significant. However, this is not surprising given that the vast majority of late-night accidents are single-car events. Figure 3.3 shows that between midnight and 5am approximately 80% of all teen accidents involve only 1 driver.[10] Additionally, previous survey research shows that young drivers are both more likely to be in single-car accidents and to cite drowsiness as the cause of such accidents (Gislason et al., 1997).

Figure 3.3: Teen Car Accidents By Hour



For this reason, I next analyze the sample of single-car accidents involving one teen driver in column (6) of Table 3.4. The estimated effect on this type of teen accident is a 26% decrease in the late-night hours. Given that the average number of late-night, single

[10]Note that this is also reflected in the very small sample in column (5). Most schools in the sample have zero-count observations for 2-car, late-night accidents in all years.

car teen accidents per school in a month is 0.039, this effect is equivalent to approximately 68 fewer accidents of this type per year for teen drivers across all of rural Kansas. This result suggests that there is a significant sleep effect on teen accident rates. This finding coincides with research on sleep apnea, which finds that sleep apnea patients had no more multi-car accidents than control drivers but were 7 times more likely to experience single-car accidents (Haraldsson et al., 1990). An alternative explanation might be that teens are simply driving less at night when start times are later in the morning. While I cannot test this directly due to a lack of age- and location-specific data on vehicle miles traveled, it is a seemingly counterintuitive theory.

Given that most late-night accidents involve only one car, there is likely very little interaction between teen drivers and adult drivers during this time window. This independence make a difference-in-differences analysis using adult drivers as the "control" group feasible. This type of analysis is attractive for 2 reasons: first, including adult accidents may better control for school-specific time trends in road safety; and second, each school is unlikely to have a total count of zero car accidents over all time periods once adult accidents are included so that very few schools will be dropped from the estimation sample.

Data on adult car accidents was provided by the Kansas Department of Transportation for drivers aged 30-35 in the years 2004 to 2014. This age group is ideal because these drivers are well past their teenage years, but also not generally old enough to have high school children of their own. As with the teen accident data, I use latitude/longitude coordinates to map each adult accident into a high school zone. This results in a count of car accidents involving an adult driver aged 30-35 for each school in each month of each year (dropping weekends and summer months). I then estimate the following model

using Poisson Fixed-Effects QML,

$$\log \mathbb{E}[A_{igym}] = \beta_1 S_{iy} + \beta_2 T_g + \beta_3 S_{iy} * T_g + X'_{iy}\eta + DST_{ym} + GDL_{ym} + D_i + D_y + D_m, \quad (3.2)$$

where the dependent variable, A_{igym}, is the count of single-car accidents involving a driver in age group $g \in \{14-18, 30-35\}$, near high school i, in year y and month m. The variable T_g is an indicator for age group $g = 14-18$ so that the coefficient, β_1 captures the effect of a delay in start time on adult accidents (this is expected to be zero), β_2 captures the baseline difference in accident rates between teens and adults, and β_3 captures the difference in the effect of a start time delay between teens and adults. The total effect of a delay in start time on teen drivers is thus captured by the sum of the coefficients $\beta_1 + \beta_3$.

Table 3.5: Difference-in-Differences Analysis of Single-Car, Late-Night Accidents

	9pm-4am	8pm-5am	7pm-6am
	(1)	(2)	(3)
Start Time	-0.095	-0.046	0.001
	(0.097)	(0.080)	(0.070)
Teen	7.010***	7.106***	7.600***
	(2.162)	(1.949)	(1.911)
Start*Teen	-0.235***	-0.240***	-0.259***
	(0.067)	(0.060)	(0.059)
Total Effect on Teen Accidents	**-0.330*** (0.109)**	**-0.285*** (0.092)**	**-0.258*** (0.088)**
N	45,828	47,736	48,798
School FEs	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes
Month FEs	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Standard errors in parentheses are clustered by school. All columns include: total enrollment, % students black, % students hispanic, urbanization indicators, indicators for the 2007 change to Daylight Savings Time and the 2010 change to teen driver's license requirements. Each column also includes total enrollment as the exposure variable with coefficient constrained to be one.

Table 3.5 displays the results of this estimation for 3 time periods: 9pm-4am, 8pm-5am, and 7pm-6am. The estimated effect of a 15 minute delay in school start time is now somewhat larger and more precise at a 26-33% decrease in late-night, single-car teen accidents. As expected, the point estimate of the effect of start time on adult single-car accidents is small and insignificant, indicating an absence of systematic, confounding factors that would cause a simultaneous decrease in both teen and adult accidents. These results further bolster the findings in Table 3.4, suggesting that there is a persistent sleep mechanism affecting teen car accident risk.

Table 3.6 shows that the main results are also robust to alternative definitions of both the morning commute and the late-night period. Columns (1) and (2) allow for a more generous definition of the morning commute window (5-10am). Meanwhile, columns (3) and (4) allow for a more flexible definition with a separate time window defined for each school based on its own start time in each year. Columns (5) and (6) show that the main results are consistent for an alternative late-night window of 7pm-6am.

Table 3.6: Robustness Check: Alternative Time Windows

	Morning Commute (5-10am)		Morning Commute (1hr before bell - 1hr after)		Late-Night (7pm - 6am)	
	2-Car	Single-Car	2-Car	Single-Car	2-Car	Single-Car
	(1)	(2)	(3)	(4)	(5)	(6)
Start Time	**0.207****	**0.068**	**0.234****	**0.037**	**-0.159**	**-0.260****
	(0.103)	**(0.092)**	**(0.113)**	**(0.108)**	**(0.116)**	**(0.108)**
School FEs	X	X	X	X	X	X
Year FEs	X	X	X	X	X	X
Month FEs	X	X	X	X	X	X
N	19,035	18,504	18,666	16,821	14,598	20,556

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Standard errors in parentheses are clustered by school. All columns include: total enrollment, % students black, % students hispanic, urbanization indicators, indicators for the 2007 change to Daylight Savings Time and the 2010 change to teen driver's license requirements. Each column also includes total enrollment as the exposure variable with coefficient constrained to be one.

Finally, to ensure further that these findings are not being driven by some unobserved factor rather than by changes in school start time, I employ a placebo test with randomly

assigned start times. For the balanced panel of 259 schools, I draw a random start time for each year from a normal distribution with a mean of 8:05am and a standard deviation of 11 minutes (this is a relatively close approximation of the actual start time distribution as seen in Figure 3.2). With this simulated data I then re-estimate the 2-car regression for the morning commute hours (column 3 of Table 3.4), the single-car estimates for late-night hours (column 6 of Table 3.4), and the difference-in-differences estimate for single-car, late-night accidents (column 1 of Table 3.5). Table 3.7 reports the point estimates of interest averaged over 1,000 iterations of simulated start time data as well as the rejection rate for a 5% t-test. The point estimates are all very close to zero and the rejection rates are all nearly 5% indicating that there is no placebo effect driving the main results. Furthermore, note that all of the point estimates from the main findings lie outside the range of even the largest estimates arising from the simulated data.

Table 3.7: Placebo Test: Randomly Assigned Start Times

	Morning Commute (6-10am) 2-Car	Late-Night (9pm-4am) Single-Car	Diff-in-Diff (9pm-4am) Single-Car
	(1)	(2)	(3)
Avg. Start Time Coefficient	-0.001	0.007	0.006
Range of Start Time Coeff.	[-0.084, 0.057]	[-0.106, 0.171]	[-0.103, 0.152]
5% Rejection Rate	0.059	0.039	0.049
Iterations	1,000	1,000	1,000

In each iteration, start times are drawn randomly for each school-year observation from a normal distribution with a mean of 8:05am and standard deviation of 11 minutes.

All iterations in all columns include: school fixed effects, year indicators, month indicators, total enrollment, % students black, % students hispanic, urbanization indicators, indicators for the 2007 change to Daylight Savings Time and the 2010 change to teen driver's license requirements as well as total enrollment as the exposure variable with coefficient constrained to be one.

3.6 Conclusion

To combat the problem of chronic sleep deprivation among today's youth, many US school districts are opting to delay high school start times, thereby allowing teens additional sleep time in the morning. However, this solution may have unexpected consequences in the form of changes to teen driving patterns and car accident rates. Changes to school start times can alter teen car accident risk both through a direct effect on sleep deprivation and indirectly through changes to the driving environment, making the direction of the overall effect theoretically ambiguous.

I utilize within-school variation in start times across the state of Kansas to identify the effect of high school start time on the average number of teen car accidents. I find evidence suggesting that this effect is positive during the morning commute hours, indicating that any effect stemming from avoided sleep loss is completely offset by the effect of shifting teen driving into a more congested hour of morning traffic patterns. The estimated effect of a 15 minute delay in high school start times is a 21% increase in morning, 2-car teen accidents. However, I also find evidence of a persistent sleep-deprivation effect by focusing on late-night, single-car accidents. At these times there is very low traffic volume, making it possible to observe the direct effect of decreased sleep deprivation on teen accidents. I find that a 15 minute delay in school start times leads to a 26% decrease in late-night teen accidents.

Taken together, these findings provide evidence that changes to school start time do have an effect on teen car accidents and that both a sleep effect and an opposing traffic congestion effect are in play. Whether these two effects balance out to be a cost to later start times or a benefit may depend on the local traffic environment. The estimates presented here may be somewhat specific to the rural, midwestern context. In a more urban setting with higher traffic volume throughout the day, there might be a much

larger positive effect during the morning commute and a much smaller negative effect in the evening hours. Additionally, urban settings have a much higher overall teen accident rate such that a 21% increase in morning teen accidents might amount to a significant cost to consider in the start time policy decision.

Appendix A

Appendix for Chapter 1

The following Stata code was used to create Table 1.2. The code estimates the model in Section 1.5 under the alternative hypothesis of two regimes using the EM algorithm and then under the null hypothesis of a single regime using the Stata `ml` command. Finally, the QLR test statistic is calculated.

```
* Estimating QLR test statistic for Bloom et al (2003)

* Log likelihood function with 2 regimes
capture program drop llf
program define llf
version 10.1
args lnf theta1 theta0 delta sigma lambda
quietly replace `lnf'=(1/_N)*((1-etahat)*(ln((2*_pi*`sigma'^2)^(-1/2))
   +((-1/(2*`sigma'^2))*(lgdp-`theta0'-`delta'*latitude)^2)+ln(1-`lambda'))
   +etahat*(ln((2*_pi*`sigma'^2)^(-1/2))+((-1/(2*`sigma'^2))*(lgdp-`theta1'
   -`delta'*latitude)^2)+ln(`lambda')))
end

* Log likelihood function for a single regime
capture program drop llfsingle
program define llfsingle
```

```
version 10.1
args lnf theta delta sigma
quietly replace `lnf'= (1/_N)*ln(((2*_pi*`sigma'^2)^(-1/2))*
   exp((-1/(2*`sigma'^2))*(lgdp-`theta'-`delta'*latitude)^2))
end

/*****************************************************/
* First estimate parameters and log likelihood for the case of 2 regimes:
* lgdp = theta0 + delta*latitude + u~N(0,sigma2) with probability (1-lambda)
* lgpp = theta1 + delta*latitude + u~N(0,sigma2) with probability lambda

/*****************************************************/
* Start with initial guess for theta0, theta1, delta, sigma2, and lambda:
reg lgdp latitude
mat beta=e(b)
svmat double beta, names(matcol)
scalar dhat=betalatitude
gen intercept=lgdp-dhat*latitude
summarize intercept
scalar t0hat=r(mean)-r(Var)
scalar t1hat=r(mean)+r(Var)
scalar shat=sqrt(r(Var))
scalar lhat=0.5
matrix gammahat=(t1hat, t0hat, dhat, shat, lhat)
di "Original guess for parameter values: "
matrix list gammahat

/*****************************************************/
* Start loop that continues until parameter estimates have converged
gen error1=10
gen error2=10
```

```
gen error3=10
gen tol=1/_N
gen count=0
gen count1=1
gen count2=1
gen count3=1
gen f1=0
gen f0=0
gen fboth=0
gen etahat=0
gen llfhat=0
gen llfnew=0
gen fdelta=0
gen fnew=0
gen Inllfnew=0
gen Inllfdelta=0
gen nd1=0
gen nd2=0
gen nd3=0
gen nd4=0
gen nd5=0

while error1>tol | error2>tol | error3>tol {

* Calculate guess for eta_t=Pr(St=1|sample)
* Calculate f(Yt|St=1, gammahat)
quietly replace f1=((2*_pi*gammahat[1,4]^2)^(-1/2))*
   exp((-1/(2*gammahat[1,4]^2))*(lgdp-gammahat[1,1]-gammahat[1,3]*
   latitude)^2)
* Calculate f(Yt|St=0, gammahat)
quietly replace f0=((2*_pi*gammahat[1,4]^2)^(-1/2))*
```

```
   exp((-1/(2*gammahat[1,4]^2))*(lgdp-gammahat[1,2]-gammahat[1,3]*
   latitude)^2)
* Calculate f(Yt|gammahat)
quietly replace fboth=gammahat[1,5]*f1+(1-gammahat[1,5])*f0
quietly replace etahat=gammahat[1,5]*f1/fboth

/*****************************************************/
* Now use etahat to create and maximize log likelihood function

ml model lf llf /theta1 /theta0 /delta /sigma /lambda
ml init gammahat, copy
ml max
mat gammanew=e(b)

/*****************************************************/
* Check whether the parameter estimates have converged
mata: st_matrix("temp", max(abs(st_matrix("gammanew")-st_matrix("gammahat"))))
quietly replace error1=temp[1,1]

* Check whether the log likelihood has converged
quietly replace llfnew=e(ll)
quietly replace llfhat=(1/_N)*((1-etahat)*(ln((2*_pi*gammahat[1,4]^2)^(-1/2))
   +((-1/(2*gammahat[1,4]^2))*(lgdp-gammahat[1,2]-gammahat[1,3]*latitude)^2)
   +ln(1-gammahat[1,5]))+etahat*(ln((2*_pi*gammahat[1,4]^2)^(-1/2))
   +((-1/(2*gammahat[1,4]^2))*(lgdp-gammahat[1,1]-gammahat[1,3]*latitude)^2)
   +ln(gammahat[1,5])))
quietly summarize llfhat
quietly replace llfhat=r(sum)
quietly replace error2=abs(llfhat-llfnew)

* Check whether the numeric derivative is zero
```

```
* Recalculate incomplete log likelihood with new gamma estimates
quietly replace f1=((2*_pi*gammanew[1,4]^2)^(-1/2))*
   exp((-1/(2*gammanew[1,4]^2))*(lgdp-gammanew[1,1]-gammanew[1,3]*latitude)^2)
quietly replace f0=((2*_pi*gammanew[1,4]^2)^(-1/2))*
   exp((-1/(2*gammanew[1,4]^2))*(lgdp-gammanew[1,2]-gammanew[1,3]*latitude)^2)
quietly replace fnew=gammanew[1,5]*f1+(1-gammanew[1,5])*f0
quietly replace Inllfnew=log(fnew)
quietly summarize Inllfnew
quietly replace Inllfnew=r(sum)/_N
* Calculate incomplete log likelihood for gamma + 0.0001
forval i=1/5 {
matrix gammadelta=gammanew
matrix gammadelta[1,`i']=gammadelta[1,`i']+.0001
quietly replace f1=((2*_pi*gammadelta[1,4]^2)^(-1/2))*
   exp((-1/(2*gammadelta[1,4]^2))*(lgdp-gammadelta[1,1]-gammadelta[1,3]*
   latitude)^2)
quietly replace f0=((2*_pi*gammadelta[1,4]^2)^(-1/2))*
   exp((-1/(2*gammadelta[1,4]^2))*(lgdp-gammadelta[1,2]-gammadelta[1,3]*
   latitude)^2)
quietly replace fdelta=gammadelta[1,5]*f1+(1-gammadelta[1,5])*f0
quietly replace Inllfdelta=log(fdelta)
quietly summarize Inllfdelta
quietly replace Inllfdelta=r(sum)/_N
quietly replace nd`i'=abs(Inllfdelta-Inllfnew)/.0001
}
quietly replace error3=max(nd1,nd2,nd3,nd4,nd5)

/***************************************************/
* Keep track of when each convergence criterion is met
quietly replace count1=count1+1 if error1>tol
quietly replace count2=count2+1 if error2>tol
```

```
quietly replace count3=count3+1 if error3>tol

* Update gammahat and overall iteration count
matrix gammahat=gammanew
quietly replace count=count+1

* End of loop
}

/***************************************************/
* Calculate final log likelihood for 2 regimes
quietly replace f1=((2*_pi*gammanew[1,4]^2)^(-1/2))*
   exp((-1/(2*gammanew[1,4]^2))*(lgdp-gammanew[1,1]-gammanew[1,3]*latitude)^2)
quietly replace f0=((2*_pi*gammanew[1,4]^2)^(-1/2))*
   exp((-1/(2*gammanew[1,4]^2))*(lgdp-gammanew[1,2]-gammanew[1,3]*latitude)^2)
gen f2reg=gammanew[1,5]*f1+(1-gammanew[1,5])*f0
gen llf2reg=ln(f2reg)
quietly summarize llf2reg
quietly replace llf2reg=r(sum)
* Output final parameter estimates
disp "Final estimated parameter values for 2 regimes:  "
matrix list gammanew
disp "Final estimated log likelihood for 2 regimes: " llf2reg
disp "Total number of loop iterations: " count
disp "Parameter values converged after "count1 " iterations"
disp "Log likelihood value converged after " count2 " iterations"
disp "Gradient of Log likelihood converged after " count3 " iterations"

/***************************************************/
* Second, estimate parameters and log likelihood for the case of only 1 regime:
```

```
* Maximize log likelihood with only 1 regime
* lgdp = theta + delta*lat + u~N(0,sigma2)
quietly summarize intercept
matrix gamma0=(r(mean), dhat, .1)
* Maximize to find new estimate of gamma
ml model lf llfsingle /theta /delta /sigma
ml init gamma0, copy
ml max
mat gammasingle=e(b)

*Calculate log likelihood for 1 regime with estimated gamma
gen llf1reg=ln(((2*_pi*gammasingle[1,3]^2)^(-1/2))*
   exp((-1/(2*gammasingle[1,3]^2))*(lgdp-gammasingle[1,1]-gammasingle[1,2]*
   latitude)^2))
quietly summarize llf1reg
quietly replace llf1reg=r(sum)
* Output final parameter estimates
disp "Final estimated parameter values for 1 regime:  "
matrix list gammasingle
disp "Final estimated log likelihood for 1 regime: " llf1reg

/************************************************/
* Finally, calculate QLR test statistic:
gen QLR=2*(llf2reg-llf1reg)
disp "Quasi-Likelihood Ratio test statistic of 1 regime: " QLR
```

Appendix B

Appendix for Chapter 2

Figure B.1: Access to Elite Colleges by Region (17 Region Classification)

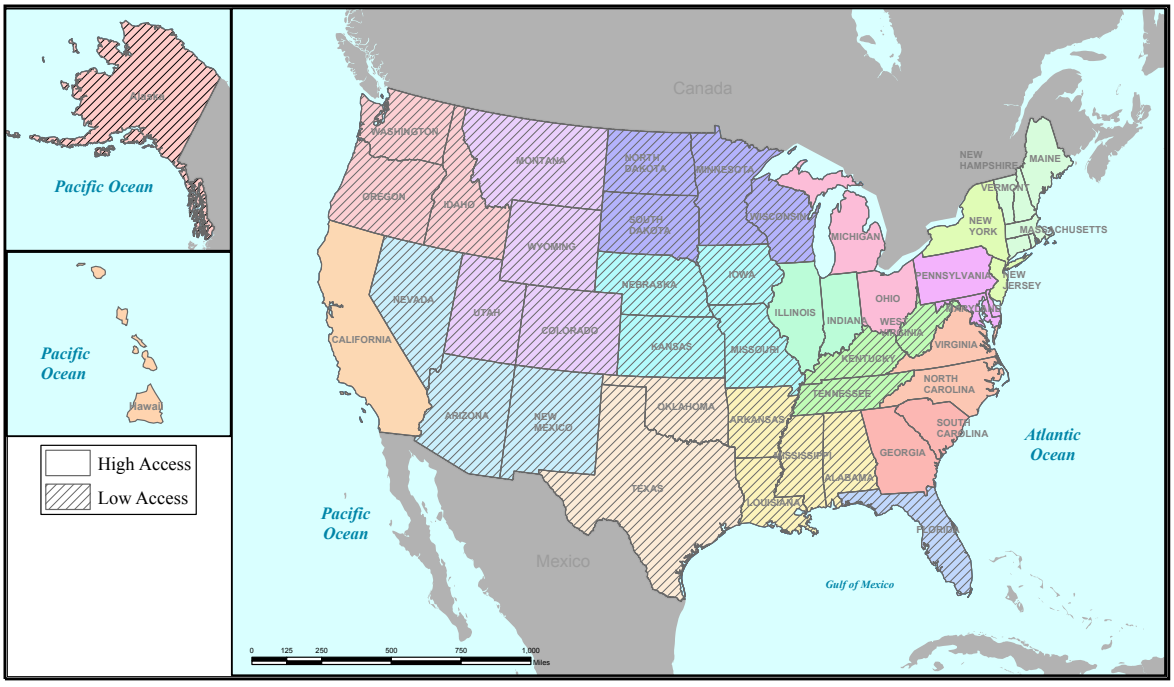


Table B.1: Elite Colleges (Top 5%)

	USN		IPEDS	
Institution Name	75th Percentile SAT Score	2012 Rank	75th Percentile SAT Score	2001 Rank
California Institute of Technology	1590	1	1580	2
Harvard University	1590	2	1580	1
Princeton University	1590	3	1540	6
Yale University	1590	4	1510	12
Harvey Mudd College	1570	5	1540	5
Massachusetts Institute of Technology	1570	6	1560	3
University of Chicago	1570	7	1490	16
Columbia University	1560	8	--	--
Dartmouth College	1560	9	1520	8
Pomona College	1550	10	1520	9
Stanford University	1550	11	1550	4
Amherst College	1530	12	1510	10
Duke University	1530	13	--	--
Rice University	1530	14	1440	35
Swarthmore College	1530	15	1530	7
University of Pennsylvania	1530	16	1490	18
Williams College	1530	17	1510	11
Brown University	1520	18	1490	15
Northwestern University	1520	19	1470	20
University of Notre Dame	1520	20	1450	26
Vanderbilt University	1520	21	1400	60
Washington University in St Louis	1520	22	1470	19
Carleton College	1510	23	1460	23
Carnegie Mellon University	1510	24	1460	24
Johns Hopkins University	1510	25	1490	17
Cooper Union for the Advancement of Science and Art	1500	26	--	--
Cornell University	1500	27	1500	13
Haverford College	1500	28	1450	27
Tufts University	1500	29	1410	54
Barnard College	1490	30	1410	55
Bowdoin College	1490	31	1440	33
Georgetown University	1490	32	1460	21
University of California-Berkeley	1490	33	1450	28
University of Southern California	1490	34	1400	59
Wellesley College	1490	35	1440	32
Claremont McKenna College	1480	36	1440	34
Middlebury College	1480	37	--	--
Washington and Lee University	1480	38	1420	48

Table B.1: Elite Colleges (Top 5%)

	USN		IPEDS	
Institution Name	75th Percentile SAT Score	2012 Rank	75th Percentile SAT Score	2001 Rank
Wesleyan University	1480	39	1460	22
Emory University	1470	40	1460	25
Hamilton College	1470	41	1370	84
Reed College	1470	42	1420	44
Rensselaer Polytechnic Institute	1470	43	1400	61
Vassar College	1470	44	1430	36
New York University	1460	45	1420	43
Oberlin College	1460	46	1420	41
University of Virginia-Main Campus	1460	47	1420	42
Brandeis University	1450	48	1430	38
College of William and Mary	1450	49	1420	40
Davidson College	1450	50	1410	51
Georgia Institute of Technology-Main Campus	1450	51	1420	47
Scripps College	1450	52	1360	87
Boston College	1440	53	1390	64
Case Western Reserve University	1440	54	1440	31
Colgate University	1440	55	1400	58
Macalester College	1440	56	1430	39
Smith College	1440	57	1370	78
University of California-Los Angeles	1440	58	1400	63
University of Rochester	1440	59	1410	50
Whitman College	1440	60	1370	73
Bryn Mawr College	1430	61	1380	69
Kenyon College	1430	62	1380	67
Northeastern University	1430	63	1230	313
Rhodes College	1425	64	1380	68
Bard College	1420	65	--	--
Bates College	1420	66	1400	62
Colby College	1420	67	1410	53
Colorado College	1420	68	1380	70
Grinnell College	1420	69	1440	30
Hendrix College	1420	70	1340	108
Mount Holyoke College	1420	71	--	--
New College of Florida	1420	72	1420	46
St Olaf College	1420	73	1360	88
United States Air Force Academy	1420	74	1360	93
University of Michigan-Ann Arbor	1420	75	1402	57
Wheaton College	1420	76	1410	52

B.1 Simultaneous Decision Model

As in Section 2.2.1, let every individual have an ability level, a_i, drawn from a continuous distribution, $f(a)$. Employers cannot directly observe an individual's ability level, but instead receive two potential signals: college quality (Q) and major choice (M). I focus here on the simple case of only 2 college types (Q_H = elite and Q_L = non-elite) and 2 major choices (M_H = STEM and M_L = non-STEM).

Students must simultaneously decide where to apply to college and which field to major in. To identify the pure signaling effect, I assume that there is no human capital accumulation due to college quality and major choice so that, in equilibrium, firms set wages equal to expected ability. The individual's decision can then be written as,

$$\max_{Q_i, M_i} \mathbb{E}[a|Q_i, M_i] - C_{Q_H}(a_i) - C_{M_H}(a_i, Q_i). \tag{B.1}$$

The function $C_{Q_H}(a_i)$ represents the effort cost of attending an elite college (relative to the cost of attending a non-elite college). The function $C_{M_H}(a_i, Q_i)$ represents the additional effort cost of choosing a STEM major, which depends on both ability and college quality. Both of these costs are decreasing in ability, $\partial C_{M_H}/\partial a_i < 0$ and $\partial C_{Q_H}/\partial a_i < 0$, as in the traditional Spence model (Spence, 1973). As Spence points out, it is this decreasing cost assumption that is critical to ensuring that college quality and major choice serve as distinguishing signals and lead to a separating equilibrium. Furthermore, I assume that $C_{Q_H}(a_i) > C_{M_H}(a_i, Q_L)$ so that the equilibrium ability sorting is restricted to the case where the highest ability students choose (Q_H, M_H), followed by the next highest ability group choosing (Q_H, M_L), then (Q_L, M_H), and finally (Q_L, M_L). Removing this condition results in a separating equilibrium in which either no students at elite colleges choose non-STEM majors, which is clearly not supported by empirical evidence, or where the average

ability of STEM students at non-elite schools exceeds that of non-STEM students at elite colleges. Using the data and definitions described in Section 2.4, I find that this type of separating equilibrium is not consistent with the raw data on SAT scores. The average SAT scores are: 1,318 for STEM majors at elite schools, 1,264 for non-STEM majors at elite schools, 1,123 for STEM majors at non-elite schools, and 1,033 for non-STEM majors at non-elite schools.

Of course, the decision of where to attend college is not based on effort cost alone. There will also be some very high ability individuals who do not attend the best possible school that they qualify for. This could be due to the high financial cost or for other reasons such as a desire to stay close to home, etc. To account for this, I allow for a uniformly distributed constraint so that some fraction, $1 - p$, of all eligible students are directly constrained from attending an elite college. The resulting separating equilibrium is shown for a uniform ability distribution in the top panel of Figure B.2.

However, this equilibrium does not yet include the college admissions decision. It is a well-known empirical fact that there exists an excess supply of applicants to elite colleges. The top 5% of colleges (as defined in Section 2.4) admitted on average only 28.5% of applicants in the Fall of 2011 with some schools admitting less than 10% of applicants. This excess supply allows elite colleges to set a strict cutoff point in the distribution of ability, a^{Q_H}, that is higher than the minimum ability student who would like to attend an elite college, $a^{Q_H} >> a_{M_L}^{Q_H}$ (See bottom panel of Figure B.2). Students must now take the college admissions cutoff point as exogenously given, and adjust their college major choices accordingly. The students' maximization problem is now represented by (2.1) and it is clear that the resulting separating equilibrium and comparative statics will follow directly from the sequential model described in Section 2.2.1.

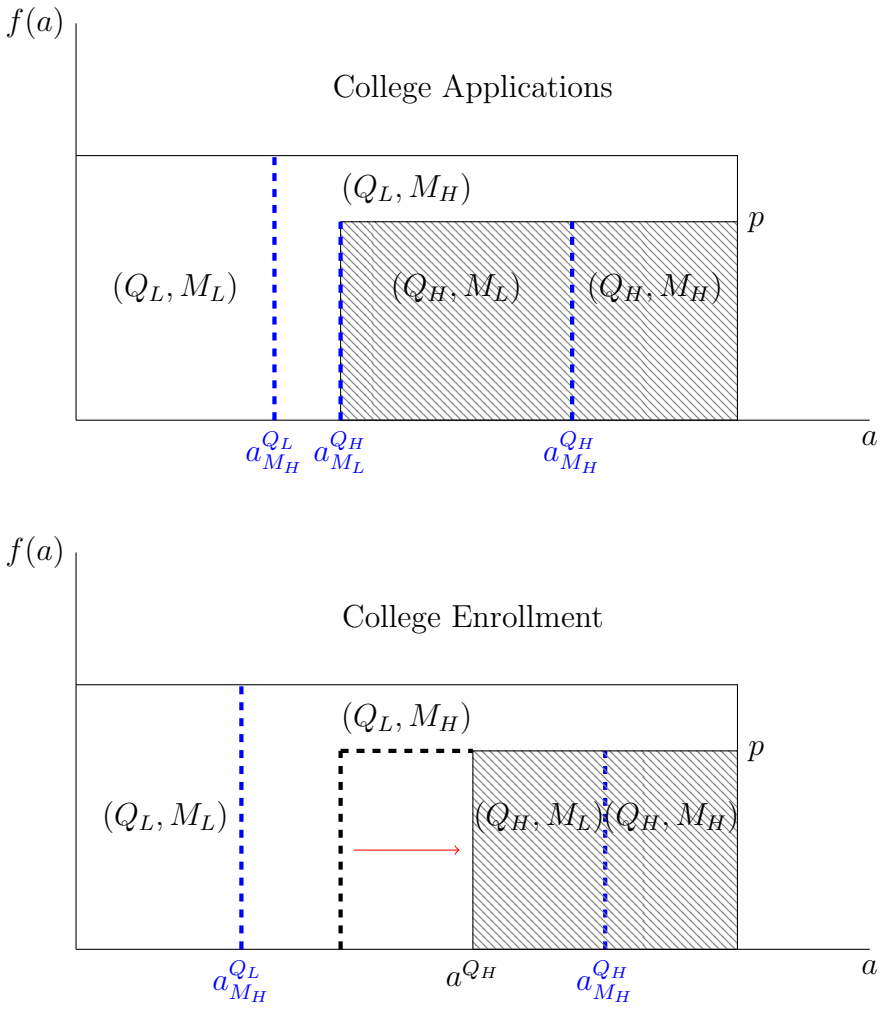


Figure B.2: Separating equilibrium for uniform ability before and after college admissions decisions

B.2 An Increasing Constraint Function

Consider a constraint function, $p(a)$ satisfying the following two conditions: (1) $p(a)$ must be increasing in ability (so that the fraction constrained, $1 - p(a)$, is smallest for the most able); and (2) $p(a) < 1$ for all ability levels in both the high and low access scenarios. Then the separating equilibrium for the model given in Section 2.2.1 must

satisfy the following modified conditions:

$$\mathbb{E}[a(p(a))|a^{Q_H} \leq a < a_{M_H}^{Q_H}] = \mathbb{E}[a(p(a))|a \geq a_{M_H}^{Q_H}] - C_{M_H}(a_{M_H}^{Q_H}, Q_H), \tag{B.2}$$

$$\mathbb{E}[a|a < a_{M_H}^{Q_L}] = \psi(a) - C_{M_H}(a_{M_H}^{Q_L}, Q_L), \tag{B.3}$$

where

$$\psi(a) = \tag{B.4}$$
$$\frac{[F(a^{Q_H}) - F(a_{M_H}^{Q_L})]\mathbb{E}[a|a_{M_H}^{Q_L} \leq a < a^{Q_H}] + (1-\theta)[1 - F(a^{Q_H})]\mathbb{E}[a(1-p(a))|a \geq a^{Q_H}]}{F(a^{Q_H}) - F(a_{M_H}^{Q_L}) + (1-\theta)[1 - F(a^{Q_H})]},$$

and

$$\theta = \int_{a^{Q_H}}^{\overline{a}} p(a)da.$$

The effect of an increase in $p(a)$ on the expected ability of non-elite, non-STEM students is now:

$$\frac{\partial \psi(a)}{\partial p(a)} = \tag{B.5}$$
$$\frac{[1 - F(a^{Q_H})][F(a^{Q_H}) - F(a_{M_H}^{Q_L})]}{\gamma}(\mathbb{E}[a|a_{M_H}^{Q_L} \leq a < a^{Q_H}] - \mathbb{E}[a(1-p(a))|a > a^{Q_H}])\frac{\partial \theta}{\partial p(a)}$$
$$+ \frac{(1-\theta)[1 - F(a^{Q_H})]}{\gamma}\frac{\partial \mathbb{E}[a(1-p(a))|a > a^{Q_H}]}{\partial p(a)}$$

where $\gamma = F(a^{Q_H}) - F(a_{M_H}^{Q_L}) + (1-\theta)[1 - F(a^{Q_H})]$. The first fraction in (B.5) is a ratio of populations, which is clearly positive: $1 - F(a^{Q_H}) > 0$; $F(a^{Q_H}) - F(a_{M_H}^{Q_L}) > 0$; $\gamma > 0$. The term $(\mathbb{E}[a|a_{M_H}^{Q_L} \leq a < a^{Q_H}] - \mathbb{E}[a(1-p(a))|a > a^{Q_H}])$ is the difference between the expected ability of STEM majors at non-elite schools who are below the admissions cutoff and the expected ability of the directly constrained students who are above the

admissions cutoff, which is clearly negative. The change in the area under the constraint function, $\partial\theta/\partial p(a)$, is positive making the entire first term of equation (B.5) negative.

The second term in (B.5) is the product of a ratio of populations (positive) and the change in the expected value of the ability of directly constrained students given a decrease in the fraction constrained from $1-p(a)$ to $1-\tilde{p}(a)$. Given that $1-p(a) \geq 1-\tilde{p}(a)$ $\forall a$, it follows that $a(1-p(a)) \geq a(1-\tilde{p}(a))$ $\forall a$ and therefore $\mathbb{E}[a(1-p(a))|a > a^{Q_H}] \geq \mathbb{E}[a(1-\tilde{p}(a))|a > a^{Q_H}]$. Given this decrease in the expected value of the ability of directly constrained students, the second term of (B.5) is also non-positive, and $\frac{\partial\psi(a)}{\partial p(a)} < 0$. Thus, the main result is unchanged and the effect of decreasing the fraction of students constrained increases the non-elite cutoff point, $\frac{da_{M_H}^{Q_L}}{dp(a)} > 0$.

B.3 Violation of the Upper Bound on Effort Cost

The separating equilibrium defined by (2.3)-(2.4) requires that the equilibrium cutoff for choosing a STEM major at non-elite schools is below the elite college admissions cutoff point, $a_{M_H}^{Q_L} \leq a^{Q_H}$. This is guaranteed by an assumption on the upper bound of the cost of choosing STEM at non-elite schools; $C_{M_H}(a^{Q_H}, Q_L) \leq a^{Q_H} - \mathbb{E}[a|a < a^{Q_H}]$, such that it is optimal to choose a STEM major for at least the most able student who is not eligible to attend an elite college. If this assumption does not hold and $a_{M_H}^{Q_L} > a^{Q_H}$ then the resulting separating equilibrium must satisfy the following conditions:

$$\mathbb{E}[a|a^{Q_H} \leq a < a_{M_H}^{Q_H}] = \mathbb{E}[a|a \geq a_{M_H}^{Q_H}] - C_{M_H}(a_{M_H}^{Q_H}, Q_H), \tag{B.6}$$

$$\phi(a) = \mathbb{E}[a|a \geq a_{M_H}^{Q_L}] - C_{M_H}(a_{M_H}^{Q_L}, Q_L), \tag{B.7}$$

where

$$\phi(a) = \frac{F(a^{Q_H})\mathbb{E}[a|a < a^{Q_H}] + (1-p)[F(a_{M_H}^{Q_L}) - F(a^{Q_H})]\mathbb{E}[a|a^{Q_H} \leq a < a_{M_H}^{Q_L}]}{F(a^{Q_H}) + (1-p)[F(a_{M_H}^{Q_L}) - F(a^{Q_H})]}. \quad \text{(B.8)}$$

Here $\phi(a)$ is the expected ability of individuals at non-elite colleges who choose non-STEM majors. This is a weighted average of the ability of all individuals who are below the admission cutoff, $a_i < a^{Q_H}$, and the individuals who are directly constrained into attending a non-elite college and choose a non-STEM major, $a^{Q_H} \leq a_i < a_{M_H}^{Q_L}$.

The effect of decreasing the fraction of constrained students can be found by taking the total derivative of (B.7) and (B.8),

$$\frac{da_{M_H}^{Q_L}}{dp} = \frac{-\partial\phi(a)/\partial p}{\partial\phi(a)/\partial a_{M_H}^{Q_L} - \partial\mathbb{E}[a|a > a_{M_H}^{Q_L}]/\partial a_{M_H}^{Q_L} + \partial C_{M_H}(a_{M_H}^{Q_L}, Q_L)/\partial a_{M_H}^{Q_L}}. \quad \text{(B.9)}$$

The quantity in the denominator, $\partial\phi(a)/\partial a_{M_H}^{Q_L} - \partial\mathbb{E}[a|a > a_{M_H}^{Q_L}]/\partial a_{M_H}^{Q_L} + \partial C_{M_H}(a_{M_H}^{Q_L}, Q_L)/\partial a_{M_H}^{Q_L}$, is negative under local stability of the equilibrium. The quantity in the numerator, $-\partial\phi(a)/\partial p$ is the reverse of the direct effect of decreasing the fraction constrained on the expected ability of STEM majors at non-elite schools,

$$-\frac{\partial\phi(a)}{\partial p} = -\frac{[F(a_{M_H}^{Q_L}) - F(a^{Q_H})]F(a^{Q_H})}{F(a^{Q_H}) + (1-p)[F(a_{M_H}^{Q_L}) - F(a^{Q_H})]}(\mathbb{E}[a|a < a^{Q_H}] - \mathbb{E}[a|a^{Q_H} < a < a_{M_H}^{Q_L}]). \quad \text{(B.10)}$$

The difference in expected abilities, $\mathbb{E}[a|a < a^{Q_H}] - \mathbb{E}[a|a^{Q_H} < a < a_{M_H}^{Q_L}]$, is clearly negative, making the above equation positive, $-\frac{\partial\phi(a)}{\partial p} > 0$. Therefore, removing the upper bound on the non-elite cost of choosing a STEM major reverses the effect of shifting the constraint on the non-elite cutoff point.

If $a_{M_H}^{Q_L} > a^{Q_H}$ then the effect of increased access to elite colleges and a smaller fraction of students who are directly constrained is a *decrease* in the non-elite major cutoff point,

$\frac{da_{M_H}^{Q_L}}{dp} < 0$. While this scenario seems unlikely given the empirical evidence (it would imply that only very high-ability students at non-elite schools participate in STEM majors),[1] it can only cause a false rejection of the asymmetric information model predictions if I do not observe a positive effect on the non-elite cutoff point in high access regions.

B.4 Differential College Admissions

Here I consider the possibility that the admissions cutoff may reflect differences in the constraint, p. In regions where the fraction of directly constrained students is high, administrators at elite colleges may respond by accepting additional students with lower ability who are not constrained so that the admissions cutoff decreases to a lower point, $\tilde{a}^{Q_H}$, (shown in Figure B.3 for a uniform ability distribution).

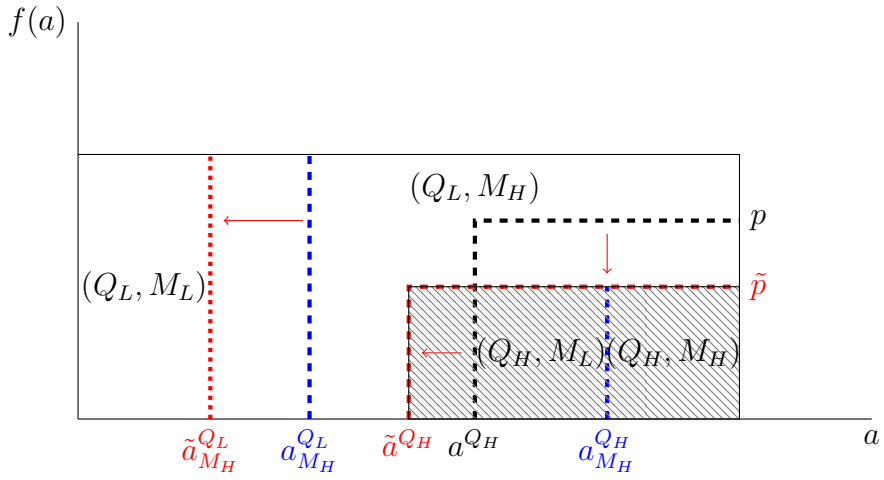


Figure B.3: Separating equilibrium for uniform ability with a shift in admissions standards

Taking the total derivative of (2.3) and (2.4) yields the shift in the non-elite major

[1] Using the data and definitions described in Section 2.4, I find that this type of separating equilibrium is not consistent with the raw data on SAT scores. The average SAT scores are: 1,318 for STEM majors at elite schools, 1,264 for non-STEM majors at elite schools, 1,123 for STEM majors at non-elite schools, and 1,033 for non-STEM majors at non-elite schools.

cutoff,

$$\frac{da_{M_H}^{Q_L}}{dp} = \frac{\partial\psi(a)/\partial p + (\partial\psi(a)/\partial a^{Q_H})(da^{Q_H}/dp)}{\partial\mathbb{E}[a|a < a_{M_H}^{Q_L}]/\partial a_{M_H}^{Q_L} - \partial\psi(a)/\partial a_{M_H}^{Q_L} + \partial C_{M_H}(a_{M_H}^{Q_L}, Q_L)/\partial a_{M_H}^{Q_L}}, \tag{B.11}$$

which now involves an additional term, $(\partial\psi(a)/\partial a^{Q_H})(da^{Q_H}/dp)$. This term depends on the direct effect on $\psi(a)$ from shifting the admissions cutoff. The direction of this effect depends on the relative positions of the expected ability of STEM majors at non-elite colleges, $\psi(a)$, and the admissions cutoff before the shift, a^{Q_H}. If $\psi(a) < a^{Q_H}$, then an increase in the admissions cutoff will add relatively high ability students to the STEM major group at non-elite colleges and $\partial\psi(a)/\partial a^{Q_H} > 0$. However, if $\psi(a) > a^{Q_H}$, then increasing the admissions cutoff will add relatively low ability students to the STEM major group at non-elite colleges and $\partial\psi(a)/\partial a^{Q_H} < 0$. The relative positions of $\psi(a)$ and a^{Q_H} will depend on both the constraint, p, and the position of the admissions cutoff within the ability distribution. Therefore, it is unclear whether this admissions response will magnify or mitigate the shift in the non-elite major cutoff due to the change in the fraction of constrained students. Importantly, this change in the admissions cutoff will have no effect on the predictions of the full information model (no shift in the non-elite major cutoff) so it can only cause a false rejection of the asymmetric information model predictions if I do not observe a positive effect on the non-elite cutoff point in high access regions.

Bibliography

(2006). Sleep in america. Summary of survey findings, National Sleep Foundation.

(2014). School start times for adolescents. Policy statement, American Academy of Pediatrics.

Altonji, J. and Pierret, C. (1997). Employer learning and the signaling value of education. In I. Ohashi, T. T., editor, *Industrial relations, incentives and employment*, pages 159–195. Macmillan Press Ltd, London.

Altonji, J. G. and Pierret, C. R. (2001). Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

Arcidiacono, P. (2004). Ability sorting and the returns to college major. *Journal of Econometrics*, 121(12):343 – 375.

Bedard, K. (2001). Human capital versus signaling models: University access and high school dropouts. *Journal of Political Economy*, 109(4):749–775.

Bettinger, E. (2010). To be or not to be: Major choices in budding scientists. In *American Universities in a Global Market*, NBER Chapters, pages 69–98. National Bureau of Economic Research, Inc.

Black, D. A., Sanders, S., and Taylor, L. (2003). The economic reward for studying economics. *Economic Inquiry*, 41(3):365–377.

Bloom, D., Canning, D., and Sevilla, J. (2003). Geography and poverty traps. *Journal of Economic Growth*, 8:355–378.

Bostwick, V. (2016). Signaling in higher education: The effect of access to elite colleges on choice of major. *Economic Inquiry*, 54(3):1383–1401.

Bostwick, V. and Steigerwald, D. (2012). Obtaining critical values for test of markov regime switching. Economics working paper series, University of California Santa Barbara.

Bostwick, V. K. and Steigerwald, D. G. (2014). Obtaining critical values for test of markov regime switching. *Stata Journal*, 14(3):481–498(18).

Carrell, S. E., Maghakian, T., and West, J. E. (2011). A's from zzzz's? the causal effect of school start time on the academic achievement of adolescents. *American Economic Journal: Economic Policy*, 3(3):62–81.

Carskadon, M. A., Wolfson, A. R., Acebo, C., Tzischinsky, O., and Seifer, R. (1998). Adolescent sleep patterns, circadian timing, and sleepiness at a transition to early school days. *SLEEP-NEW YORK*, 21:871–881.

Carter, A. and Steigerwald, D. (2012). Testing for regime switching: A comment. *Econometrica*, 80(4):1809–1812.

Carter, A. and Steigerwald, D. (2013). Markov regime-switching tests: Asymptotic critical values. *Journal of Econometric Methods*, 2(1):25–34.

Cho, J. and White, H. (2007). Testing for regime switching. *Econometrica*, 75:1671–1720.

Clark, D. and Martorell, P. (2014). The signaling value of a high school diploma. *Journal of Political Economy*, 122(2):pp. 282–318.

Danner, F. and Phillips, B. (2008). Adolescent sleep, school start times, and teen motor vehicle crashes. *Journal of Clinical Sleep Medicine*, 4(6):533–535.

Daymonti, T. N. and Andrisani, P. J. (1984). Job preferences, college major, and the gender gap in earnings. *Journal of Human Resources*, 19(3):408–428.

Do, C. (2004). The effects of local colleges on the quality of college attended. *Economics of Education Review*, 23(3):249 – 257.

Donald, S. G. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics*, 89(2):221–233.

Dorans, N. (1999). Correspondences between act and sat i scores. Report 99-1, College Entrance Examination Board.

Durmer, J. S. and Dinges, D. F. (2005). Neurocognitive consequences of sleep deprivation. *Seminars in neurology*, 25(1).

Edwards, F. (2012). Early to rise? the effect of daily start times on academic performance. *Economics of Education Review*, 31(6):970–983.

Fang, H. (2006). Disentangling the college wage premium: Estimating a model with endogenous education choices. *International Economic Review*, 47(4):1151–1185.

Fischer, S. (2015). The downside of good peers: How classroom composition differentially affects men's and women's stem persistence. Technical report, University of California, Santa Barbara.

Gislason, T., Tomasson, K., Reynisdottir, H., Bjornsson, J. K., and Kristbjarnarson, H. (1997). Medical risk factors amongst drivers in single-car accidents. *Journal of Internal Medicine*, 241(3):217–223.

Griffith, A. L. and Rothstein, D. S. (2009). Cant get there from here: The decision to apply to a selective college. *Economics of Education Review*, 28(5):620 – 628.

Grogger, J. and Eide, E. (1995). Changes in college skills and the rise in the college wage premium. *The Journal of Human Resources*, 30(2):pp. 280–310.

Haraldsson, P. O., Carenfelt, C., Diderichsen, F., Nygren, A., and Tingvall, C. (1990). Clinical symptoms of sleep apnea syndrome and automobile accidents. *ORL*, 52:57–62.

Hausman, J., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application to the patents-r & d relationship. *Econometrica*, 52(4):pp. 909–938.

Hinrichs, P. (2011). When the Bell Tolls: The Effects of School Starting Times on Academic Achievement. *Education Finance and Policy*, 6(4):486–507.

Hussey, A. (2012). Human capital augmentation versus the signaling value of mba education. *Economics of Education Review*.

James, E., Alsalam, N., Conaty, J. C., and To, D.-L. (1989). College quality and future earnings: Where should you send your child to college? *American Economic Review*, 79(2):pp. 247–252.

Jenni, O., Achermann, P., and Carskadon, M. (2005). Homeostatic sleep regulation in adolescents. *SLEEP*, 28(11):1446–1454.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413.

Lang, K. and Kropp, D. (1986). Human capital versus sorting: The effects of compulsory attendance laws. *The Quarterly Journal of Economics*, 101(3):609–624.

Luppino, M. and Sander, R. (2015). College major peer effects and attrition from the sciences. *IZA Journal of Labor Economics*, 4(4).

Martiniuk, A., Senserrick, T., Lo, S., and et. al. (2013). Sleep-deprived young drivers and the risk for crash: The drive prospective cohort study. *JAMA Pediatrics*, 167(7):647–655.

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*, 29(6):923 – 934.

Paglin, M. and Rufolo, A. M. (1990). Heterogeneous human capital, occupational choice, and male-female earnings differences. *Journal of Labor Economics*, 8(1):pp. 123–144.

Riley, J. G. (1979). Testing the educational screening hypothesis. *Journal of Political Economy*, 87(5):S227–S252.

Shults, R. A., Olsen, E., and Williams, A. F. (2015). Driving among high school students united states, 2013. *Morbidity and Mortality Weekly Report*, 64(12).

Smith, A. (2015). Spring forward at your own risk: Daylight saving time and fatal vehicle crashes. *American Economic Journal: Applied Economics*, 8(2):65–91.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374.

Tyler, J. H., Murnane, R. J., and Willett, J. B. (2000). Estimating the labor market signaling value of the ged. *The Quarterly Journal of Economics*, 115(2):431–468.

Vorona, R. D., Szklo-Coxe, M., Wu, A., Dubik, M., Zhao, Y., and Ware, J. (2011). Dissimilar teen crash rates in two neighboring southeastern virginia cities with different high school start times. *Journal of Clinical Sleep Medicine*, 7(2):145–151.

Wahlstrom, K. (2002). Changing times: Findings from the first longitudinal study of later high school start times. *NASSP Bulletin*, 86(633).

Wahlstrom, K., Dretzke, B., Gordon, M., Peterson, K., Edwards, K., and Gdula, J. (2014). Examining the impact of later high school start times on the health and academic performance of high school students: A multi-site study. Technical report, Center for Applied Research and Educational Improvement. St Paul, MN: University of Minnesota.

Winston, F. e. a. (2007). Driving: Through the eyes of teens. Technical report, The Childrens Hospital of Philadelphia.

Wolpin, K. I. (1977). Education and screening. *American Economic Review*, 67(5):949–958.

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, 90(1):77 – 97.