Three Essays on the Economics of Education

by

Joshua Milton Hyman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Public Policy and Economics)
in the University of Michigan
2013

Doctoral Committee:

Professor Susan M. Dynarski, Chair
Professor John Bound
Professor Brian A. Jacob
Professor Jeffrey A. Smith

DEDICATION

I dedicate this dissertation to my loving wife, Caroline Theoharides.

ACKNOWLEDGEMENTS

I am grateful to the members of my dissertation committee – Susan Dynarski, Brian Jacob, Jeff Smith, and John Bound – for being excellent teachers and advisors. John Bound provided thoughtful and helpful comments. Jeff Smith met with me frequently, and read several drafts, providing detailed suggestions after each read. Brian Jacob encouraged me to apply to the Ford School, and provided mentorship and support throughout my many years. Susan Dynarski made me the researcher and scholar that I am today. Through collaboration she taught me the research process, through her emotional support she taught me to be confident in my work, and through her feedback and advice she was critical in the completion of my dissertation.

Mike Gideon has been with me from the first day of graduate school, and I am indebted to him for his friendship, encouragement, and valuable feedback. Thanks to Mary Corcoran for her support during the first year of the program. I am grateful for helpful conversations with Charlie Brown, Sebastian Calonico, Steve DesJardins, John DiNardo, Tom Downes, Rob Garlick, Andrew Goodman-Bacon, Steve Hemelt, Kevin Stange, Lesley Turner, Elias Walsh, and seminar participants at the University of Michigan.

Thanks to ACT Inc. and the College Board for the use of their data in Chapters 1 and 3. In particular, I thank Ty Cruce, John Carroll, and Julie Noble at ACT Inc. and Sherby Jean-Leger at the College Board. I am grateful for data and assistance from my partners at the Michigan Department of Education, Center for Educational Performance and Information, and Michigan Consortium for Educational Research. Specifically, I thank Ken Frank, Tom Howell, Venessa Keesler, Joseph Martineau, and Barbara Schneider. The research reported in Chapters 1 and 3 was supported by the Institute for Education Sciences, U.S. Department of Education, through Grant R305E100008. I thank Jayne Zaharias-Boyd of HEROS and the Tennessee Department of Education for allowing the match between the STAR and National Student Clearinghouse data for Chapter 2. The Education Research Section at Princeton University generously covered the cost of this match.

My parents, Arthur and Sherry Hyman, have been unwavering sources of support during the writing of this dissertation, as they have been my entire life. My brother, Daniel Hyman, encouraged me throughout graduate school, and in particular during the difficult first year of

classes. Finally, my wife, Caroline Theoharides, has been there for me every day, editing my paper drafts and presentation slides, and providing me with thoughtful comments and suggestions. This dissertation would not have been finished if not for her encouragement, support, and understanding.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

ACT for All: The Effect of Mandatory College Entrance Exams on Postsecondary Attainment and Choice

Abstract

Recent experimental studies of low-cost interventions that reduce administrative or informational barriers to college enrollment have shown cost-effective increases in educational attainment. Yet it remains to be seen whether these interventions can be implemented effectively at a large scale. Since 2001, eleven states have incorporated the ACT or SAT into their eleventh grade standardized testing regime, thus requiring and paying for all students to take a college entrance exam. I examine the effect of this inexpensive and large-scale policy on postsecondary enrollment, persistence, and choice. I first exploit the implementation of the reform to measure the *pre-policy* supply of college-ready students not taking a college entrance exam. I show that for every ten poor students taking the ACT or SAT pre-policy and scoring at a college-ready level, there are five additional poor students taking neither exam who *would* score at a college-ready level. I then compare changes in college-going from before to after the policy among students at high schools without an ACT test center pre-policy relative to students at schools with a pre-policy center. This strategy exploits the lower baseline ACT-taking, and thus larger dosage of the policy at schools without a pre-policy center. I find that the policy increases statewide enrollment at four-year institutions by 0.6 percentage points (2%), with effects concentrated among students with a low-to-mid-level predicted probability of taking the ACT in the absence of the policy. Among students enrolled in the poorest third of high schools, the effect is 1.3 percentage points (6%). I find similar effects on enrolling through the fourth year of college, suggesting that marginal students persist.

1.1. Introduction

Inequality in educational attainment has widened substantially over the past few decades. Not only do minority and low-income students attend college in lower proportions than their majority and higher-income counterparts, but conditional on enrolling, these students are also less likely to persist through college and complete a degree (Bailey and Dynarski, 2011). While certainly not every low-income and minority student would benefit from postsecondary education, recent research suggests that a non-trivial number of high achieving, disadvantaged students either do not attend college, or attend a less selective school than they could (Dillon and Smith, 2013; Avery and Hoxby, 2012; Bowen, Chingos, and McPherson, 2009; Pallais and

Turner, 2006). Policies that induce low-income students to attend and persist at appropriately selective institutions could have substantial implications for reducing educational inequality.

Many policies and interventions aim to increase the educational attainment of disadvantaged students. Policies such as Head Start and class size reduction, which aim to increase the human capital of students, as well as policies such as student aid, that reduce the cost of college, have all been shown to successfully increase postsecondary attainment (Deming, 2009; Dynarski, Hyman, and Schanzenbach, forthcoming; Deming and Dynarski, 2010). However, these policies are all quite expensive, costing tens of thousands of dollars to induce one additional student to enroll in college (Dynarski, Hyman, and Schanzenbach, forthcoming). Recently, interventions aimed at reducing informational and administrative barriers to college enrollment have found large effects at a fraction of the cost of the more traditional tools mentioned above (Hoxby and Turner, 2013; Bettinger, et al., 2012; Carrell and Sacerdote, 2013). It remains to be seen whether these low-cost policies can be implemented effectively at scale.

In this paper, I examine the impacts of an inexpensive policy aimed at boosting postsecondary attainment that is currently operating at scale. Eleven states require, and pay for, college entrance exams (i.e., the ACT or SAT) for all public school eleventh graders. Given that it costs less than $50 per student for states to implement this policy,[1] very small effects on college-going would suffice for the policy to be as cost-effective as traditional student aid. This paper answers the question: What is the effect of mandatory college entrance exams on postsecondary enrollment, persistence, and choice? I use a new longitudinal data set provided by the Michigan Consortium for Educational Research (MCER), containing all Michigan public high school students. The data include demographics, eighth and eleventh grade statewide assessment scores, and information on postsecondary enrollment. I merge in ACT and SAT scores for all test-takers during the sample period.

As a first step toward predicting whether we might expect to observe postsecondary effects, I use the post-policy ACT score distribution to deduce what fraction of pre-policy non-takers would score at a college-ready level if they took the exam. I show that for every ten poor students taking a college entrance exam and scoring college-ready, there are an additional five poor students who do not take the test but who would score college-ready if they did. This represents a contribution to the emerging literature on "undermatch." Avery and Hoxby (2012)

[1] From author's communications with state departments of education.

focus on the supply of disadvantaged students who take a college entrance exam and score in the top ten percent of all ACT- or SAT-takers, but do not apply to selective colleges. I use a lower threshold of "high-achieving," and look back further in the pipeline of the college application process, to show that there is a large supply of disadvantaged students who would score well enough to enroll in a selective four-year college, but who are dropping out of the application process prior to even taking a college-entrance exam.

To examine the effects of the mandatory ACT policy on postsecondary outcomes, I acquire an historic list of all ACT test centers in Michigan. I compare changes in college-going from before to after the implementation of the policy for students in schools without a test center in the pre-policy period relative to students in schools that had a test center. This exploits the fact that schools without a test center pre-policy had lower test-taking rates and thus experience a larger treatment dosage from the policy. I use propensity score matching to restrict my analysis to a sample of schools with and without a test center that have similar observed characteristics. My results are robust to a number of specification checks, including the creation of treatment and comparison groups based on driving distance from a student's home to the nearest pre-policy test center, rather than the school-level test center measure.

I find that there is a 0.6 percentage point (2%) increase in the statewide four-year college enrollment rate due to the policy. This is the first credible estimate of the postsecondary enrollment effects of the mandatory college entrance exam policy. Although two recent studies have provided estimates on enrollment effects (Goodman, 2012; Klasik, forthcoming), these results are not robust to basic specification changes. Goodman (2012) uses aggregate institution-year level data from the Integrated Postsecondary Education Data System (IPEDS) and finds much larger enrollment effects than does the present paper. In Appendix 1.1, I reconstruct her data set and replicate her results, showing that when I use either an arguably more valid set of control variables, or a more accurate college enrollment measure as the dependent variable, her results become substantially attenuated and statistically indistinguishable from zero. [2]

I find that the overall enrollment effects of the policy mask important heterogeneity, with larger effects (1.3 points, 5%) for students with a low-to-mid-level ex-ante probability of taking

[2] Klasik (forthcoming) uses similar data and methods as Goodman (2012). Klasik's results are larger than Goodman's (2012), are statistically imprecise, and vary substantially across treatment state and specification. In similar analyses (available upon request) I use the Current Population Survey (CPS), American Community Survey (ACS), and National Longitudinal Survey of Youth (NLSY) 1997 restricted access geocode data finding similarly imprecise results that vary dramatically across treated state, data source, and specification with no clear pattern.

the ACT in the absence of the policy. Effects are also larger among males (0.9 points, 3%), poor students (1.0 points, 6%), and students at schools with a high poverty share (1.3 points, 6%). Because the two aforementioned studies use aggregate state-level data, they cannot estimate heterogeneity in effects of the policy by student or school characteristics. By using micro-data, I am able to show that this policy is in fact effective at reducing inequality, with effects on college enrollment concentrated among economically disadvantaged students and poor schools.

Given that educational inequality is also pervasive in terms of persistence (Bailey and Dynarski, 2011), to accurately assess the effectiveness of the mandatory ACT policy, it is crucial to test if the marginal student induced into college by the policy actually persists through college. Because my data follow students over time, my study is the first to estimate persistence through college as a result of the policy. I find that the effects on enrolling and spending four years in college are similar to the enrollment effects, implying that marginal enrollees persist.

The remainder of this paper is structured as follows: In section 1.2, I discuss the mandatory college entrance exam policy and costs associated with taking the test. Section 1.3 describes the data used in my analysis. Section 1.4 examines the extent of college-ready students not taking a college entrance exam pre-policy. Section 1.5 examines the effects of the policy on college enrollment, persistence, and choice. Section 1.6 discusses the interpretation of my results in light of possible supply-side capacity constraints, and the fact that I estimate a local average treatment effect (LATE). Finally, Section 1.7 concludes with a comparison of the cost and benefit of mandatory college entrance exams to other policies that boost postsecondary attainment.

1.2. Costs, Information, and Mandatory College Entrance Exams

The ACT and SAT are college admission exams required for admission to nearly all four-year institutions across the country.[3] Historically, these exams have been taken almost exclusively by students considering applying to a four-year institution. However, since 2001, eleven states have implemented free and mandatory college entrance exams for all high school juniors.[4] Table 1.1 lists the states that have adopted this policy, which exam they use (ACT or

[3] Exceptions are primarily for-profit institutions, specialty or religious institutions, and institutions that admit all or nearly all applicants.

[4] Several more states are in the process of passing legislative bills to implement the reform or are implementing the reform in a pilot phase in a handful of districts.

SAT), and the year that the first eleventh grade cohort was (or will be) exposed to the policy. Nearly all states have adopted the ACT rather than the SAT.

The state-mandated ACT and SAT are the official exams used for college admission purposes. In the absence of the statewide policy, the ACT and SAT are offered on Saturday mornings, cost students between $30 and $50, and require students to travel to the nearest test center.[5] State-mandated exams are given during the school day at no financial cost to the student, and are almost universally at a student's high school. As with the standard ACT and SAT, students can select colleges to which they send their scores. Students are mailed their own official score report several weeks after they take the exam.

Speculating about how mandatory college entrance exams could have an effect on the college enrollment decisions of students requires a discussion of the forces behind the growing inequality in educational attainment. There are (at least) three primary hypotheses for why disadvantaged students enroll in college at lower rates than their more advantaged peers. The first two, credit constraints and poor academic preparation, are well-studied, and can explain much but not all of observed educational inequality. One less well understood hypothesis involves administrative barriers and information constraints: low-income students and their parents may not fully understand or be aware of the benefits and accessibility of higher education, and also may not be able to navigate the college and financial aid application process. While mandatory college entrance exams have the potential to impact college enrollment decisions through any of these channels, the reduction of administrative barriers and informational constraints arguably seems most relevant, as discussed below.

Recent research has shown that small changes to the structure of choice-making, such as changes in the default choice, can have large behavioral effects in various policy domains like retirement savings plans (Beshears et al., 2009; Madrian and Shea, 2001). Similarly, a small change to the structure of the college entrance exam score report sending process was shown to have large effects on the number of score reports students sent (Pallais, 2011). Requiring all students to take a college entrance exam is a substantial change to the structure of the four-year college application process, but given the remaining large obstacles to enrolling at a four-year school, it is ambiguous whether this "nudge" should translate into increased enrollment.

[5] Fee waivers are available for low-income students but take-up is low, perhaps because it requires paperwork on the part of the student and coordination with high school counselors.

Mandatory college entrance exams reduce the monetary, psychic, and time cost of applying to college. While spending $30 to $50 and five hours on a Saturday represents a very small share of the overall cost of applying to and attending college, these monetary and time costs can represent a real hurdle to many low-income students, particularly if taking the test requires seeking time off from employment. Further, approximately half of public school students do not attend a high school with a test center in the school, which means that they would have to find and travel to the nearest test center. Offering the exam for free on a school day reduces or eliminates these costs to the student.[6]

Mandatory college entrance exams also have the potential to alleviate information constraints in the college application process. Students taking the ACT or SAT may learn about college accessibility, as after the test they receive mailings from postsecondary institutions across the country. Test-takers may also learn about their college-going ability. The score on these tests provide students with a signal of their likelihood of being admitted to, and succeeding at, a four-year college or university. Under a policy of mandatory college entrance exams, all students receive this signal of their college ability, whether or not they were planning to apply to a four-year institution.[7]

Finally, mandatory college entrance exams may also increase information about the college application process by altering behavior at the school level. In many mandatory college entrance exam states, Adequate Yearly Progress (AYP) of the school depends on how students perform on the exam. Consequently, teachers, school administrators, and guidance counselors have an incentive to educate students about the ACT and the SAT, and why they are important,

[6] A recent working paper isolates the cost of taking a college entrance exam in a nearby location, as opposed to one's high school. Bulman (2013) finds that the opening of an SAT test center in a school has large effects on SAT-taking, and on educational attainment. This suggests that the binding cost or barrier faced by a marginal student deciding whether to take a college entrance exam is not the direct financial or time cost of taking the test itself, but rather the cost or lack of information due to not having a test center in one's school. That paper also examines the effects in three school districts (Stockton, CA, Palm Beach, FL, and Irving, TX) of implementing a free SAT. He finds four-year enrollment effects of the policies on the order of 15 percent. While these effects are substantially larger than those I estimate, a single district in the state offering the SAT for free is quite a different policy than a statewide implementation of a mandatory exam.

[7] The literature is mixed on whether this signal should affect behavior. Stinebrickner and Stinebrickner (2012) show that learning about ability plays a central role in the college drop-out decision. In the secondary school context, Jacob and Wilder (2010) show that students do update their expectations based on the acquisition of new information about their ability, but that over four-fifths of high school seniors think that they will acquire a bachelor's degree (an extreme overestimate). The fact that high school students tend to overestimate the likelihood that they will earn a degree suggests that new information from ACT scores would not be an expected channel through which the mandatory ACT policy would boost college enrollment.

in order to boost student efforts. In practice, most schools have at least some resources available to help students prepare for the tests, while some schools with greater resources offer test preparation sessions, or entire classes devoted to preparing for the exams.[8] More broadly, this policy has the potential to increase the college-going culture at the school, which has been shown to be an important instrument in increasing the postsecondary attainment of disadvantaged students (Jackson, 2010).

1.3. Data

For the majority of analyses in this paper, I use a new data set containing all students attending Michigan public high schools in six recent eleventh grade cohorts (2003–2004 through 2008–2009).[9] The data contains time-invariant demographic information such as sex, race, and date of birth, as well as time-varying characteristics such as free and reduced-price lunch status, limited-English-proficiency (LEP) status, special education (SPED) status, and student's home address. The data also contain eighth grade and eleventh grade state-assessment results. For the cohorts of students after implementation of the mandatory ACT exam, the eleventh grade assessment results include ACT scores.

Student level postsecondary enrollment information is obtained by matching students to the National Student Clearinghouse (NSC). The NSC is a non-profit organization that houses postsecondary enrollment information on over ninety percent of undergraduate enrollment nationwide. Matching is completed using name and date of birth.[10] School- and district-year level characteristics from the Common Core of Data (CCD) are merged to the dataset based on where and when students are enrolled in high school.

I have acquired and merged on several other key pieces of information. First, using student name, date of birth, sex, race, and eleventh grade home zip code, I matched the student-level Michigan data to micro-data from ACT Inc. and the College Board on every ACT-taker and SAT-taker in Michigan over the sample period. This allows me to observe ACT-takers before the policy, as well as students who took the SAT instead of the ACT pre-policy. I also acquired from ACT Inc. a list of all ACT test centers in Michigan over the sample period, including their

[8] From author's discussions with high school guidance counselors.
[9] These data were provided by the Michigan Department of Education, Center for Educational Performance and Information, and Michigan Consortium for Educational Research.
[10] See Dynarski, Hemelt, and Hyman (2013) for a detailed discussion of the NSC matching process and coverage rates.

addresses and their open and close dates. I geocoded student home addresses during eleventh grade, and the addresses of these test centers, to calculate the driving distance from the student's eleventh grade home to the nearest ACT test center.[11]

Table 1.2 shows sample means before and after implementation of the mandatory ACT. I condition my sample on reaching the spring semester of eleventh grade, which is the semester when the eleventh grade state assessment is given. Michigan was hit hard by the economic recession during the sample period: the percentage of eleventh graders eligible for free lunch rose from 24% to 32%, while the percentage that are black increased from 15.5% to 18%. The local unemployment rate rose from 7.3% to 9.1%.[12] Educational attainment was fairly stable over the period, with high school graduation at 84.4% and college enrollment increasing slightly from 57% to 59%.[13]

Prior to the mandatory ACT policy, 56% of students took the ACT. The percentage increased to 91% after the policy.[14] ACT-taking rates tend to increase more for those groups of students who have lower rates prior to the policy. This is particularly pronounced among students eligible for free or reduced-price lunch, whose rate of ACT-taking more than doubles from 35% to 85%.[15]

In Michigan, 6.4% of students take the SAT before the mandatory ACT policy is implemented. Interestingly, this drops nearly in half to 3.3% after the policy.[16] Nearly all of the students taking the SAT pre-policy (5.8 of the 6.4%) also take the ACT. Unsurprisingly, all

[11] In the rare case when an eleventh grade home address is missing, I use the home address during the surrounding grades. For 2 percent of the sample, the address is missing during all grades in high school. In these instances, a distance to the nearest ACT center is not assigned.

[12] Unemployment rates at the city (when available) or county level are from the Bureau of Labor Statistics.

[13] I define a student as enrolling in college if he or she enrolls before October 1st of the second fall following on-time high school graduation. This definition ensures that the measure is consistent across cohorts as I do not observe more than two years of enrollment for the most recent cohort. This variable can be thought of as a liberal measure on-time college enrollment that captures students graduating high school on time and taking a gap year before enrolling, or students who take an extra year to graduate high school and then enroll the following fall.

[14] The reason for the 9 percentage points of non-compliance is primarily drop-out. Of the 91% of students in my sample who reach 12th grade, the test-taking rate is 95%. Of those who graduate high school, the test-taking rate is 97.8%. The remaining non-compliance is mostly due to students taking the special education version of the eleventh grade test that does not include the ACT. Of the 80% of the sample who are high school graduates, and who do not take the special education version of the test, the fraction with a valid ACT score is 98.9%. While 95% of students in each school are required to take the eleventh grade assessment for NCLB purposes, it is not technically a graduation requirement, hence this small remaining gap.

[15] The exception to this pattern is black students, who have a lower test-taking rate than whites pre-policy, but whose test-taking rate increases by a similar number of percentage points. This is due to the large proportion of black students who drop out of high school without taking the eleventh grade assessment.

[16] An examination of the percentage by cohort reveals that the drop occurs suddenly at the time of the reform and is not due to a negative pre-trend.

students who take the SAT post-policy also take the ACT. Finally, students taking the SAT in Michigan are quite high achieving, with a mean score converted to the ACT metric of 24.7 (out of 36) pre-policy, relative to the mean of 20.7 among ACT-takers.

In some analyses, I use ACT micro-data from several other states. I acquired from ACT Inc. a one-in-four sample of all ACT test-takers scheduled to graduate high school during 1999–2011 in the four states that implemented mandatory ACT-taking prior to 2009, and in fourteen nearby comparison states.[17] Before acquiring the data, I chose the comparison states using a propensity score estimation procedure, and pre-treatment state-year level covariates. I do not describe the data in detail or provide summary statistics, because I only use these data briefly in this paper. A more detailed description of this data is available from the author upon request.

1.4. Selection into College Entrance Exam-Taking

In this section, I use the exogenous increase in college entrance exam-taking due to the mandatory ACT policy, and the resulting nearly complete distribution of scores, to examine selection into college entrance exam-taking.

1.4.1. The Supply of High-Achieving Non-Takers

I begin my analysis of selection into college entrance exam-taking in Michigan by predicting the ACT score distribution that would be observed among non-takers during the pre-policy period if they were to take the ACT. I do this by subtracting the number of test-takers scoring at each ACT score during the pre-policy period from the number scoring at each score in the post period, when nearly all students take the test. Students who take the SAT but not the ACT pre-policy are included in all of my analyses, which prevents inaccurate categorization of students in the pre-policy period who took the SAT instead of the ACT, as non-takers.[18] Later in this section I describe the sensitivity of the results to omitting the SAT data.

Figure 1.1a shows this exercise graphically: the dashed line plots the frequency distribution of scores pre-policy, which appears to be approximately normal with a slightly thicker right tail. The solid line shows the distribution of scores post-policy, which is larger because there are many more test-takers, and is substantially skewed to the left, reflecting the

[17] The treatment states are Colorado, Illinois, Michigan, and Kentucky. The comparison states are Alabama, Arkansas, Idaho, Iowa, Ohio, Oklahoma, Minnesota, Mississippi, Missouri, Montana, New Mexico, Utah, West Virginia, and Wisconsin.

[18] For students taking the ACT multiple times, I use their first score. For students taking the SAT but not the ACT, I include their SAT score scaled to the ACT metric. For students taking both tests, I use their first ACT score.

lower average scores of students induced into test-taking. The dotted line shows the difference between these two lines, and represents the predicted score distribution of all pre-policy non-takers under the assumptions that (1) the average size of the cohorts is the same pre-and post-policy, (2) the composition of public school students and other factors in Michigan affecting ACT scores is stable over the sample period, and (3) all students take the ACT in the post period. As we have already seen, none of these assumptions is strictly true, so I adjust my procedure in a number of ways.

To ensure that the changing cohort size and composition of Michigan students is not leading to differences in the score distributions, I reweight the post-policy cohorts of students following DiNardo, Fortin, and Lemieux (DFL 1996) to resemble the pre-policy students according to their observed characteristics. Specifically, I estimate using OLS:

$$PRE_{isd} = \alpha + \beta_1 X_{isdt} + \beta_2 S_{st} + \beta_3 D_{dt} + \varepsilon_{isdt} \,, \qquad (1)$$

where PRE_{isd} is an indicator for student *i* in school *s* in district *d* being in the pre-policy period. *X* is a vector of individual-level covariates, *S* is a vector of school-year level covariates, and *D* is a vector of district-year level covariates.[19] I predict $\hat{PRE}_{isd}$, which is the propensity score of being in the pre-policy period. The DFL weight equals: $\frac{\hat{PRE}_{is}}{(1-\hat{PRE}_{is})}$, which I then censor at its 1st and 99th percentile.[20] Each pre-policy score receives a weight of 1, and each post-policy score receives its censored DFL weight. I normalize the DFL weights in the post-policy period to have a mean equal to 0.963, which is the proportional size of the three combined pre-policy cohorts relative to the three combined post-policy cohorts. To compute the distribution of latent scores, I sum the weights in the post period at each ACT score, and subtract the sum of the weights at each score in the pre period.

Figure 1.1b plots the re-weighted post-policy score distribution (the solid line). Assume that after DFL-reweighting the only difference between the pre- and post-policy cohorts is that nearly everyone takes the ACT in the post period. Then the difference in the number of students

[19] *X* includes LEP, SPED, free lunch status, race dummies, and sex. *S* includes fraction on free lunch, fraction black, number of eleventh graders, and pupil-teacher ratio. *D* includes the district-level versions of the variables in *S* plus student-guidance counselor ratio, dummies for urban / rural status, and the local unemployment rate. All interactions of student-level covariates with each other and with the school- and district-year level covariates are included. The R-Squared from the regression is 0.149.

[20] Results are not sensitive to whether or not the weights are censored.

scoring at each ACT score bin should reflect the distribution of unobserved latent scores of the students who did not take the exam before it was mandatory. Examining how the difference between the dashed and solid lines varies between Figures 1.1a and 1.1b provides a visual depiction of how adjusting for cohort composition and size affects the results. There is a small but noticeable increase in the gap between these two lines in the upper part of the distribution. Because the post-policy sample has a higher fraction minority and free-lunch, the DFL-reweighting weights white and non-free lunch students higher, slightly shifting the post-policy distribution upward. I adjust for cohort size and composition in the remainder of the analyses in Section 1.4.

While the latent scores of pre-policy non-takers (Figure 1.1b, dotted line) are generally lower than the scores of those taking the test (dashed line), there is a long right tail of students who do not take a college entrance exam, but would score well if they did. In the first two columns of Table 1.3, I show the mean, standard deviation, and various percentiles of the distribution of scores of takers and non-takers in the pre-policy period. The mean score among students taking the ACT or SAT in the pre period is 20.7, and the mean latent score among non-takers is 16.8.[21]

While the mean among non-takers is nearly four points lower (a highly statistically significant difference), there is a non-trivial fraction of non-takers with college-ready latent scores. I use a score of 20 as a threshold of college-readiness, which is the 25th percentile of all students in Michigan in the pre-policy sample who attend and graduate from a four-year post-secondary institution.[22] ACT Inc. cites a score of 20 as likely qualifying a student for admission to a "traditional" four-year institution. The choice of 20 reflects a threshold that represents students with a good chance of admittance to, and success at, a reasonably selective four-year institution. I also show the results as calculated using a threshold of 22, which the ACT cites as likely qualifying a student for admission to a selective four-year institution.[23] I show the sensitivity of the results to other score thresholds in section 1.4.4.

[21] A Kolmogorov-Smirnov nonparametric test of the equality of the distribution of the observed scores among takers and the latent scores among non-takers is rejected with a p-value of 0.000.

[22] It is also the 25th percentile of first-year students attending Central Michigan University and Western Michigan University, two selective, though not flagship, state universities.

[23] These figures are taken from ACT Inc. (2002). A score of 18–21 likely qualifies a student for admission to non-selective institutions, 20–23 to traditional institutions, 22–27 to selective institutions, and 27–31 (or higher) to highly selective institutions.

In Table 1.3, column 1, I show that almost 118,000 students, or 58% of ACT/SAT-takers pre-policy, score at or above 20. Almost 27,000 students not taking the exam, or 21%, would score at this level based on the distribution of latent scores. This means that if all students took the exam, we would see a 22.7% increase in the number of students scoring college ready (26,717 / 117,953). Put differently, for every 100 students taking the test and scoring college-ready, there exist another 23 students not taking the test who would score-college ready. I refer to this 0.227 as the "proportion of college-ready non-takers to takers." When I consider an ACT score threshold of 22 rather than 20, this proportion decreases somewhat to 0.192: for every 100 students who take the ACT or SAT and score at a level likely qualifying them for admission to a selective college, there exist another 19 students not taking either test who would score at this level.

I calculate standard errors for the proportion of college-ready non-takers to test-takers, and for the percent of test-takers and non-takers that are college ready. I compute these standard errors by running 200 bootstrapped replications of the above exercise and calculating the statistics after each replication. The standard deviation of the statistic across these replications is the estimated standard error of the statistic. I report the standard errors in Table 1.4. The standard errors for the 0.227 proportion, the 58.2 percent of test-takers that are college-ready, and the 21.3 percent of non-takers that are college-ready are 0.016, 0.9, and 1.5, respectively. Thus, the 95% confidence interval around the 0.227 proportion of college-ready non-takers to takers is between 0.196 and 0.259.[24]

1.4.2. Threats to Validity

The inclusion of SAT-takers in my analysis is important for the accuracy of the results, and has been omitted in a previous study conducting a similar exercise (Goodman, 2012). When I calculate the proportion of non-takers to takers, but include missing scores for students who took the SAT but not the ACT, I calculate a slightly higher proportion of 0.241. This small increase is a result of the 0.5% of eleventh graders in Michigan pre-policy who took the SAT but not the ACT. While the change from the 0.227 proportion to 0.241 is small, the fact that this

[24] Each bootstrapped replication resamples schools from the original data to allow for correlation of the error term within schools. The main assumption for the validity of the bootstrapped standard errors is that the original sample is representative of the population of interest. This is convincing because the sample is indeed the population of all Michigan public school students, which is the population of interest. See Efron and Tibshirani (1993) for details. Because the standard errors are more conservative, I conduct the bootstrapped replications after having already created the DFL weights using the original sample. Re-estimating the DFL regression after each of the bootstrap replications produces standard errors for the statistics reported above equal to 0.012, 0.9, and 1.0, respectively.

small upward bias is induced by only one half of one percent of students taking the SAT instead of the ACT, implies that the proportion of college-ready non-takers to takers is quite sensitive to observing SAT-taking behavior. This discrepancy is linearly scalable: if 5% of students in a state take the SAT and not the ACT, as opposed to 0.5%, then not observing SAT-takers would lead a researcher to overstate the proportion by 14 percentage points. The lack of SAT-takers could explain the larger results found by Goodman (2012), which focus on Colorado, in which a third of high school graduates took the SAT before the policy.

Non-compliance in the post-policy period is a remaining threat to the validity of this exercise, as not all students take the ACT. Consequently, the dotted line in Figure 1.1 actually shows the distribution of latent ACT scores of non-takers who would take the exam under a mandatory ACT policy. One way to test the sensitivity of the results to non-compliance is to restrict the sample in both the pre- and post-policy period to high school graduates not taking the special education version of the eleventh grade exam. Among this sample, the ACT-taking rate in the post-policy period is near 100%.[25]

I show the moments and percentiles of the ACT score distributions for this sample in Table 1.3, columns 4 and 5. The distribution of takers pre-policy (column 4) is virtually identical to the entire sample. The distribution of non-takers (column 5) is also very similar, though shifted slightly higher than the entire sample. The proportion of college-ready non-takers to test-takers is also similar at 0.206 (standard error of 0.015) compared to 0.227 among the entire sample. Using an ACT score threshold of 22 instead of 20 leads to a proportion of 0.180, relative to 0.192 in the original sample. These proportions are not statistically different from those estimated using the original sample.

Another way to test the sensitivity to non-compliance is to use the entire sample of post-policy test-takers to predict the ACT scores of the post-policy non-compliers, using observed characteristics. I estimate the following equation using students in the post-policy period:

$$ACT_{isdt} = \alpha + \beta_1 X_{isdt} + \beta_2 S_{st} + \beta_3 D_{dt} + \varepsilon_{isdt} \,, \qquad (2)$$

[25] One reason for non-compliance in the post period in addition to high school dropout and taking the special education version of the eleventh grade test is migration out-of-state. A small fraction of students may enroll in their spring eleventh grade semester but then move out-of-state mid-semester before the eleventh grade test is offered in March. While my college-enrollment measure captures out-of-state attendance, I only observe ACT-taking for students who take the exam in Michigan. This could cause bias in my analysis of ACT-taking if students moving out-of-state are a selected sample. Testing the sensitivity of the results to non-compliance in the post period also effectively tests against bias due to out-of-state migration.

where ACT_{isdt} is the ACT score of student *i* in school *s* in district *d* in cohort *t*. The other variables are as in Equation (1). I predict $\hat{ACT}_{isdt}$ out of sample for the non-compliers who do not take the ACT post-policy, and then proceed with the exercise as before. As shown in column 8 of Table 1.3, the distribution of scores shifts downward slightly. The proportion of college-ready non-takers to test-takers increases somewhat from 0.227 to 0.260, but again these are not statistically different. The proportion when using an ACT score threshold of 22 rather than 20 increases from 0.192 to 0.217. These increases are due to the fact that while most of the non-compliers in the post period would score below the college-ready threshold, some would not. These high-scoring non-compliers boost the number of college-ready non-takers, thus increasing the proportion.[26]

The predicted ACT scores using observed characteristics are likely to be an upper bound on the true latent scores of the non-compliers, given that students who drop out of school are likely to be of lower ability on unobserved characteristics. This sensitivity check shows that even given an upper bound on how these non-compliers would score, the statistic of interest (0.260) is similar to that estimated without the predicted scores of the non-compliers (0.227%). Given that using a sample with near 100% compliance, or predicting the scores of non-compliers, does not yield substantively different main results than using the original sample, I proceed with the remainder of the Section 1.4 analysis using the original sample without predicting the scores of non-compliers.

1.4.3. Geographic Generalizability: Are the Results Michigan-Specific?

It is possible that the latent ability distribution of students not taking a college entrance exam varies across regions of the country. Thus, one concern with my analysis is that the results regarding the supply of college-ready students not taking a college entrance exam would not generalize to other states. To address this issue, I mimic the above analysis using ACT micro-data from the three other earliest-adopting mandatory ACT states. My strategy is identical, except that I cannot use DFL-reweighting to account for changing student composition over time because I observe only those students who take the ACT, rather than the full sample of eleventh graders, before and after the reform,

[26] Using this logic, the 0.227 estimated using the main sample in columns 1 and 2 is a lower bound in the sense that any non-compliers in the post period scoring >=20 will boost this proportion since the number of pre-policy takers earning >=20 remains unchanged.

I instead weight ACT-takers using state-year CCD counts of eleventh graders by sex, race, and poverty status. ACT-takers in the pre period receive a weight of 4 because the data are a 25% sample. ACT-takers in the post period receive a weight calculated using the following formula:

$$Weight_{grp}^{s} = 4 * \frac{N_g^{s,pre}}{N_g^{s,post}} * \frac{N_r^{s,pre}}{N_r^{s,post}} * \frac{N_p^{s,pre}}{N_p^{s,post}}, \quad (3)$$

where $Weight_{grp}^{s}$ is the weight given to a post policy ACT-taker in state *s* with sex *g*, race *r*, and poverty status *p*. $N_g^{s,pre}$ is the number of eleventh graders with sex *g* in state *s* during the three years prior to the mandatory ACT policy. $N_r^{s,post}$ is the number of eleventh graders with race *r* in state *s* during the three years after the implementation of mandatory ACT in that state. Poverty status in the CCD is proxied for using free lunch eligibility.[27] I attempt to replicate this proxy in the ACT micro-data using self-reported family income and free lunch income eligibility thresholds by year.[28] I normalize the post-policy weights to have a mean equal to $(4 * \frac{N^{pre}}{N^{post}})$ to control for changes in school-age population size in the state from before to after the policy change.

Figure 1.2 shows the results of this analysis. Looking first at panel (b), the distributions of ACT scores look very similar to those shown in Figure 1.1. In the previous analysis using the census of ACT and SAT takers, I calculate a proportion of college-ready non-takers to takers of 0.227. The proportion calculated using the sample of ACT-micro data and CCD counts is slightly higher at 0.244. As a reminder, when I calculate this proportion using the Michigan data without including the SAT data, the proportion is 0.241, almost identical to that using the ACT micro-data.

Panel (a) combines results for Illinois and Kentucky. The picture looks very similar to that of Michigan. The rate of college-ready non-takers to takers in this combined sample is 0.232, very similar to the 0.244 calculated for Michigan. When splitting out Illinois from Kentucky, we see that the rate is slightly higher in Illinois (Panel c, 0.250) and slightly lower in

[27] Unlike sex and race, the number of students on free lunch in the CCD is available only as school totals, not by grade. I create the total number of eleventh graders receiving free lunch in a state by assuming that the fraction of eleventh graders receiving free lunch in a school is the same as the total fraction of the school receiving free-lunch.
[28] While several studies have shown a substantial degree of measurement error in ACT/SAT self-reported family income measures (Avery and Hoxby, 2012; Card and Payne, 2002), this strategy is the best available given that I do not observe free lunch status in the ACT micro data.

Kentucky (Panel d, 0.208). The similarity of these results suggests that the analysis I am conducting is generalizable at least to two of the other earliest adopting mandatory states.[29]

1.4.4. Who Are the High-Achieving Non-Takers?

I have shown that there is a non-trivial supply of "high-achieving non-takers," or students who do not take a college entrance exam but would score college-ready if they did. It is important to understand whether this supply varies across different subgroups of the student population. This heterogeneity has implications for which groups of students might experience larger impacts of the mandatory ACT policy on postsecondary outcomes. Moreover, if there is a larger supply of these students among disadvantaged populations, this would support explanations for the income-gap in college enrollment, such as information barriers and complexity in the financial aid and college application process.

In Figure 1.3, I plot the distributions of post-policy scores, pre-policy scores of test-takers, and predicted pre-policy scores of non-takers separately by sex, race, and free-lunch status. The first noticeable difference when comparing the frequency distributions of black to white students, or poor to non-poor students, is the far smaller number of disadvantaged students taking a college entrance exam; and, among those taking the exam, the lower scores earned. As the differences in the supply of college-ready non-takers relative to college-ready takers is difficult to discern visually, I report the results numerically in Table 1.4.

Column 1 of Table 1.4 replicates the results reported in the first two columns of Table 1.3. Examining heterogeneity by race shows a slightly lower proportion of college-ready non-takers to takers among black students as compared to white students. However the standard errors on the statistics for the black students are quite large. Looking by sex and using a college-readiness threshold of 20 in Table 1.4, there appears to be a somewhat larger (and statistically different) supply of male college-ready non-takers relative to female college-ready non-takers. For every 100 male students who take a college entrance exam and score college-ready, there are approximately 27 male students who do not take the exam but would score at a college-ready level. The comparable statistic among females is 20.5. The most dramatic heterogeneity is seen

[29] I exclude Colorado from this analysis because unlike the other three mandatory ACT states, there is a high rate of SAT-taking pre-policy (32% of high school graduates relative to 12% in Illinois, 8% in Kentucky, and 10% in Michigan, as published by ACT Inc.). High achieving students who take the SAT and not the ACT pre-policy will be incorrectly categorized as college-ready non-takers. As expected, I calculate a proportion of college-ready non-takers to takers of 0.41 in Colorado, nearly twice as large as in the other mandatory ACT states.

by poverty status. The proportion of non-poor, college-ready non-takers to takers is near the levels we have seen thus far at 0.217. The proportion for poor students is 0.480. For every 100 poor students taking a college entrance exam and scoring at a college-ready level, there are nearly 50 poor students who do not take the exam but would score college-ready.[30]

This large supply of college-ready poor students not taking a college entrance exam provides evidence that the supply of "missing one-offs" that has been identified in recent literature (Avery and Hoxby, 2012; Dillon and Smith, 2013; Bowen, Chingos, and McPherson, 2009) exists earlier in the college application pipeline. Indeed, these high-achieving students have not even made it past the earliest hurdles in the college application process, but would be qualified to enroll at a reasonably selective four-year college.

Given the large supply of college-ready non-takers among poor students, examining heterogeneity within this group of poor students has potential policy relevance. I split poor students by their urban / rural status and by their eighth grade test score on the state assessment. If promising non-takers are concentrated geographically, this would provide policymakers with a more targeted population at which to aim their reforms. Conditioning on whether students earn high or low eighth grade test scores is particularly policy-relevant, because teachers and guidance counselors can use these scores to determine their investment of resources during high school.

I find that the proportion of college-ready non-takers to takers is very high among poor urban students (0.54), and among poor students with below average eighth grade test scores (0.65). For every 10 such students taking the ACT or SAT and scoring college-ready, there are between 5 and 7 who do not take the exam but would score college-ready. There are smaller but still large supplies of these students among poor-rural students (0.39) and among poor students with above average eighth grade test scores (0.37). These results suggest that teachers and guidance counselors should not assume that disadvantaged students who score poorly on state assessments would not be qualified to enroll in a four-year college, if set on the proper path.[31]

[30] For each calculation by subgroup, I restrict the sample to students in that group and create a new set of DFL weights scaled to adjust for the different sample sizes pre- versus post-policy. Thus, the larger number of free lunch or minority students in the post period does not mechanically lead to a larger proportion of college-ready non-takers to takers among these groups.

[31] I also examine heterogeneity by school-level characteristics such as school poverty share, and whether the school was a pre-policy ACT test center, two subgroups that are of interest later in the paper. I find a slightly larger proportion among high poverty high schools and among schools that were not a pre-policy center, but the results are not statistically significantly different across the groups.

In Figure 1.4a, I examine the sensitivity of the results to the choice of college-readiness threshold. The X-axis is the ACT score used as the threshold. I additionally label most ACT scores with an institution in Michigan that has this score as the 25th percentile for entering students.[32] The Y-axis gives the proportion of students who do not take a college entrance exam, but would score college-ready relative to the number of college-ready test-takers. The solid line with circle-shaped markers represents the calculation using all students. The circle at ACT score 20 gives the original 0.227 shown in Table 1.3. The square marker at ACT score 20 gives the 0.480 for poor students.

While the proportion of college-ready non-takers to takers among the overall sample is relatively stable across the choice of college-readiness threshold (approximately between 0.2 and 0.3), the result among students qualifying for free-lunch varies greatly depending on the choice of threshold. The proportion is about 2/3 when using a college-readiness threshold of 18, but is less than 1/3 when using a threshold in the mid-20's. Interestingly, when we look by urban / rural status among poor students, the proportion of college-ready non-takers to takers remains quite high: around 0.50 to 0.55 across ACT score thresholds in the mid 20's. This is consistent with the results in Table 1.4 (row 1, columns 8 and 9), showing that the fraction of poor rural students who take a college entrance exam and score college-ready is much higher than the fraction among poor urban students. This pattern of results suggests that high-achieving poor students not embarking on the path toward four-year college enrollment is more prominent in urban than in rural areas.

1.4.5. Are They Actually College-Ready?

Students with high latent college entrance exam scores might not take the test because they have additional information suggesting that conditional on receiving a high score, they would still be ineligible for admission to a four-year college. I test whether I observe the same proportion of college-ready non-takers to takers, conditioning on a high school GPA above some threshold that indicates a reasonable likelihood of success at a four-year college. I also condition on whether a student has taken a course-load in high school that indicates college-readiness. If I find that the previously estimated proportions mostly disappear once I condition on a good high school GPA, or a challenging high school course-load, then perhaps there is less room than

[32] The University of Michigan – Ann Arbor, with a 25th percentile ACT score of 27, is not included in the figure.

previously thought for the mandatory ACT policy to have an impact on college-going. The signal of college-readiness from ACT may come too late for these students—their path is already set.[33]

Figure 1.4b plots the fraction of college-ready non-takers to takers using an ACT score college-readiness threshold of 20, and high school GPA above various thresholds.[34] Because 11.5% of ACT-takers do not report their GPA, I plot bounds on the proportion of college-ready non-takers to takers, assuming missing GPA's are A's versus assuming they are D's. The bars reflect the midpoint of the bounds, and the error bands reflect the bounds themselves. The 0.227 presented in Table 1.3 without conditioning on GPA is within the bounds of the first bar, in the figure that counts students as college-ready if they earn above a 2.0 GPA. However, the proportion decreases linearly with GPA threshold: the proportion is between 0.15 and 0.18 using a GPA threshold of greater than 3.0, and falls to between 0.11 and 0.14 using a GPA threshold of greater than 3.5. Further conditioning on taking a challenging curriculum changes these proportions by very little so I exclude these results.

The rates also fall when conditioning on high school GPA among poor students. The proportion is between 0.20 and 0.33 using a GPA threshold of greater than 3.5. Given my goal of picking a college-readiness threshold that measures likely success at a reasonably selective four-year college, a GPA threshold of greater than 3.0 is consistent with prior literature.[35] In this range, the proportion of college-ready non-takers to takers for poor students is between 0.25 and 0.33. This range is substantially lower than the 0.48 estimated without conditioning on GPA, suggesting that many of the high-achieving students not taking a college entrance exam might be making a more informed decision than their latent test score would suggest: even if they took the exam, their GPA might disqualify them for admission to a quality four-year institution. Still, for every 100 poor students with a greater than 3.0 high school GPA, who take the ACT or SAT and score college-ready, there are between 25 and 33 similarly high achieving poor students who do not take the exam, but would score college-ready. This substantial supply of high-achieving low-

[33] The high school GPA and course-taking information that I use for this exercise is self-reported on the ACT Student Questionnaire. It has been shown to correspond very closely to courses taken and grades earned as reported on student transcripts (Freeberg, 1988; Sawyer et al., 1988; Valiga, 1987). As a measure of a college-ready curriculum, I use what ACT calls a "core curriculum:" whether students have taken (or plan to take) four years of English and at least three years each of math, science, and social studies.

[34] I use the same process, including the DFL re-weighting, as described in Section IVa. The only difference is that I tally the number of ACT-takers pre- and post-policy by their GPA in addition to their score.

[35] In Roderick et al. (2009), their metric for qualifying to enroll in a selective four-year institution for students scoring between 18-20 on the ACT is earning a high school GPA of 3.0 or higher.

income individuals certainly seem eligible for four-year postsecondary enrollment, but are not making it past the earliest hurdles in the college application process.

1.5. Effects of the Mandatory ACT Policy on Postsecondary Outcomes

In this section, I examine the effects of the mandatory ACT policy in Michigan on postsecondary enrollment, persistence, and choice.[36]

1.5.1. Effects on College Enrollment and Choice

The simplest methodology for examining the effect of the mandatory ACT policy on college enrollment is to examine how enrollment changes from before to after the policy. As previously shown in Table 1.2, the average postsecondary enrollment rate among the three pre-policy cohorts in my sample is 0.57. The average rate among the three post-policy cohorts is 0.589, or 1.9 percentage points higher. The increase in the enrollment rate at four-year colleges is 1.0 percentage points. These increases, however, may not represent the true impact of the mandatory ACT policy. The sinking economy, shifting demographic composition, similarly timed education reforms, and any other factors changing over this time period, could be affecting the college enrollment behavior of Michigan students.[37]

To mitigate the biases resulting from these omitted factors, I estimate the causal impact of mandatory ACT-taking on postsecondary enrollment in Michigan using a treatment-comparison research design. Specifically, I compare changes in college attendance between the pre- and post-policy periods in schools that did not have an ACT test center in the school pre-policy, to those that did. I estimate the following equation using OLS:

$$Y_{isdt} = \beta_0 + \beta_1 Post_t + \beta_2 NoCenter_{isd} + \beta_3 (Post_t \times NoCenter_{isd}) + \beta_4 X_{isdt} + \alpha_s + \varepsilon_{isdt}\,, \qquad (4)$$

where Y_{isdt} is a college enrollment outcome for student *i* in school *s* in district *d* in cohort *t*. *Post* is a dummy for attending eleventh grade during the post period, *NoCenter* is a dummy for attending a school without an ACT test center pre-policy (which drops out when I include school

[36] Appendix 1.1 presents an analysis of the effect of the policy in other treated states using data from IPEDS. I also conduct analysis using a number of other nationally representative data sets. The results are statistically imprecise with no clear pattern across treated state or specification. They are available upon request.

[37] The Michigan Promise Scholarship was a short-lived merit scholarship that offered up to $4,000 toward college for the last pre-mandatory ACT policy cohort and first two post-policy cohorts. Preliminary findings suggest little to no impact of the policy on college-going (Dynarski, et al. 2013). The Michigan Merit Curriculum, also implemented around this time, is a reform increasing the course requirements necessary to graduate high school. However, the first cohort exposed to the policy is eleventh graders in 2010 who are not in my sample.

fixed effects), X is a vector of student-level and school- and district-year level covariates, and α is a full set of school fixed effects.[38] ε is the error term, clustered at the school level. β_3 is the coefficient of interest, and gives the effect of the policy in schools with no test center relative to those with a center.

The motivation behind the above strategy is the hypothesis that schools without a test center will experience a slightly larger increase in ACT-taking associated with the implementation of the mandatory ACT policy than will their pre-existing test center counterparts. The identifying assumption behind my estimation strategy is that any differential changes in college enrollment after the mandatory ACT policy between the students in these two groups of schools are due to the differential effects of the policy on ACT-taking in these schools. Other similarly timed statewide education reforms or factors that are changing over time, and that could affect college-going, are assumed to affect the two types of schools equally.

In Table 1.5, I examine empirically whether the identifying assumption is likely to be true. Columns 1 and 2 show student-weighted sample means of schools with and without a test center before the mandatory ACT policy. Slightly over half of students in Michigan attend a school with a test center, even though there are double the number of schools without a center. Not only are schools with test centers much larger, but they tend to enroll students with higher academic achievement, higher ACT-taking rates, and higher educational attainment. Schools with a test center are more likely to be in an urban or suburban area, and less likely to be in a rural area.

It is not surprising that schools with a center are quite different than those without, as becoming a test center is primarily a demand-driven phenomenon. To become a test center a teacher, counselor, or administrator from the school fills out an online form. They agree to be open on at least one testing day per year, must expect at least 35 students on the testing day, and must have the proper room conditions and seating arrangements, which are then verified by an ACT official.

Given the difference-in-difference research design, it is not a threat to the validity of my estimates if the two types of schools are different, but rather if the schools are changing

[38] Unless otherwise noted, X includes student-level variables sex, race, free lunch status, LEP, SPED, eighth grade test scores, school-year level variables fraction black, free lunch, number of eleventh graders, average eighth grade scores, and the same district-year level covariates plus guidance counselor to pupil ratio, dummies indicating urban / rural status, and the local unemployment rate.

differentially over time. In columns 4 and 5 of Table 1.5, I plot means at the two types of schools in the post period, and the difference-in-difference estimate in column 7. There is some evidence that the populations of these schools are changing differentially over time. There is a statistically significant increase in free lunch status for schools without a center over time, relative to schools with a center, and a statistically significant decrease in enrollment.

To ensure that the compared schools with and without a test center are similar except for their test center status, I use propensity score matching on a series of school- and district-year level observed characteristics to create a sample of matched test center and non-test center schools. I run a probit regression of whether a school has a test center on school- and district-level pupil-teacher ratio, percent free or reduced price lunch, grade 11 enrollment, and fraction black. I also include (a) average school-level eighth and eleventh grade state assessment scores; (b) dummies for whether the school is in a suburban area, small town, or rural area; (c) the growth rate in the school's eleventh grade enrollment; (d) the district-year level guidance counselor to student ratio; and (e) the local unemployment rate. The fitted value from that regression is the propensity score.

I use nearest neighbor matching (without replacement), which matches each school with a test center to the non-test center school with the closest propensity score. Because some of the schools with a test center have extremely high propensity scores where there are few similar non-test center schools, I trim the ten percent of schools with the highest propensity scores.[39] In my sample, nearest neighbor matching tends to produce the best balance of covariates, but I show that my results are not sensitive to either propensity score reweighting, or to other methods of matching such as kernel or caliper matching, that have been shown to produce superior results in some contexts (Busso, DiNardo, and McCrary, 2013; Heckman et al., 1997).[40] Similarly, trimming fewer of the center-schools produces similar results but inferior covariate balance.[41]

Table 1.6 again shows sample means by test center status pre- and post-policy, but now including only the propensity score matched sample of schools. The balance of covariates by test

[39] These tend to be very large schools in suburban areas.

[40] Radius, or caliper, matching matches test center schools to all non-test center schools within a specified range of propensity scores (I use one percent). Kernel matching does not explicitly drop schools but rather for each test center school weights nearby non-test center schools using a kernel function (I use an Epanechnikov kernel).

[41] If I trim the sample by twenty percent, my college enrollment results display the same pattern of heterogeneity and are slightly larger in magnitude. If I do not trim any of the test center schools with the highest propensity scores, the balance of covariates across the two types of schools is substantially worse and the pattern of heterogeneity is again the same, but slightly smaller in magnitude.

center status pre-policy is much better, but still not completely balanced. More importantly, there is now no evidence that the schools in this matched sample are trending differentially with respect to their composition.[42] None of the covariates have a statistically significant difference-in-difference estimate. Rates of ACT-taking at schools without a pre-policy center nonetheless increase by 4 percentage points after the policy relative to schools with a pre-policy center. This 4 percentage point gap arguably captures the effect of having a test center in one's high school on test-taking. There is no difference-in-difference effect on high school graduation or overall college enrollment, but a marginally statistically significant 0.8 percentage point increase in four-year enrollment.

To further explore the validity of the difference-in-difference methodology, I plot college attendance rates of schools with and without a test center in the matched sample by cohort. Figure 1.5 shows that trends in college enrollment are nearly identical across the two types of schools prior to the mandatory ACT policy. This is reassuring, as it suggests that college enrollment would have continued to trend in parallel in the absence of the policy, satisfying one of the key identifying assumptions in my estimation strategy. The level of four-year college enrollment is higher in the schools with a test center, presumably reflecting that some of the students induced into taking the ACT, by having a center in their school, subsequently enroll in a four-year college.

The top panel of Figure 1.5 shows that there is a small increase post-policy in the level of college enrollment in the schools with no test center pre-policy, relative to those with a center pre-policy. Panel (b) shows that there is an almost one percentage point jump in four-year enrollment in schools without a pre-policy center relative to those with a center, cutting the gap between the two types of schools by more than half. There is also a slight relative drop in two-year college enrollment, suggesting that part of the increase in four-year enrollment is due to changes in where students attend college as opposed to increased enrollment.

I now turn to estimating the effect of the policy on college enrollment using the regression-adjusted difference-in-difference Equation (4). Table 1.7, row 1, column 1 shows

[42] It is also important to note that most stories involving differences in unobservables biasing the effects would provide a downward bias on the results. If particularly active or motivated teachers, counselors, or administrators are those who initiate a test center at a school, it seems likely that such staff would more effectively implement the mandatory ACT policy than staff at non-center schools. In fact, the author's discussions with guidance counselors and visits to high schools suggest the existence of heterogeneity in the implementation of the mandatory ACT policy, such as differences in the levels of awareness and culture surrounding the test, and in test-preparation.

little effect of the policy on enrollment. Column 2 adds covariates and column 3 adds school fixed effects. Columns 4 and 5 use the other matching methods. The point estimate on any college enrollment is between 0.3 and 0.5 percentage points, is statistically insignificant, and is fairly stable across the columns. In columns 6–10, I examine effects on whether a student enrolls at a four-year institution. There is a 0.8 percentage point effect of the policy, with a standard error of 0.4 percentage points. Adding covariates does not alter the estimate but the inclusion of school fixed effects lowers the coefficient to 0.6 percentage points. This represents a 1.9% increase in the four-year enrollment rate, off of the pre-policy mean of 32.1%. The effect is identical using kernel matching, and slightly higher using propensity score reweighting, but is not particularly sensitive.[43]

The coefficient on the *Post* dummy in column 8 of Table 1.7 indicates that there is a 1.1 percentage point, statistically significant increase in four-year enrollment post policy among students at schools with a test center pre-policy. The 0.6 percentage point increase for the no-test-center schools is above and beyond this 1.1 percentage point increase. I discuss the effects of the policy entirely discounting the 1.1 percentage point effect, because I cannot disentangle the effects of the policy for schools with a pre-policy center from other factors changing over time. In this sense, the effect I discuss in the above paragraph represents a lower bound on the true statewide effect of the policy, if we think that any of the 1.1 percentage point effect for the schools with a center is due to the mandatory ACT policy.

Finally I examine the effect of the policy on two-year enrollment.[44] There is a small, negative and statistically insignificant point estimate for two-year enrollment. It is stable as I add in covariates, school fixed effects, and use different matching and reweighting methods.

1.5.2. Heterogeneity of Impacts

While all students are affected by the mandatory ACT policy as they are all now required to take the test, there is no reason to think that all students will be equally impacted by the

[43] I also examine whether the policy increases enrollment at selective four-year colleges. I define selective colleges as those with Barron's quality rankings of Most Competitive, Highly Competitive, or Very Competitive. In Michigan, 11% of students in the matched sample of schools enroll at such colleges pre-policy, relative to the 32% that enroll at any four-year college. Unfortunately the results are too imprecise to detect a selective college enrollment effect. In results not shown, I observe a very small and statistically insignificant effect of the policy on selective college enrollment. I similarly find no statistically significant impact of the policy on out-of-state enrollment.

[44] I define two-year enrollment as enrolling in a two-year school and not a four-year school, so that two- and four-year enrollment are mutually exclusive.

policy. Many students would have taken the ACT regardless of the policy. Other students are forced to take the ACT, but are so academically unprepared—or otherwise off the path of application to college—that being forced to take the exam will have no impact on their educational plans.

To hone in on the marginal student most impacted by this policy, I create an index measuring the predicted probability that a student would take the ACT based on the pre-policy relationship between ACT-taking and student-level observed demographic characteristics. Specifically, I estimate the following equation using OLS:

$$TAKE_{isdt} = \beta_0 + \beta_1 X_{isdt} + \alpha_s + \varepsilon_{isdt} \,, \tag{5}$$

where X includes all main effects and interactions of sex, race, free and reduced-price lunch status, and LEP and SPED status. α is again a full set of school fixed effects.[45] I estimate this equation and then predict $\widehat{TAKE}$ for all students pre- and post-policy, which is their pre-policy predicted probability of ACT-taking.

Figure 1.6a breaks students into vigintiles (twenty quantiles) based on this index, and plots mean ACT-taking rates of students in pre-policy cohorts (solid line) and of students in post-policy cohorts (dashed line). The distance between the two lines in this figure could be thought of as the dosage of the treatment, in the sense that it gives the change in the ACT-taking rate for students with a given probability of taking the ACT pre-policy. Table 1.8 reports the difference-in-difference effects of the policy on ACT-taking and college enrollment by quintiles of this predicted probability index. Among all students, there is a 3.4 percentage point effect of the policy on ACT-taking in non-test center high schools, relative to test center schools (column 1, row 1).[46] The largest increases in ACT-taking occur among students with the lowest pre-policy probability.

The remaining rows of Table 1.8, column 1 replicate the preferred specification from Table 1.7. Despite the large impact on ACT-taking among students with a very low pre-policy probability, the effects on four-year enrollment are near zero for this group. Effects are largest on four-year college enrollment for students with a low or mid-level probability. The effects are

[45] Table 1.A.1 reports the results from this regression. The results are nearly identical when using probit or logit.
[46] Pre-policy means of the dependent variable are reported in italics beneath the standard errors.

then near zero again for students in the top two quintiles of the probability index.[47] In figure 1.6b, I plot the pre-policy raw four-year college enrollment rates for each vigintile of the predicted probability of ACT-taking (solid line). I then estimate Equation (4) separately for each vigintile and add the difference-in-difference coefficient to the pre-policy rate (dashed line). As seen in Table 1.8, the enrollment effects are entirely concentrated within the second and third quintiles of the predicted probability index.

To increase precision and collapse students into a group that seems marginal, and a group whose college enrollment behavior seems relatively unaffected by the policy, I combine the low and middle students together, and the very low, high, and very high students together. I call this latter group the "tails" of the distribution. Among students in the low to middle range of the predicted probability index (between the two vertical lines in figure 1.6), there is a 1.4 percentage point increase in college enrollment due to the policy among students at schools without a test center, relative to schools with a center. This effect is driven entirely by a 1.3 percentage point increase in enrollment at four-year colleges, representing a five percent increase. There is little evidence of any effect on four-year enrollment among students in the tails of the predicted probability, nor on two-year enrollment in either of the groups.

To guide policy, it would be helpful to examine which types of students along specific observed dimensions are most impacted by the mandatory ACT. Table 1.9 examines the effect of the policy by race, sex, and poverty status. The effects are largest among students who have the lowest rates of postsecondary attainment pre-policy. Males experience a 0.9 percentage point increase in four-year enrollment (standard error of 0.4 points). Students eligible for subsidized meals experience a similarly sized effect, representing a 6% increase relative to their pre-policy mean. Black students also appear to experience a relatively large effect of the policy; however, the results are imprecisely estimated. As with students overall, while the effect of the policy on two-year enrollment appears to be negative, none of these subgroups experience a statistically significant change in two-year enrollment.

Finally, I examine the effects by school poverty share. This is an even more policy-relevant dimension, as education policies are easier to implement at the school level than to apply to only students with particular characteristics. I split students into terciles based on the

[47] Results are similar when dividing the predicted probability index by tercile or quartile and are available upon request.

share of students in their school that qualify for free or reduced-price lunch. I then combine students in the low and middle poverty schools, and compare to the effects on those in high poverty schools. Students in high poverty schools experience an increase in four-year enrollment of 1.3 percentage points (standard error of 0.6 percentage points). This effect represents a 5.7% increase in four-year enrollment. There is no impact among students at schools with low or middle levels of poverty, and no statistically significant impacts on two-year enrollment in either group.

All of the margins of heterogeneity explored thus far have exploited the demographic characteristics of students and where they attend school. Another interesting question is whether the effects of this policy vary by student ability. This dimension of heterogeneity is related to the "undermatch" phenomenon. If this policy induces students to enroll who are high achieving but not on the path to a four-year college, then it will likely improve the match between student ability and postsecondary plans.

In Table 1.10, I examine the effect of the policy by eighth grade test score, which proxies for student ability. I use the simple average of a student's math and English Language Arts (ELA) scores, and standardize these average scores within cohort and year to have a mean of zero and a standard deviation of one.[48] I then split students into those who have an above average test score, and those with a below average score.[49] Not all eleventh graders in my sample were in Michigan public schools during eighth grade. Ten percent of the sample is missing eighth grade score. Thus, I first replicate my previous results using this smaller sample of students (column 1). The effects are nearly identical: the four-year college enrollment effect is now 0.7 percentage points, again a 2% increase.

Columns 2 and 3 show results for high and low ability students, respectively. The coefficients are very similar across the two groups, and neither is statistically significant. Due to the lower pre-policy mean enrollment rate among the low-scoring students, the effect in percentage terms for this group is much larger than for the high scoring students. Given that the effect of the policy on four-year enrollment is concentrated among relatively disadvantaged students at poor schools, this is perhaps unsurprising as these students tend to have lower eighth grade scores. To examine whether high and low achieving students are differentially affected by

[48] The individual math and ELA are also standardized before taking their average.
[49] As I further break these groups by predicted probability of ACT-taking, finer splits by eighth grade score than above and below the average yield imprecise estimates.

the policy controlling for demographics, I split the students by whether they attend a high poverty school. Column 5 shows that the effect on four-year enrollment rates among students attending high poverty schools is indeed concentrated among those with above average eighth grade scores.[50] These students experience a 2.3 percentage point increase due to the policy (standard error of 1.2 points).[51] However, the 1.1 percentage point gain among the low achieving counterparts (column 6), is greater in percent terms. It is thus difficult to conclude whether the policy is disproportionally affecting high ability students.

1.5.3. Do Marginal Enrollees Drop Out?

While college entry has been rising in recent decades, college completion has remained flat (Bound, Lovenheim, and Turner, 2010). A key concern with a policy such as the mandatory ACT is that if it induces students on the margin to attend college, these students may not persist. If this is the case, then the effects on four-year enrollment rates would overstate the benefits of the program. Alternatively, the fact that the policy seems to increase enrollment at four-year colleges, with some possible switching from two-years, is possibly optimistic for expecting the enrollment increases to translate into persistence. Bound, Lovenheim and Turner (2010) show that the recent decline in college completion has been driven by students attending nonselective institutions such as community colleges, where there are fewer resources per student, leading to less support and increased dropout rates.[52]

In Table 1.11, I examine the effects of the policy on the share of students who enroll in a four-year college and persist to the second, third, and fourth year. As a reminder, the definition of enrollment is whether a student enrolls by the second fall following on-time high school graduation. Students in the most recent cohort who enrolled in college during the second fall after on-time high school graduation have only had time to progress through their first year of college. Consequently, this exercise requires dropping post-policy cohorts from the sample. Row 1 (columns 1–3) reports the previously estimated four-year enrollment results for all cohorts overall, and by school poverty share. The second (non-missing) row for those columns shows the effects on enrolling after dropping the most recent post-policy cohort. The effect on overall enrollment is 0.7 percentage points and statistically significant at the 10% level.

[50] I use the overall mean standardized score of zero as opposed to the within-group mean score.
[51] Similarly, a linear interaction of the treatment and eighth grade score has a near zero coefficient for the overall sample, but a positive and statistically significant coefficient among students at high poverty schools.
[52] Other studies also find that students who start at a community college are less likely to eventually earn a bachelor's degree than those starting at a four-year college (Reynolds, 2012; Long and Kurlaender, 2009).

The next row shows the effect on enrolling and persisting to the second year. The overall effect is somewhat attenuated from 0.7 to 0.5 percentage points, and loses statistical significance. However, because the fraction of students pre-policy enrolling and persisting to the second year is smaller than the fraction enrolling, the effect in percentage terms is 1.8%, only slightly smaller than the 2.2% effect overall. This suggests that students induced by the policy to enroll in a four-year school persist to the second year of college at almost the same rate as inframarginal students. I then examine the effect of the mandatory ACT policy on enrollment and persistence to the third and fourth years of college. The effect on enrolling and persisting to the third year is 0.7 percentage points, identical to the overall effect estimated using only one post-policy cohort. Given, the difference in pre-policy means, this implies that students induced to enroll by the policy actually persist at higher rates than inframarginal students.

Examining persistence to the fourth year requires redefining the enrollment measure. In columns 4–6, I restrict enrollment to be by the first fall following on-time high school graduation. Using this measure of enrollment, and all three post cohorts, the overall effect is the same as before (0.6 percentage points) and statistically significant at the 10% level. The effects on persistence to years two and three are identical in percent terms to the enrollment effects using the respective samples. And the effect on enrolling and persisting to year four is slightly greater in percentage terms than the enrollment effect. Though statistically imprecise, these results suggest that marginal students persist through college at the same rate as their inframarginal peers. This result is consistent with other interventions that provide information and dismantle barriers to the college application process, and find that students induced to enroll persist at high rates (Bettinger et al., 2012; Carrell and Sacerdote, 2013).

While the ultimate postsecondary outcome is whether a student completes college, the implementation of the policy is too recent to accurately assess if there are increases in degree completion. In the bottom row of Table 1.11, in column 4, I report the overall effect on graduating from a four-year college within four years of scheduled on-time high school graduation. Only 9% of the sample pre-policy earns a bachelor's degree in four years with this rate doubling when examining five-year completion rates. Unfortunately, I cannot observe five-year graduation rates as the first cohort exposed to the policy is currently in their fifth year of

college.[53] The results for four-year graduation rates are quite imprecise, but there is suggestive evidence that the enrollment and persistence effects translate into degree completion. The overall effect is 0.3 percentage points, which is a 3% increase relative to the pre-policy mean.[54]

1.5.4. Robustness Checks

I present several robustness checks to examine the sensitivity of my estimates. In Table 1.12, columns 1–3, I present estimates that control for pre-trending of the outcome variable. Given the relatively few data points in the sample before and after the policy change (three before, three after) these are not my preferred specifications. Yet it is informative to present the results as a specification check. I estimate the following equation:

$$Y_{isdt} = \beta_0 + \beta_1 Post1_t + \beta_2 Post2_t + \beta_3 Post3_t + \beta_4 (Post1_t \ x \ NoCenter_{isd}) + \beta_5 (Post2_t \ x \ NoCenter_{isd}) + \beta_6 (Post3_t \ x \ NoCenter_{isd}) + Cohort_t + \beta_7 X_{isdt} + \alpha_s + \varepsilon_{isdt}\ , \quad (6)$$

where *Post1*, *Post2*, and *Post3* are dummies for the three post periods, and *Cohort* is a linear time trend. The reported coefficient is $\frac{\beta 4+\beta 5+\beta 6}{3}$, or the average effect of being in the post period conditional on the covariates, school fixed effects, and pre-trend. Estimating the post dummies separately assures that the trend is estimated using only the pre-policy period.

The results controlling for pre-trends are very similar to the main results, though slightly attenuated. The point estimate for the overall sample is 0.5 percentage points. The effect among students attending a high poverty high school is 1.2 percentage points. The results are similar but much less precise when I control for school-specific time-trends.

My next robustness check uses a different method of constructing the treatment and comparison groups. Instead of grouping students by whether they attend a high school with or without a test center pre-policy, I use a student's home address during the eleventh grade, and the address of the nearest pre-policy test center, to create a driving distance from the student's home to that test center. I then separate students by whether they are less than or greater than five miles

[53] I receive National Student Clearinghouse data in the summer, which means that for a cohort currently enrolled in their fifth year of college, I observe whether they enroll in year four and graduate after their fourth year.

[54] The effects on persistence by school poverty largely reflect the pattern seen overall, though the persistence effect is more attenuated among the high poverty group. The effect on graduation nearly entirely disappears among this group. The overall point estimate appears to be driven by the students in the low / middle poverty schools for whom the enrollment effect is quite small.

to the nearest pre-policy center (the pre-policy mean).[55] This strategy serves as a test of the external validity of the matched sample to the entire Michigan sample, as well as a test of the sensitivity of the results to the different method of constructing the treatment / comparison group.[56]

Columns 4–6 of Table 1.12 present the results using the driving distance analysis for students in the matched sample of schools.[57] The effects on four year enrollment are larger and more precise than when using the test center measure and the pattern of heterogeneity is the same. There is a 1.0 percentage point increase in four-year enrollment overall, and a 1.9 point increase among students at high poverty schools. Both effects are statistically significant at the 1% level. The effect on two-year enrollment is negative as in the main analysis, however, it is larger in magnitude, and is statistically significant.

In columns 7–9, I do not restrict the analysis to students in the matched sample of schools. The results are slightly attenuated but are very similar. The effect on four-year enrollment for all students is 0.9 percentage points (standard error of 0.3 points). The pattern of heterogeneity is the same, though less extreme. Using this distance-based methodology, it appears that the effects of the policy estimated on the matched sample extrapolate to the entire population of Michigan. I cannot say with certainty whether these results would have external validity in other states, given differences in higher education systems and other factors. However, the fact that the analysis in Section 1.4 finds that the supply of college-ready students not taking an exam pre-policy is nearly identical across other mandatory ACT states, suggests that the postsecondary results may be generalizable as well.

1.6. Discussion

The effects estimated in this paper using the difference-in-difference design are not average treatment effects (ATE) for the entire population of Michigan, or even for the entire

[55] Results are nearly identical when I instead use three miles (the pre-policy median) to do the split. See Figure 1.A.1 (map) for means of student-level driving distance pre-policy by school district.

[56] I prefer to use the school-level test center method as my main strategy, and to leave the distance method as a robustness check for two reasons: 1) any attempt to separate students by distance into treatment and comparison groups is arbitrary since distance is a continuous measure, and 2) it is easier to understand the selection process of schools becoming test centers than of students living close to or far from a test center. Thus, I can more convincingly sign any possible bias due to selection on unobserved characteristics when using the test center strategy than when using the distance strategy.

[57] I drop the 2% of the sample with missing home addresses.

matched sample of schools. They are local average treatment effects (LATE) estimated for a specific and marginal group of students. The LATE is the expected outcome gain, or ATE, for those induced to receive treatment through a change in the instrument (Imbens and Angrist, 1994). In this context, these are post-policy ACT-takers who were enrolled in a high school without a pre-policy center, would not have taken a college entrance exam pre-policy in their high school, but would have if enrolled at a high school with a center. I estimate the effect of the policy as the effect of inducing these students to take the ACT. This means that to get a treatment on the treated (TOT) estimate for this specific group of students, I can think of the difference-in-difference increase in ACT-taking as a first stage by which I can scale up the effects on four-year enrollment.

Scaling the effects on four-year enrollment by the first-stage effect on ACT-taking yields a TOT effect suggesting that 18% of students induced into taking the ACT by the policy would subsequently enroll in a four year college (= 0.6 / 3.4). This percentage is larger among the subgroups of students that experience the largest impacts of the policy on 4-year enrollment.[58] These fairly large results are plausible given that marginal students picked up by LATEs in the context of education policies often realize large treatment effects (Card, 1995). However, these results are unlikely scalable to the entire population of students induced to take the ACT by the mandatory ACT policy. If the results were scalable, we would expect to see statewide increases in four-year enrollment rates due to this policy greater than 6 percentage points, given the 35 percentage point increase in ACT-taking. Michigan experienced nowhere near this dramatic increase in four year enrollment post-policy.

Although my estimated reduced form four-year college enrollment effect is a LATE that cannot be extrapolated to the entire population, it is a policy-relevant parameter. A state considering implementing the mandatory ACT policy could expect to observe this effect among high schools without a pre-existing test center. There could also be an additional effect for all schools that I am unable to identify with the difference-in-differences design.

A remaining threat to validity that I cannot rule out involves supply-side capacity constraints on the side of colleges. If there are a fixed number of slots, at least in the short-run,

[58] Results are the same for a more formal two-stage-least-squares analysis of the effect of taking the ACT on enrollment, where the excluded instrument is the interaction of a dummy for being in the post-policy period with a dummy for being enrolled in a school without a pre-policy center. Results of this analysis are available from the author upon request.

then it is possible that colleges could be accepting more applications from students in schools with no pre-policy center, and accepting fewer applications from students enrolled at high schools with a pre-policy center. This would positively bias my estimated treatment effect, artificially suggesting an enrollment effect, when there is only a compositional effect of who is enrolling. While I cannot conclusively rule out this story, there is little reason to think that in the matched sample of schools, students would be displaced at higher rates from schools with a pre-policy test center than from schools without one. The two types of schools are similar across observed characteristics and have similarly sized supplies of college-goers pre-policy who could be potentially displaced by the new enrollees.

Though I argue that capacity constraints should equally affect the treatment and comparison groups in my analysis, I do not assert that they should be taken lightly. Bound and Turner (2007) show that a 10% increase in a state's cohort size leads to a 4% decrease in the fraction of students earning a BA from that state. They find that part of this cohort crowding effect is due to decreased completion rates at non-selective colleges. These colleges respond to the increased demand by enrolling more students, but subsequently have lower per-student resources to serve these students and help them through college. The other portion of the decrease in the fraction of students earning a BA is due to decreased enrollment rates at flagship public universities that want to maintain high per-student resources, and so do not expand slots to accommodate the increased demand.

These results suggest several testable implications for my analysis.[59] First, the types of institutions in which there is plausibly greater demand due to the mandatory ACT policy are those that have binding capacity constraints. If a substantial portion of the four-year enrollment effect that I estimate in this paper is due to increases at non-selective four-year colleges, than this could provide evidence that students induced into selective colleges by the mandatory ACT policy are displacing inframarginal students into less selective colleges.

Second, if cohorts tend to be increasing in size, as is the case in my sample from the pre to post period, this should support the possibility of capacity constraints. If on the other hand, cohorts were decreasing in size (as they are in the last two cohorts in my sample and younger, recent cohorts), then this could leave more slots open for students induced in by the mandatory

[59] In a future draft of this paper, I test these implications empirically to examine whether capacity constraints appear to be an issue in this context.

ACT. This is particularly likely since public universities are required to enroll a certain fraction of student from in-state. Furthermore, it implies that variation across states in the effectiveness of the mandatory college entrance exam policy at boosting postsecondary enrollment rates could stem from whether states are in a period of population expansion or contraction.

Finally, the discussion of capacity constraints and cohort crowding brings up the discussion of equity and efficiency in the sorting of students into postsecondary education. Even if there is one-for-one crowd out due to the policy where every student induced in displaces a student who would have attended, this would still be a policy that reduces inequality in educational attainment as I find that those induced in are disadvantaged and from high-poverty schools. Prior to the policy, there is discrimination in the postsecondary market in the sense that college-ready students for some reason other than their productivity are not enrolling in college. I show that after the barrier has been lifted, underprivileged students enroll at higher rates, so that the policy increases equity, even if it doesn't increase enrollment rates. Similarly, if colleges pre-policy are letting in students on the margin who are less well-qualified than those admitted after the policy, then the mandatory ACT policy would increase efficiency in the sorting of students into college as well as equity.

1.7. Conclusion

Nearly a dozen states have incorporated the ACT or SAT into their eleventh grade statewide assessment, requiring that all public school students take a college entrance exam. In this paper, I use the implementation of this policy to show that there are many students, particularly low-income students, who would score quite well on a college entrance exam but who do not take the test. For every ten poor students who take a college entrance exam and score college-ready, there are an additional five poor students who do not take the test but would score college-ready.

I compare changes in college-going rates pre- and post-policy among students at schools that did not have an ACT test center pre-policy to those that did, showing that the policy increases four-year enrollment rates by 0.6 percentage points or 2%. This increase is 1.3 points for students with a low- to mid-level predicted probability of taking the ACT in the absence of the policy. It is similarly high among males (0.9 points), poor students (1.0 point), and among students in the poorest third of high schools (1.3 points). Though estimated imprecisely, the

effect on enrolling in a four-year college for up to four years is similar, implying that the students induced to attend college by the policy persist at the same rate as inframarginal college-goers.

While these increases in the four-year college enrollment rate might not appear to be dramatically large, relative to other educational interventions this policy is inexpensive and currently being implemented on a large scale. The direct costs to states of a mandatory ACT policy include: (1) the per-student test fee, which for spring 2012 was $32 (a $2 discount off the price a student would pay privately);[60] (2) a statewide administration management fee, which is approximately another dollar per student; and (3) the costs associated with trainings, meetings, and other logistical issues, which comes to less than a dollar per student.[61] While (2) and (3) seem to vary by state, the total cost is less than $50 per student for all mandatory ACT states. Further, this is an upper bound, given that the actual cost to a state is the direct cost of the policy minus the cost to design, administer, and grade the portions of the eleventh grade exam displaced by the ACT.[62]

To show the relative cost-effectiveness of the mandatory ACT policy at increasing postsecondary attainment, I compare this policy to other educational interventions that increase college-going. Following Dynarski, Hyman, and Schanzenbach (forthcoming), I create an index of cost-effectiveness by dividing a policy's cost by the proportion of students it induces into college. For example, assuming a $50 per student cost and an increase in the four-year college enrollment rate of 0.6 percentage points, the amount spent by the mandatory ACT policy to induce a single child into college is $8,333 (=$50 / 0.006).

Dynarski (2003) shows that for $7,000 in traditional student aid, about two-thirds of the treated students attending college were inframarginal, while the other one-third were induced into college by the scholarship. The cost per student induced into college in that study is therefore $21,000, because three students take the scholarship for every one induced in. Reduced class size during elementary school in the Tennessee STAR experiment boosted college enrollment by 3 percentage points (Dynarski, Hyman, and Schanzenbach, forthcoming). The cost

[60] If a state chooses to include the writing portion of the ACT, the cost is an additional $15 per test.

[61] All mandatory ACT cost numbers in this section come from communication between the author and staff at state departments of education. All costs of other policies in this section are in 2007 dollars and come from Levine and Zimmerman (2010) unless otherwise noted. The costs of the early childhood programs and STAR have been discounted back to age zero using a 3% discount rate. Costs of mandatory ACT and other high school and college interventions have not been discounted.

[62] States must still administer another eleventh grade assessment to supplement the ACT and include social studies and other requirements of No Child Left Behind. However, certainly it is cheaper to design and grade fewer tests.

per student was $12,000, resulting in a cost per student induced into college of $400,000 (=$12,000 / 0.03). Deming (2009) shows a college enrollment effect of Head Start, the federally funded preschool program, near 6 percentage points. The cost per student induced into college by attending Head Start is therefore $133,000 (=$8,000 / 0.06).

These more traditional education policies are far more expensive than the mandatory ACT policy. Alternatively, Bettinger et al. (2012) randomly offer families assistance filling out the FAFSA, which costs $88 per student, and increases college enrollment by 8 percentage points. The cost per student induced into college is therefore $1,100 (=$88/0.08). This policy is extremely cost effective; more so than the mandatory ACT. However, it is unclear whether this policy could be successfully operated on a scale as large as the mandatory ACT policy.[63] For instance, class size reduction modeled after the success of STAR was less effective at the state level than in the experiment (Jepsen and Rivkin, 2009). Relative to other interventions operating on a large scale, such as traditional student aid, the mandatory ACT policy seems very cost-effective.[64]

Furthermore, if the mandatory ACT policy was targeted at the poorest third of schools, the cost per student induced into college would drop to under $4,000. One policy discussion moving forward could center around whether states could implement this policy only in certain schools, or whether this is logistically infeasible or jeopardizes the effectiveness of the policy. Doing so would require states to use a different statewide test for accountability purposes. Regardless of targeting, the mandatory ACT policy appears to be a relatively low-cost intervention operating on a large-scale that increases postsecondary attainment.

Still, the mandatory ACT is far from a cure-all. The results in Section 1.4 suggest that requiring all students to take a college entrance exam increases the supply of poor students scoring at a college-ready level on the exam by nearly 50%. Yet the policy increases the number of poor students enrolling at a four-year institution by only 6%. In spite of the policy, there remains a large supply of disadvantaged students who are high-achieving, and not on the path to

[63] Another recent intervention offers students college application fee waivers and information regarding college application strategies and financial aid (Hoxby and Turner, 2013). I exclude this study from the cost / benefit comparison because they focus on changes in the number and selectivity of colleges applied to, and the selectivity of the college in which a student enrolls, rather than simply college attendance.

[64] One caveat to this cost / benefit analysis is that the marginal student may vary across studies: the student who is induced to attend college by assistance filling out the FAFSA may differ from who is induced from smaller classes. The above differences in cost / benefit could reflect that it is more difficult to induce certain students. Ideally I would like to compare the cost of inducing the same student into college.

enrolling at a four-year college. The important question still stands: What policies beyond simply requiring these students to take a college entrance exam will help them to achieve their educational potential, and reduce the income-gap in postsecondary attainment? Mandatory ACT helps, but is not the final act.

References

ACT Inc. 2002. "Interpreting ACT Assessment Scores." Research Publication. http://www.act.org/research/researchers/briefs/2002-1.html#UItAIYq5fw [accessed May 4, 2013].

Avery, Christopher and Caroline Hoxby. 2012. "The Missing 'One-Offs': The Hidden Supply of Low-Income, High-Achieving Students for Selective Colleges." *NBER working paper* 18586.

Bailey, Martha J. and Susan M. Dynarski. 2011. "Gains and Gaps: A Historical Perspective on Inequality in College Entry and Completion." In Greg Duncan and Richard Murnane, eds. *Social Inequality and Educational Disadvantage*, Russell Sage.

Beshears, John, James J. Choi, David Laibson, and Bridgette C. Madrian. 2009. "The Importance of Default Options for Retirement Saving Outcomes: Evidence from the United States." In Jeffrey Brown, Jeffrey Liebman, and David A. Wise (Eds.) *Social Security policy in a changing environment.* Chicago: University of Chicago Press.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119(1): 249–275.

Bettinger, Eric P., Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu. 2012. "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment." *Quarterly Journal of Economics* 127(3): 1205–1242.

Bound, John, Michael Lovenheim, and Sarah E. Turner. 2010. "Why Have College Completion Rates Declined? An Analysis of Changing Student Preparation and Collegiate Resources." *American Economic Journal: Applied Economics* 2(3): 1–31.

Bound, John, and Sarah E. Turner. 2007. "Cohort Crowding: How Resources Affect College Attainment." *Journal of Public Economics* 91: 877–899.

Bowen, W., M. Chingos, M. McPherson. 2009. *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton: Princeton University Press.

Bulman, George. 2013. "The Effect of Access to College Assessments on Enrollment and Attainment." Working paper.

Busso, Matias, John DiNardo, and Justin McCrary. 2013. "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects." Working Paper.

Card, David. 1995. "Earnings, Schooling, and Ability Revisted." In Solomon Polachek, ed., *Research in Labor Economics*, vol. 14, Greenwich Connecticut: JAI Press.

Card, David and A. Abigail Payne. 2002. “School Finance Reform, the Distribution of School Spending, and the Distribution of Student Test Scores.” *Journal of Public Economics* 83: 49–82.

Carrell, Scott and Bruce Sacerdote. 2013. “Do Late Interventions Matter Too? An Experiment to Raise College Going Among NH High School Seniors.” *NBER Working Paper* #19031.

Deming, David. 2009. “Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 2009 1 (3): 111–134.

Deming, David and Susan M. Dynarski. 2010. “Into College, Out of Poverty? Policies to Increase the Postsecondary Attainment of the Poor.” In Phillip B. Levine and David J. Zimmerman, eds., *Targeting Investments in Children: Fighting Poverty When Resources Are Limited*, 283–302. Chicago: University of Chicago Press.

Dillon, Eleanor and Jeffrey Smith. 2013. “The Determinants of Mismatch Between Students and Colleges.” Work in progress.

Dynarski, Susan M. 2003. “Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion.” *American Economic Review* 91 (1): 279–288.

Dynarski, Susan M., Steven Hemelt, and Joshua M. Hyman. 2013. “Data Watch: Using National Student Clearinghouse (NSC) Data to Track Postsecondary Outcomes.” Working paper.

Dynarski, Susan M., Joshua M. Hyman, and Diane Whitmore Schanzenbach. Forthcoming. “Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion.” *Journal of Policy Analysis and Management*.

Dynarski, Susan M., Ken Frank, Brian Jacob, and Barbara Schneider. 2013. “The Effect of the Michigan Promise Scholarship on Educational Outcomes.” Work in progress.

Efron, Bradley and Robert Tibshirani. 1993. “An Introduction to the Bootstrap.” *Monographs on Statistics and Applied Probability* 57, Chapman Hall.

Freeberg, N. 1988. "Accuracy of Student Reported Information." College Board Report Number 88-5.

Goodman, Sarena. 2012 “Learning from the Test: Raising Selective College Enrollment by Providing Information.” Working paper.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme.” *Review of Economic Studies* 64 (4): 605–654.

Hoxby, Caroline and Sarah Turner. 2013. “Expanding College Opportunities for High-Achieving, Low Income Students.” Stanford University working paper.

Imbens, Guido W. and Joshua Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62(2): 467–475.

Jacob, Brian A. and Tamara Wilder. 2010. “Educational Expectations and Attainment.” *NBER* Working Paper #15683.

Jackson, C. Kirabo. 2010. “A Little Now for a Lot Later: A Look at a Texas Advanced Placement Incentive Program.” *Journal of Human Resources* 45(3): 591–639.

Jepsen, Christopher and Steven Rivkin. 2009. "Class Size Reduction and Student Achievement: The Potential Tradeoff Between Teacher Quality and Class Size." *Journal of Human Resources* 44 (1): 223–250.

Klasik, Daniel. Forthcoming. "The ACT of Enrollment: The College Enrollment Effects of State-Required College Entrance Exam Testing." *Educational Researcher*.

Levine, Phillip B. and David J. Zimmerman, eds. 2010. *Targeting Investments in Children: Fighting Poverty When Resources Are Limited*. Chicago: University of Chicago Press.

Long, Bridget. T. and Michal Kurlaender. 2009. "Do Community Colleges Provide a Viable Pathway to a Baccalaureate Degree?" *Educational Evaluation and Policy Analysis* 31 (1): 30–53.

Madrian, B. C. and Shea, D. F. 2001. "The power of suggestion: Inertia in 401(k) participation and savings behavior." *The Quarterly Journal of Economics* 116(4): 1149–1187.

Pallais, Amanda. 2011. "Small Differences that Matter: Mistakes in Applying to College." Working paper.

Pallais, Amanda and Sarah Turner. 2006. "Opportunities for Low Income Students at Top Colleges and Universities: Policy Initiatives and the Distribution of Students." *National Tax Journal* 59(2): 357–386.

Reynolds, C. Lockwood. 2012. "Where to Attend? Estimating the Effects of Beginning College at a Two-Year Institution." *Economics of Education Review* 31(4): 345–362.

Roderick, M., Nagaoka, J., Coca, V., and Moeller, E. 2009. "From high school to the future: Making hard work pay off." *Consortium for Chicago School Research*. Available:http://ccsr.uchicago.edu/publications/high-school-future-making-hard-work-pay [accessed May 4, 2013].

Sawyer, Richard, Joan Laing, and Walter Houston. 1988. "Accuracy of Self-Reported High School Course and Grades of College-Bound Students." ACT Research Report No. 88–1. Iowa City, IA: ACT, Inc.

Stinebrickner, Todd R. and Ralph Stinebrickner. 2012. "Learning about Academic Ability and the College Dropout Decision," *Journal of Labor Economics* 30(4): 707–748.

Valiga, Michael J. 1987. "Accuracy of Self-Reported High School Course and Grade Information." ACT Research Report No. 87–1. Iowa City, IA: ACT, Inc.

Appendix 1.1: Replication of Goodman (2012)

A recent working paper using aggregate institution-year level data from the Integrated Postsecondary Education Data System (IPEDS) finds much larger effects from the mandatory ACT policy than does my paper (Goodman, 2012). In this appendix, I replicate the author's analysis and show that using equally or arguably more valid specification choices, I find that the results become substantially attenuated and statistically indistinguishable from zero. This is an important finding given the large difference in magnitude between our results.

I reconstruct the sample from Goodman (2012) using data from IPEDS. IPEDS collects annual information from every university, college, and vocational and technical postsecondary institution nationwide that participates in federal student aid programs. Goodman (2012) defines enrollment as first-time, first-year enrollment of degree- or certificate-seeking students. Her measure is disaggregated by state of residence when the student was admitted. This information is reported in even years only since 1994. Unless otherwise noted, in my construction of the data and discussion of the methodology, my choices follow Goodman (2012).

Also following Goodman (2012), I merge the IPEDSs data to Barron's college quality rankings, and to several state-year level covariates derived from the March Current Population Survey.[65] I designate a selective college as one that has a Barron's quality ranking higher than the lowest ranking of "uncompetitive." Goodman also examines several increasingly selective definitions of selectivity, which I do not report, but my replication of the author's results and my subsequent null finding are consistent for these other definitions of selectivity. The state-year level covariates that I construct from the CPS include the unemployment rate among adults aged 16 or older, the fraction of adults aged 25 or older with a BA, the population of 16 year olds lagged by two years, the fraction of all residents who are poor, and the fraction who are minorities.

Goodman uses a difference-in-difference methodology to examine the effects of the mandatory ACT policy in the two earliest adopting states, Colorado and Illinois. The control states she uses are the 22 other states in which the ACT is the dominant college entrance exam,

[65] The author does not indicate which monthly CPS survey she uses to create these covariates, so I assume it is the main March survey. The summary statistics of these covariates are nearly identical to those in her paper.

excluding Michigan since its first high school graduating class exposed to the policy is in 2008, the final year in the sample.[66] The author's primary estimating equation is:

$$ln(ENR)_{st} = \alpha + \beta ACT_{st} + \lambda_t + \gamma_s + X_{st} + \varepsilon_{st}\,, \quad (13)$$

where $ln(ENR)_{st}$ is logged first-time enrollment in state s and year t, ACT_{st} is an indicator for being a treated state in the post-policy period, λ_t and γ_s are a set of year and state fixed effects, respectively, and $X_{st.}$ is the set of state-year level covariates. ε_{st} is the error term, which I cluster at the state level following Bertrand et al. (2000). The identifying assumption is that any changes in the dependent variable experienced by students in treated states, relative to students in comparison states, after the mandatory ACT policy was implemented in 2002, are due to the policy. Under this assumption, β gives the effect of the policy.

In a separate specification, Goodman includes state-specific linear time trends to account for pre-trending of the dependent variable, allowing for treatment and control states to follow different trends:

$$Y_{st} = \alpha + \beta ACT_{st} + \lambda_t + \gamma_s + \gamma_s * Trend_t + X_{st} + \varepsilon_{st}\,, \quad (14)$$

where $Trend_t$ is the linear time tend. Following the author, when I estimate Equation (14) I use unclustered heteroskedasticity-robust standard errors, because they are more conservative than clustering by state.

I report the main results transcribed from Goodman (2012) in the first row of Table 1.A.2. In columns 1 and 2, the author finds no effects of the policy on overall college enrollment. However, in columns 3 through 8, the author finds statistically significant effects of the policy on selective college enrollment ranging between 10% and 16%. Column 4 controls for all of the covariates listed above except for the two year lagged population of 16 year olds. Column 5 controls for the log of this population. Column 6 controls for all of the covariates and the logged population. Given the apparent effects on selective enrollment but not overall enrollment, Column 7 includes the logged total enrollment in the state-year as a further control for any factors affecting postsecondary enrollment. Column 8 adds the covariates from the CPS to this specification.

[66] In her main result, the author excludes 2010 due to the implementation of the policy in Kentucky and Tennessee in that year.

I present my best attempt at replicating the author's results in the second row. The point estimates and standard errors on overall enrollment (columns 1 and 2) are identical. Turning to selective enrollment, the results are very similar, but not identical. One possible explanation for these small discrepancies could be differences between our versions of the Barron's data, or in our matching of the Barron's data to the IPEDS. Regardless, the magnitudes of the coefficients, standard errors, and the pattern of results across the various specifications are very similar to those reported in Goodman (2012).

The first way in which I change the author's specification is by adding an additional set of four covariates to the four she includes. I add the median household income from the CPS as an additional control beyond the unemployment rate already included, in an attempt to further capture the financial situation of families and opportunity cost of attending college. I then add three variables from the Common Core of Data (CCD), a survey of all public K-12 schools in the US administered annually by the National Center for Education Statistics. These variables are (1) average in-state tuition at public four-year universities, which again is meant to capture the cost of enrolling in college; (2) average per-pupil, public school, K-12 expenditure, which is meant to capture the educational environment of the state; and (3) the lagged number of eleventh graders enrolled in public schools as a second measure of cohort size (in addition to the two-year lagged count of 16 year olds) that also reflects what proportion of the school-aged population has dropped out by eleventh grade, is enrolled in a private school, or is homeschooled.

The third row of Table 1.A.2 includes these additional controls. In columns 4, 6, and 8 we see that when controlling for these extra measures, the percent increase in the number of students enrolling at a selective college drops to near zero. The effect loses statistical significance in columns 6 and 8, and maintains statistical significance at the 10% level in column 8. A possible threat to validity in this sort of difference-in-difference design is if the treatment states tend to adopt other education reforms, or otherwise invest in education in the years after the mandatory ACT policy. The drop in results when controlling for these education-related variables suggests this may be the case. When I add these additional covariates individually, the drop in the point estimates is most dramatic with the inclusion of the lagged number of eleventh graders.[67] This measure of cohort size could be picking up effects of other changes in the

[67] The result is the same whether or not I log the count of eleventh graders, and when I estimate the equation separately for Illinois or Colorado.

education system that are affecting high school dropout, homeschooling, and private school enrollment not picked up by the more broad measure of overall population.

The second way in which I change the author's specification is by using a different measure of college enrollment.[68] Goodman uses first-time, first-year (i.e., freshman) enrollment regardless of when a student graduated high school. This adds measurement error to the difference-in-difference research design, because some of the students enrolling during the post-policy period would have graduated high school before the policy was in effect. I instead use the same first-time enrollment variable, but counted only for those students who graduated high school within the last twelve months. This measure of *immediate* enrollment more accurately assigns enrollment to the pre- versus post-policy cohorts. As a large fraction of first-time enrollees graduated high school more than twelve months in the past, the difference between these two variables is non-trivial. The average state-year in my sample had approximately 35,000 first-time students and 24,000 first-time, immediate enrollees. The Digest of Education Statistics, put out annually by NCES, corroborates that immediate, first-time enrollees comprise roughly two-thirds of all first-time enrollment.[69]

Goodman's only justification for using the non-immediate first-time enrollment is a footnote: "IPEDS also releases counts for the number of first-time, first-year enrollees that have graduated high school or obtained an equivalent degree in the last 12 months, but these are less complete." The author does not include a citation for her statement that the data are incomplete, or expand on the way in which the data are incomplete. The author also does not mention that the use of non-immediate enrollment could bias her results.

I check the completeness of the immediate enrollment variable and find little evidence that it is incomplete. The institution-year level data include the number of first-time enrollees from each state (Goodman's measure) and the number of first-time, immediate enrollees from each state. I calculate that 1.2% of college-years have a missing value for the immediate enrollment variable for at least one of the fifty states, when the non-immediate enrollment variable from the same state is not missing. Most of these cases occur when there are only a handful of first-time students enrolled from the state. The total first-time enrollment of these

[68] I also test the sensitivity of her results to weighting by population and to using states in the same census region as control states (Goodman shows results using border states as controls) instead of all ACT states. The selective college enrollment results are uniformly smaller in magnitude when weighted by population or when using these other sets of control states, but they are similar and still statistically significant.

[69] http://nces.ed.gov/programs/digest/ [accessed May 4, 2013].

college-year-sending states that are missing immediate enrollment constitute 0.1% of all first-time enrollment in the sample.[70] Furthermore, the Digest of Education Statistics reports annually on both the first-time and immediate, first-time enrollment counts from IPEDS by year, and makes no mention of any incompleteness in the immediate enrollment data. Finally, nowhere on the IPEDS website does it mention incompleteness of this data. Communication between myself and IPEDS staff revealed that they have no knowledge of the by-state immediate enrollment measure being less complete than the overall by-state first-time enrollment. They also stated that missing values in this particular question of the survey are interpretable as zeros, which is not the case in other parts of the survey.

The fourth row of Table 1.A.2 uses the logged number of immediate first-time enrollees as the dependent variable. The point estimates for selective college enrollment are consistently smaller, but well within the confidence interval of Goodman's results.

After reporting her main results, Goodman shows that the effects become largely attenuated and lose statistical significance when controlling for state-specific time trends (column 9, Table 1.A.2). The author then shows that when she uses a different specification, the results are less sensitive to the inclusion of the time trends. This specification allows for separate effects of the policy in the first three years post-policy versus later years:

$$Y_{st} = \alpha + \beta_1(ACT_s * (<4 \text{ Years of policy})_t) + \beta_2(ACT_s * (\geq 4 \text{ Years of policy})_t) + \lambda_t + \gamma_s + X_{st} + \varepsilon_{st}, \quad (15)$$

where $(<4 \text{ Years of policy})_t$ is a dummy for being in the year 2002 or 2004 and $(\geq 4 \text{ Years of policy})$ is a dummy for being in the year 2006 or 2008. I report the author's results from regressions using Equation (15) in Table 1.A.3, columns 1 and 2. In column 1, Goodman reports a 10% increase in selective enrollment in the first years of the policy, and a 19% effect in the later years. Controlling for state-specific time trends in column 2 attenuates these results but they are still in the range of 6%–14% and are jointly statistically different from zero. In columns 3 and 4 I replicate these results, finding nearly identical point estimates and standard errors without the time trends (column 3), and similar point estimates and identical standard errors with the time

[70] I calculate this statistic by creating the college-year level total first-time enrollment totaled over those college-year-sending states with missing immediate enrollment. I then sum this over all college-years. I divide this number by total first-time enrollment summed over all college-year-sending states regardless of missing immediate enrollment status.

trends (column 4). Finally, I show the effect using this same specification but replacing the enrollment dependent variable with the measure of immediate enrollment. Comparing column 5 to column 3, the coefficients drop by one-half and one-third for the immediate and delayed effects respectively. More importantly, when including the time trends, the results vanish entirely and are statistically indistinguishable from zero.[71]

Goodman notes that the attenuation of her initial results when including the state time trends, and the finding of larger impacts in later post-policy years, is consistent with the initial small effect of the policy growing over time. Another explanation for this pattern is that Colorado and Illinois happen to have selective enrollment patterns that are trending upward over time relative to the control states. The fact that the effects disappear when I use immediate first-time enrollment, rather than all first-time enrollment, and control for state-specific time trends, is consistent with a story where the increases reflect shifts in educational investment and policy other than the mandatory ACT policy. It appears that the increases Goodman (2012) observes are due to other policies that are inducing older, non-traditional, first-time enrollees to attend college, not just students recently graduated from high school who were affected by the mandatory ACT policy.

[71] Goodman (2012) does not report results using this specification and controlling for covariates.

Table 1.1. States With Mandatory College Entrance Exam

	State	Exam	Year Implemented
1	Colorado	ACT	2001
2	Illinois	ACT	2001
3	Maine	SAT	2006
4	Michigan	ACT	2007
5	Kentucky	ACT	2008
6	Tennessee	ACT	2009
7	Delaware	SAT	2011
8	North Carolina	ACT	2012
9	Louisiana	ACT	2013
10	Wyoming	ACT	2013
11	Alabama	ACT	2014

Notes: Year refers to the first spring that 11th graders were (or will be) required to take the exam.

Table 1.2. Sample Means of Michigan Eleventh Grade Student Cohorts

	All Cohorts (2004-2009)	Pre-ACT Cohorts (2004-2006)	Post-ACT Cohorts (2007-2009)	Difference: (3)-(2)	P-Value: (4)=0
	(1)	(2)	(3)	(4)	(5)
Demographics					
Female	0.498	0.498	0.498	-0.001	0.436
White	0.764	0.778	0.751	-0.027	0.000
Black	0.167	0.155	0.179	0.024	0.000
Hispanic	0.033	0.031	0.035	0.004	0.000
Other race	0.035	0.036	0.035	0.000	0.258
Free or reduced lunch	0.283	0.241	0.322	0.080	0.000
Special Education	0.123	0.124	0.122	-0.002	0.041
Limited English	0.021	0.020	0.023	0.002	0.000
Local unemployment	8.25	7.34	9.13	1.79	0.000
Driving miles to nearest ACT test center	3.60	4.72	2.52	-2.20	0.000
Educational Attainment					
Reaches twelfth grade	0.908	0.904	0.912	0.009	0.000
Graduates high school	0.844	0.844	0.844	0.000	0.923
Enrolls in any college	0.580	0.570	0.589	0.020	0.000
Enrolls in four-yr college	0.314	0.309	0.319	0.010	0.000
ACT score	19.6	20.7	18.9	-1.9	0.000
ACT-Taking Rate:					
All students	0.739	0.558	0.912	0.354	0.000
Males	0.706	0.507	0.898	0.392	0.000
Females	0.771	0.611	0.926	0.315	0.000
Blacks	0.647	0.456	0.806	0.350	0.000
Whites	0.761	0.583	0.939	0.355	0.000
Free or reduced lunch	0.644	0.350	0.852	0.503	0.000
Non-free lunch	0.778	0.625	0.940	0.316	0.000
<Median grade eight score	0.662	0.401	0.902	0.501	0.000
>Median grade eight score	0.868	0.766	0.961	0.195	0.000
Missing grade eight score	0.101	0.123	0.079	-0.044	0.000
Took SAT	0.048	0.064	0.033	-0.030	0.000
Took SAT & ACT	0.045	0.058	0.033	-0.025	0.000
SAT Score	25.0	24.7	25.8	1.1	0.000
Students per cohort	122,301	119,928	124,586		
Total Students	733,466	359,752	373,714		

Notes: The sample is all first-time eleventh graders in Michigan public high schools during 2003-04 through 2008-09 conditional on reaching their eleventh grade spring semester. Free lunch, special education, and limited English proficiency status are all as of eleventh grade. Driving miles to nearest ACT test center are measured from a student's home address duing eleventh grade to the nearest ACT test center open during that year. First score is used for students taking the ACT multiple times. SAT score is scaled to ACT metric. College enrollment is measured as of 16 months (October 1st) following scheduled on-time high school graduation. Eighth grade score is the average of scores on the 8th grade math and writing exams, standardized at the subject-cohort level.

Table 1.3. Comparison of Distributions of Observed and Latent ACT Scores Pre-Policy, By Sensitivity Check

	Entire Sample			Exclude High School Dropouts and SPED Test-Takers			Include Predicted Scores of Non-Compliers in Post-Period		
	Takers	Non-Takers	Difference	Takers	Non-Takers	Difference	Takers	Non-Takers	Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Moments									
Mean	20.7	16.7	4.0	20.8	17.1	3.7	20.8	16.6	4.1
Standard Deviation	4.6	4.3	0.3	4.6	4.5	0.1	4.6	4.3	0.3
Percentiles									
5th	14	12	2	14	12	2	14	11	3
25th	17	14	3	17	14	3	17	14	3
Median	20	16	4	21	16	5	20	16	4
75th	24	19	5	24	19	5	24	19	5
95th	29	25	4	29	26	3	29	25	4
Scoring >=20									
Percent	58.2	21.3	36.8	58.9	23.7	35.1	58.2	19.8	38.4
Number	117,953	26,717	91,236	114,075	23,454	90,621	117,953	30,707	87,246
Non-Takers / Takers	0.227			0.206			0.260		
Scoring >=22									
Percent	41.9	13.0	28.9	42.4	15.0	27.5	41.9	11.9	30.0
Number	84,996	16,319	68,677	82,261	14,794	67,467	84,996	18,471	66,525
Non-Takers / Takers	0.192			0.180			0.217		

Notes: The sample is all first-time, public school Michigan eleventh graders in years 2004-2009, conditional on reaching spring of eleventh grade. Latent scores of non-takers estimated as explained in text and reweighted to adjust for cohort size and composition following DiNardo, Fortin and Lemieux (1996). Columns (4)-(9) test sensitivity of the analysis to non-compliance in the post period: Columns (4)-(6) exclude high school dropouts and students taking the special education version of the 11th grade test. Columns (7)-(9) include predicted scores of non-compliers based on observed characteristics of ACT-takers in the post period.

Table 1.4. Heterogeneity in the Pre-Policy Supply of College-Ready Students Not Taking a College Entrance Exam

								Among Poor Students			
	All	White	Black	Female	Male	Non-Poor	Poor	Urban	Rural	High Gr. 8 Scores	Low Gr. 8 Scores
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Scoring College-Ready (ACT>=20)											
Percent of ACT-Takers	58.2	64.5	19.4	57.4	59.2	62.5	33.5	20.0	47.0	56.7	16.4
	(0.9)	(0.5)	(2.1)	(0.9)	(0.8)	(0.7)	(1.2)	(2.0)	(0.9)	(1.0)	(0.8)
Percent of Non-Takers (latent score)	21.3	23.9	4.5	21.9	20.8	27.6	11.3	8.8	12.1	32.1	5.4
	(1.5)	(1.5)	(1.4)	(1.8)	(1.3)	(1.8)	(0.5)	(1.0)	(1.0)	(1.3)	(0.3)
Proportion of Non-Takers to Takers	0.227	0.221	0.175	0.193	0.265	0.217	0.480	0.544	0.392	0.365	0.651
	(0.016)	(0.013)	(0.060)	(0.016)	(0.017)	(0.014)	(0.032)	(0.092)	(0.043)	(0.024)	(0.054)
Scoring College-Ready (ACT>=22)											
Percent of ACT-Takers	41.9	47.0	10.1	40.7	43.4	45.7	20.6	10.7	30.3	37.8	7.9
	(0.8)	(0.6)	(1.2)	(0.9)	(0.8)	(0.8)	(0.8)	(1.2)	(0.8)	(0.9)	(0.4)
Percent of Non-Takers (latent score)	13.0	14.6	1.9	13.3	12.8	18.1	5.0	4.5	4.7	16.9	1.5
	(1.4)	(1.4)	(1.0)	(1.6)	(1.2)	(1.8)	(0.4)	(0.7)	(0.8)	(1.2)	(0.2)
Proportion of Non-Takers to Takers	0.192	0.185	0.137	0.164	0.222	0.195	0.343	0.518	0.234	0.287	0.383
	(0.020)	(0.016)	(0.065)	(0.020)	(0.020)	(0.018)	(0.034)	(0.095)	(0.049)	(0.027)	(0.055)

Notes: The sample is all first-time, public school Michigan eleventh graders in years 2004-2009, conditional on reaching spring of eleventh grade. Latent scores of non-takers estimated as explained in text and reweighted to adjust for cohort size and composition following DiNardo, Fortin and Lemieux (1996). Free lunch status measured as of eleventh grade. Standard errors in parentheses calculated using 200 bootstrap replications.

Table 1.5. Sample Means Pre- and Post-Policy, by Pre-Policy Test Center Status

	Before Mandatory ACT Policy			After Mandatory ACT Policy			Diff-in-Diff
	No Center	Center	Difference	No Center	Center	Difference	(6)-(3)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Demographics							
Black	0.124	0.166	-0.043*	0.145	0.180	-0.035	0.008
Hispanic	0.032	0.029	0.003	0.037	0.032	0.005	0.002
Free lunch	0.248	0.220	0.028*	0.331	0.292	0.040**	0.012*
Eighth grade scores	-0.009	0.071	-0.080**	-0.025	0.056	-0.080**	0.000
Pupil-teacher ratio	20.6	21.8	-1.2	19.8	20.1	-0.3	0.9
Grade 11 enrollment	216.6	345.1	-128.5***	223.3	360.1	-136.8***	-8.3*
Local unemployment	7.57	7.11	0.45*	9.26	8.83	0.43	-0.024
Urban area	0.543	0.711	-0.167***	0.551	0.714	-0.163***	0.004
Rural area	0.457	0.289	0.167***	0.449	0.286	0.163***	-0.004
Educational Attainment							
Take ACT or SAT	0.540	0.607	-0.067***	0.927	0.932	-0.005	0.061***
Graduate High School	0.847	0.876	-0.029***	0.847	0.879	-0.032***	-0.003
Enroll in any college	0.554	0.611	-0.056***	0.576	0.631	-0.055***	0.001
Enroll in four-year college	0.292	0.343	-0.050***	0.306	0.352	-0.046***	0.004
Enroll in two-year college	0.262	0.268	-0.006	0.270	0.279	-0.009	-0.003
Number of Schools	523	251		518	251		
Number of Students	165,009	181,463		168,825	186,468		

Notes: The sample is all first-time, public school Michigan eleventh graders in years 2004-2009, conditional on reaching spring of eleventh grade. "No Center" and "Center" refer to whether or not a high school was an ACT test center before the mandatory ACT policy. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Table 1.6. Sample Means Pre- and Post-Policy for Matched Sample of Schools, by Pre-Policy Test Center Status

	Before Mandatory ACT Policy			After Mandatory ACT Policy			Diff-in-Diff
	No Center	Center	Difference	No Center	Center	Difference	(6)-(3)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Demographics							
Black	0.120	0.169	-0.043*	0.138	0.184	-0.046	0.003
Hispanic	0.031	0.029	0.002	0.036	0.033	0.003	0.000
Free lunch	0.238	0.240	-0.001	0.312	0.319	-0.008	-0.006
Eighth grade scores	0.045	0.024	0.021	0.015	0.004	0.011	-0.010
Pupil-teacher ratio	20.5	22.2	-1.8	19.8	20.1	-0.3	1.5
Grade 11 enrollment	259.6	288.8	-29.2**	268.1	303.9	-35.8**	-6.7
Local unemployment	7.43	7.32	0.107	9.08	9.03	0.051	-0.057
Urban area	0.553	0.669	-0.116**	0.562	0.673	-0.111**	0.005
Rural area	0.447	0.331	0.116**	0.438	0.327	0.111**	-0.005
Educational Attainment							
Take ACT or SAT	0.566	0.592	-0.026*	0.941	0.928	0.012*	0.039***
Graduate High School	0.869	0.869	0.000	0.871	0.871	-0.001	-0.001
Enroll in any college	0.575	0.597	-0.022	0.599	0.616	-0.017	0.005
Enroll in four-year college	0.313	0.327	-0.015	0.329	0.335	-0.007	0.008*
Enroll in two-year college	0.262	0.270	-0.007	0.271	0.281	-0.011	-0.003
Number of Schools	226	226		226	226		
Number of Students	122,566	142,418		125,447	146,382		

Notes: The sample is all first-time, public school Michigan eleventh graders in years 2004-2009, conditional on reaching spring of eleventh grade. The sample is restricted to the 226 schools without a pre-policy ACT test center and the 226 schools with a pre-policy test center matched using nearest neighbor matching. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Table 1.7. The Effect of the Mandatory ACT on Postsecondary Enrollment

	Dep. Var. = Any Enrollment					Dep. Var. = Four-Year Enrollment					Dep. Var. = Two-Year Enrollment				
	Nearest Neighbor Matching			Kernel Matching	P-Score Weighting	Nearest Neighbor Matching			Kernel Matching	P-Score Weighting	Nearest Neighbor Matching			Kernel Matching	P-Score Weighting
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Post * No Test Center In School	0.005	0.003	0.003	0.003	0.004	0.008*	0.008*	0.006	0.006	0.007*	-0.003	-0.005	-0.003	-0.002	-0.002
	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Post	0.019***	0.029***	0.016***	0.015***	0.014***	0.008**	0.016***	0.011***	0.011***	0.011***	0.012***	0.013***	0.005	0.004	0.004
	(0.003)	(0.005)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.003)
No Test Center in School	-0.022	-0.014				-0.015	-0.014*				-0.007	0.000			
	(0.015)	(0.011)				(0.015)	(0.008)				(0.012)	(0.011)			
Covariates	N	Y	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y	Y	Y
School Fixed Effects	N	N	Y	Y	Y	N	N	Y	Y	Y	N	N	Y	Y	Y
Pre-Policy Mean		0.587		0.590	0.588		0.321		0.320	0.317		0.266		0.270	0.271
Sample Size		536,813		614,974	701,765		536,813		614,974	701,765		536,813		614,974	701,765

Notes: The sample is all first-time, public school Michigan eleventh graders in years 2004-2009, conditional on reaching spring of eleventh grade. For columns (1)-(3), (6)-(8), and (11)-(13), the sample is restricted to the 226 schools without a pre-policy ACT test center and the 226 schools with a pre-policy test center matched using single nearest neighbor matching without replacement. An epanechnikov kernel and bandwidth of 0.06 is used in columns (4), (9), and (14). Each column is a separate linear probability model regression. Postsecondary enrollment is measured as of October 1st following scheduled on-time high school graduation. Standard errors in parentheses are clustered at the school level. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Table 1.8. Using Students' Predicted Probability of ACT-Taking Pre-Policy to Narrow in on the Marginal Student

		Pre-Policy Probability(Take ACT)						
	All	Very Low	Low	Middle	High	Very High	Low/Middle	Tails
Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Take ACT	0.034***	0.044***	0.038***	0.028***	0.007	0.007	0.032***	0.036**
	(0.013)	(0.012)	(0.010)	(0.006)	(0.007)	(0.012)	(0.006)	(0.018)
	0.580	*0.199*	*0.457*	*0.600*	*0.710*	*0.835*	*0.531*	*0.618*
Enroll in:								
Any College	0.003	-0.001	0.013	0.014**	-0.008	0.003	0.014**	-0.003
	(0.004)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)	(0.006)	(0.004)
	0.587	*0.305*	*0.497*	*0.616*	*0.676*	*0.765*	*0.559*	*0.608*
Four-Year College	0.006	-0.002	0.013**	0.012**	0.001	0.001	0.013**	0.000
	(0.004)	(0.005)	(0.007)	(0.006)	(0.008)	(0.008)	(0.005)	(0.004)
	0.321	*0.077*	*0.207*	*0.305*	*0.398*	*0.553*	*0.259*	*0.369*
Two-Year College	-0.003	0.001	-0.000	0.001	-0.010	0.002	0.001	-0.003
	(0.004)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.005)	(0.004)
	0.266	*0.227*	*0.290*	*0.311*	*0.277*	*0.212*	*0.301*	*0.239*
Covariates	Y	Y	Y	Y	Y	Y	Y	Y
School Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Sample Size	536,813	86,136	117,944	117,381	104,082	111,270	235,325	301,488

Notes: The sample is all first-time, public school Michigan eleventh graders in years 2004-2009, conditional on reaching spring of eleventh grade. The sample is restricted to the 226 schools without a pre-policy ACT test center and the 226 schools with a pre-policy test center matched using nearest neighbor matching. Each point estimate is from a separate linear probability model, difference-in-difference regression. Standard errors in parentheses are clustered at the school level. Pre-policy dependent variable means are in italics. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Table 1.9. Heterogeneity in the Effect of the Mandatory ACT By Student Demographics and School Poverty Share

								School Poverty Share	
	All	White	Black	Female	Male	Non-Poor	Poor	Low/Middle	High
Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Enroll in:									
Any College	0.003	0.003	0.003	-0.000	0.005	-0.001	0.016**	-0.000	0.009
	(0.004)	(0.004)	(0.011)	(0.005)	(0.005)	(0.004)	(0.007)	(0.004)	(0.007)
	0.587	*0.605*	*0.515*	*0.622*	*0.552*	*0.640*	*0.415*	*0.634*	*0.494*
Four-Year College	0.006	0.005	0.009	0.002	0.009**	0.004	0.010**	0.001	0.013**
	(0.004)	(0.004)	(0.009)	(0.005)	(0.004)	(0.004)	(0.005)	(0.004)	(0.006)
	0.321	*0.334*	*0.256*	*0.350*	*0.291*	*0.370*	*0.164*	*0.368*	*0.228*
Two-Year College	-0.003	-0.002	-0.006	-0.002	-0.004	-0.005	0.006	-0.001	-0.004
	(0.004)	(0.004)	(0.009)	(0.005)	(0.004)	(0.004)	(0.006)	(0.004)	(0.006)
	0.266	*0.271*	*0.259*	*0.272*	*0.261*	*0.271*	*0.251*	*0.266*	*0.267*
Covariates	Y	Y	Y	Y	Y	Y	Y	Y	Y
School Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sample Size	536,813	417,851	83,061	268,573	268,240	384,331	148,147	358,113	178,700

Notes: The sample is all first-time, public school Michigan eleventh graders in years 2004-2009, conditional on reaching spring of eleventh grade. The sample is restricted to the 226 schools without a pre-policy ACT test center and the 226 schools with a pre-policy test center matched using nearest neighbor matching. Each point estimate is from a separate linear probability model, difference-in-difference regression. Free lunch is measured as of eleventh grade. Standard errors in parentheses are clustered at the school level. Pre-policy dependent variable means are in italics. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Table 1.10. Does the Mandatory ACT Policy Induce Higher Ability Students to Attend College?

	All Students			High Poverty Schools			Low/Middle Poverty Schools		
	All	High Scores	Low Scores	All	High Scores	Low Scores	All	High Scores	Low Scores
Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Enroll in:									
Any College	0.004	0.002	0.005	0.010	0.020*	0.005	0.000	-0.003	0.006
	(0.004)	(0.005)	(0.006)	(0.008)	(0.010)	(0.009)	(0.005)	(0.005)	(0.007)
	0.604	*0.718*	*0.473*	*0.569*	*0.654*	*0.429*	*0.647*	*0.737*	*0.507*
Four-Year College	0.007*	0.004	0.006	0.015**	0.023**	0.011*	0.001	-0.002	0.002
	(0.004)	(0.006)	(0.004)	(0.007)	(0.012)	(0.006)	(0.005)	(0.006)	(0.005)
	0.332	*0.470*	*0.173*	*0.240*	*0.383*	*0.152*	*0.377*	*0.497*	*0.190*
Two-Year College	-0.003	-0.002	-0.001	-0.005	-0.003	-0.006	-0.001	-0.001	0.004
	(0.004)	(0.004)	(0.005)	(0.006)	(0.006)	(0.007)	(0.004)	(0.005)	(0.006)
	0.272	*0.247*	*0.300*	*0.275*	*0.271*	*0.278*	*0.270*	*0.240*	*0.317*
Covariates	Y	Y	Y	Y	Y	Y	Y	Y	Y
School Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sample Size	486,522	247,731	238,791	158,477	57,091	101,386	328,045	190,640	137,405

Notes: The sample is as in Tables 1.6 - 1.9 except that it excludes the 10% of students with missing eighth grade test scores. Each point estimate is from a separate linear probability model, difference-in-difference regression. Scores are the average of standardized math and English eighth grade test scores. High and low scores are above and below the mean. Standard errors in parentheses are clustered at the school level. Pre-policy dependent variable means are in italics. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Table 1.11. Examining if Four-Year Enrollment Effects Persist, by Timing of Enrollmer

	Enroll Within Two Years			Enroll Within One Year		
		School Poverty			School Poverty	
	All	High	Low/Mid	All	High	Low/Mid
Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)
All Three Post Cohorts						
Enroll	0.006	0.013**	0.001	0.006*	0.013**	0.001
	(0.004)	(0.006)	(0.004)	(0.003)	(0.006)	(0.004)
	0.321	*0.228*	*0.368*	*0.291*	*0.204*	*0.336*
Persist to Year Two				0.005	0.008*	0.001
				(0.003)	(0.004)	(0.004)
				0.256	*0.193*	*0.305*
First Two Post Cohorts						
Enroll	0.007*	0.012*	0.003	0.006*	0.012**	0.002
	(0.004)	(0.006)	(0.005)	(0.004)	(0.006)	(0.005)
	0.321	*0.228*	*0.368*	*0.291*	*0.204*	*0.336*
Persist to Year Two	0.005	0.011*	0.001			
	(0.004)	(0.005)	(0.005)			
	0.278	*0.180*	*0.328*			
Persist to Year Three				0.005	0.008	0.002
				(0.003)	(0.005)	(0.004)
				0.239	*0.147*	*0.286*
First Post Cohort Only						
Enroll	0.007	0.017**	-0.000	0.006	0.010	0.003
	(0.005)	(0.007)	(0.006)	(0.004)	(0.007)	(0.005)
	0.321	*0.259*	*0.369*	*0.291*	*0.204*	*0.336*
Persist to Year Three	0.007	0.009	0.006			
	(0.004)	(0.008)	(0.005)			
	0.258	*0.228*	*0.368*			
Persist to Year Four				0.006	0.005	0.005
				(0.004)	(0.005)	(0.005)
				0.227	*0.136*	*0.273*
Graduate in Four Years				0.003	0.001	0.003
				(0.002)	(0.003)	(0.003)
				0.091	*0.040*	*0.116*
Sample Size, Three Pre Cohorts Plus:						
All Post Cohorts	536,813	178,700	358,113	536,813	178,700	358,113
First 2 Post	448,234	150,023	298,211	448,234	150,023	298,211
First Post Only	357,181	120,037	237,144	357,181	120,037	237,144
Covariates	Y	Y	Y	Y	Y	Y
School Fixed Effects	Y	Y	Y	Y	Y	Y

Notes: The sample is as in Tables 1.6 -1. 9. Each point estimate is from a separate linear probability model, difference-in-difference regression. Standard errors in parentheses are clustered at the school level. Pre-policy dependent variable means are in italics. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Table 1.12. Robustness Checks: Controlling for Pre-Trend and Using Student-Level Driving Distance to Nearest Pre-Policy Test Center

	Controlling for Pre-Trend			Using Student-Level Distance, Matched Sample of High Schools			Using Student-Level Distance, All High Schools		
		School Poverty			School Poverty			School Poverty	
	All	High	Low/Middle	All	High	Low/Middle	All	High	Low/Middle
Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Take ACT	0.033***	0.050***	0.022	0.048***	0.039***	0.055***	0.045***	0.029***	0.055***
	(0.013)	(0.015)	(0.015)	(0.009)	(0.014)	(0.011)	(0.008)	(0.011)	(0.009)
	0.580	0.470	0.636	*0.580*	0.470	0.637	*0.559*	0.431	0.622
Enroll in Any College	0.001	0.007	-0.002	0.003	0.012*	-0.001	0.002	0.001	0.002
	(0.004)	(0.007)	(0.004)	(0.004)	(0.007)	(0.004)	(0.003)	(0.006)	(0.003)
	0.587	*0.494*	*0.634*	*0.587*	*0.494*	*0.635*	*0.570*	*0.460*	*0.625*
Enroll in Four-Year College	0.005	0.012*	0.000	0.010***	0.019***	0.005	0.009***	0.010*	0.007**
	(0.004)	(0.006)	(0.004)	(0.004)	(0.007)	(0.004)	(0.003)	(0.006)	(0.003)
	0.321	*0.228*	*0.368*	*0.321*	*0.227*	*0.369*	*0.309*	*0.210*	*0.359*
Enroll in Two-Year College	-0.004	-0.005	-0.002	-0.008**	-0.007	-0.007*	-0.007**	-0.008	-0.005
	(0.004)	(0.006)	(0.004)	(0.003)	(0.006)	(0.004)	(0.003)	(0.005)	(0.003)
	0.266	*0.267*	*0.266*	*0.266*	*0.267*	*0.266*	*0.261*	*0.250*	*0.266*
Covariates	Y	Y	Y	Y	Y	Y	Y	Y	Y
School Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sample Size	536,813	178,700	358,113	526,652	175,348	351,304	719,092	238,329	480,763

Notes: The sample is all first-time, public school Michigan eleventh graders in years 2004-2009, conditional on reaching spring of eleventh grade. Columns (1)-(6) restrict the sample to the set of 552 matched high schools that comprise the sample for Tables 6-11. In columns (4) - (9) the 2% of the sample with missing home address are excluded. In these columns each point estimate is from a separate linear probability model, difference-in-difference regression where the treatment (control) group is students living further (closer) than 5 driving miles from their home to the nearest pre-policy ACT center. Standard errors in parentheses are clustered at the school level. Control means are in italics. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Table 1.A.1. Coefficients from Pre-Policy ACT-Taking Prediction Equation

	Dep Var = Take ACT
Free Lunch	-0.221***
	(0.006)
Female	0.084***
	(0.003)
Black	-0.071***
	(0.009)
Hispanic	-0.118***
	(0.010)
Other Race	0.012
	(0.009)
Lunch*Female	-0.001
	(0.004)
Lunch*Black	0.124***
	(0.007)
Lunch*Hispanic	0.091***
	(0.014)
Female*Black	0.003
	(0.006)
Female*Hispanic	-0.009
	(0.009)
Test of Equality of School FE's	
F-Stat	11,326
P-Value	0.000
R-Squared	0.222
Sample Size	354,923

Notes: The sample is as in Tables 6 - 9, except excludes the approximately 5,000 eleventh graders in spring 2006 who took the mandatory ACT in its pilot phase. Free lunch status measured as of eleventh grade. Several other student level variables and interactions as well as school- and district-year level covariates interacted with the student-level covariates are included in the regression but excluded from this table. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Table 1.A.2. Replicating Goodman (2012) Effect of Mandatory-ACT Policy in CO and IL on Postsecondary Enrollment Using IPEDS

Dependent Variable:	ln(Overall Enrollment)		ln(Selective Enrollment)								ln(4-Year Enrollment)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Goodman (2012)	0.054	0.032	0.159***	0.138**	0.111***	0.102***	0.142***	0.128***	0.043	NA	NA	NA
	(0.142)	(0.138)	(0.057)	(0.053)	(0.034)	(0.033)	(0.018)	(0.021)	(0.044)	NA	NA	NA
Replication	0.054	0.032	0.164**	0.132**	0.156**	0.127**	0.144***	0.121***	0.028	0.030	0.064**	0.011
	(0.142)	(0.138)	(0.063)	(0.057)	(0.061)	(0.057)	(0.019)	(0.022)	(0.041)	(0.043)	(0.026)	(0.022)
Replication using additional covariates		-0.039		0.027		0.027		0.035*		0.013	-0.013	
		(0.103)		(0.031)		(0.031)		(0.020)		(0.040)	(0.028)	
Replication using first-time, *immediate* enrollment	0.100	0.055	0.157**	0.109**	0.149**	0.104**	0.088**	0.072**	-0.039	-0.044	0.048	-0.040*
	(0.123)	(0.110)	(0.058)	(0.047)	(0.056)	(0.047)	(0.032)	(0.032)	(0.029)	(0.030)	(0.040)	(0.020)
Control States	22	22	22	22	22	22	22	22	22	22	22	22
N (number of state-years)	192	192	192	192	192	192	192	192	192	192	192	192
Includes Covariates	N	Y	N	Y	N	Y	N	Y	N	Y	Y	N
Includes ln(total enrollment)	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y
Includes ln(population)	N	N	N	N	Y	Y	N	N	N	N	N	N
State-Specific Time Trend	N	N	N	N	N	N	N	N	Y	Y	N	Y

Notes: The level of observation is the state-year. The sample is the twenty-four states where the majority of students take the ACT, excluding Michigan. Each cell reports a difference-in-difference coefficient from a separate regression comparing Colorado and Illinois to the remaining states. Covariates included in Goodman (2012) are unemployment rate, fraction of adults with a B.A., fraction poor, and fraction minority. "Additional covariates" include median household income, average tuition at a public four-year university, K-12 expenditure per student, and lagged number of public school eleventh graders. All postsecondary enrollment is first-time enrollment. Immediate enrollment is first-time enrollment of students who graduated high school within the past twelve months. Following Goodman (2012), standard errors in parentheses are clustered at the state level except for when state-specific time trends are included, in which case errors are corrected for heteroskedasticity but not clustered. *** = significant at the 10% level, ** = 5% level, * = 1% level.

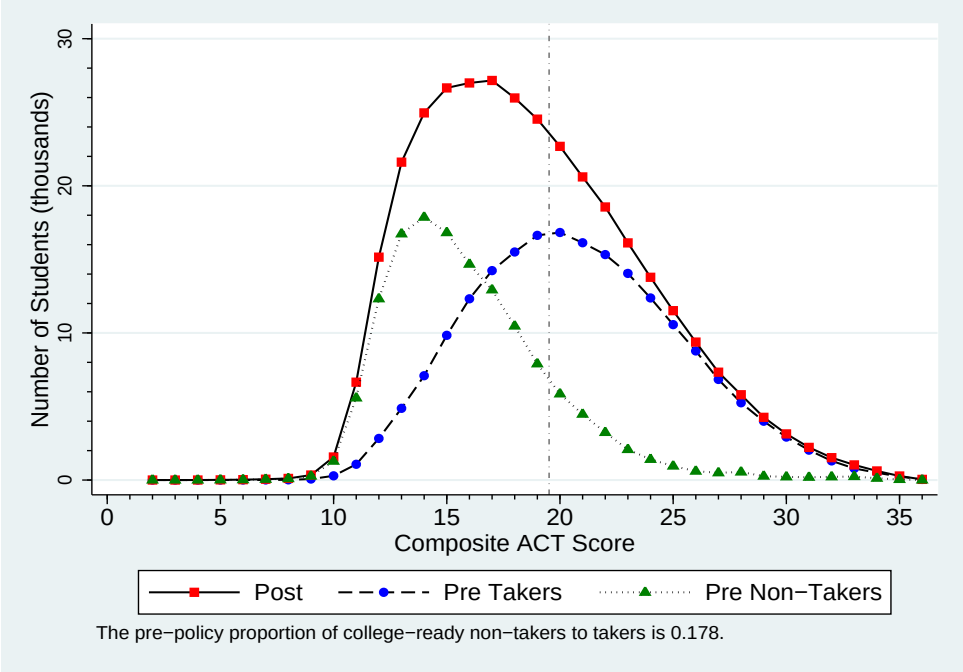
Table 1.A.3. Effect of Mandatory-ACT Policy in CO and IL on Selective Enrollment Using IPEDS

	Goodman (2012)		Replication		Replication Using Immediate Enrollment	
	(1)	(2)	(3)	(4)	(5)	(6)
$Treatment_s$*(<4 years of policy)$_t$	0.095***	0.065**	0.092***	0.050*	0.044	-0.025
	(0.014)	(0.029)	(0.014)	(0.028)	(0.030)	(0.025)
$Treatment_s$*(>=4 years of policy)$_t$	0.190***	0.140***	0.195***	0.122***	0.133***	0.020
	(0.029)	(0.034)	(0.030)	(0.034)	(0.035)	(0.031)
Joint Significance Test						
F-Statistic	30.962	9.9425	29.559	7.233	19.131	3.452
P-Value	0.000	0.000	0.000	0.001	0.000	0.034
Control States	22	22	22	22	22	22
N (num. of state-years)	192	192	192	192	192	192
Includes ln(total enrollment)	Y	Y	Y	Y	Y	Y
State-Specific Time Trend	N	Y	N	Y	N	Y

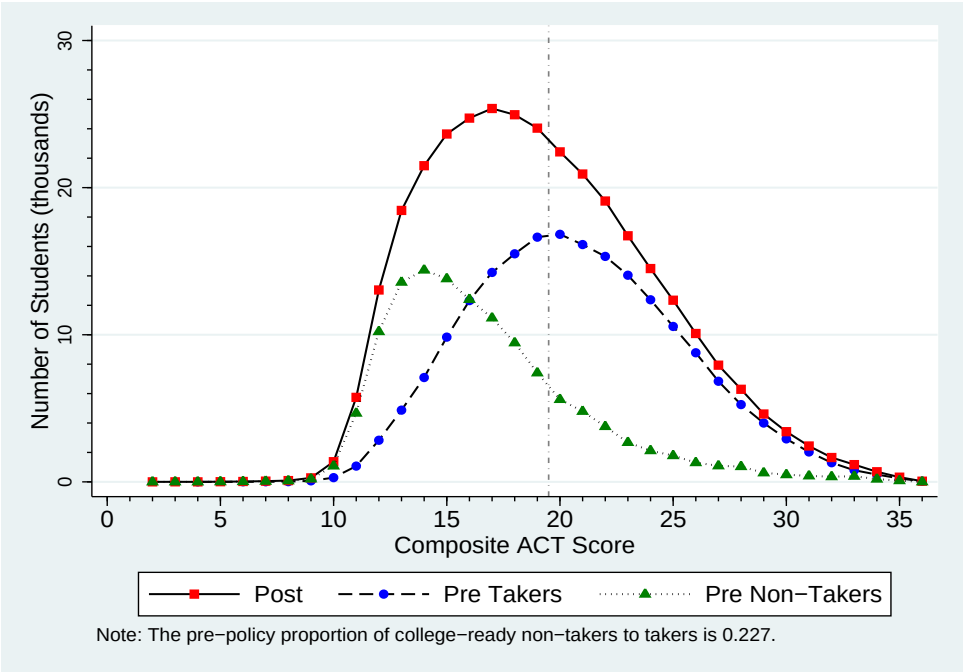
Notes: The level of observation is the state-year. The sample is the twenty-four states where the majority of students take the ACT, excluding Michigan. All postsecondary enrollment is first-time enrollment. Immediate enrollment is first-time enrollment of students who graduated high school within past twelve months. *** = significant at the 10% level, ** = 5% level, * = 1% level.

Figure 1. 1: ACT Score Distributions Pre- and Post-Mandatory ACT Policy in Michigan

(a) Unadjusted



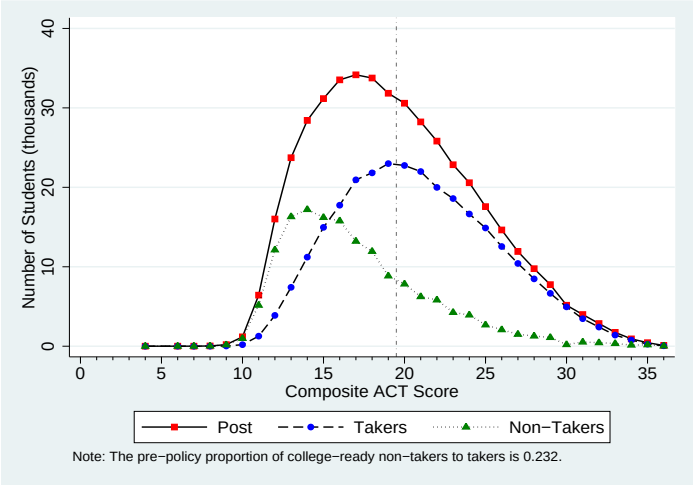
(b) Adjusted for Cohort Size and Composition



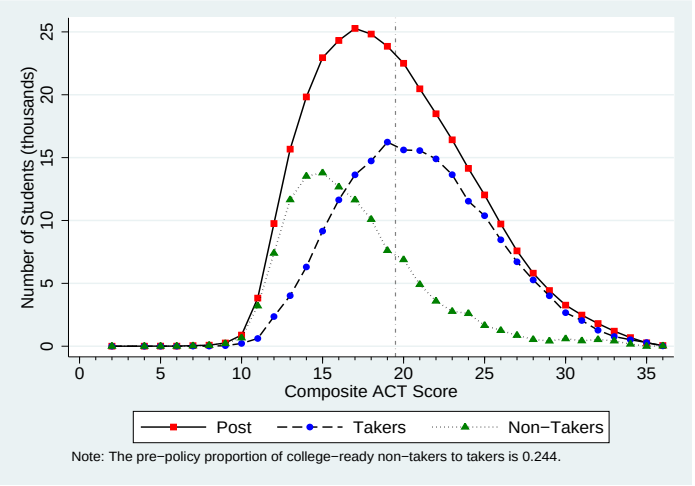
Notes: Figure (a) shows the distribution of ACT scores post-policy, pre-policy, and the difference, which is the latent score distribution among non-takers in the pre-period. Figure (b) reweights the post-policy sample to resemble the pre-period sample according to observed characteristics following DiNardo, Fortin, and Lemieux (1996). The weights are normed so the pre- and post-policy samples are of equal size.

Figure 1. 2: Diagnosing External Validity Using National ACT Micro-Data

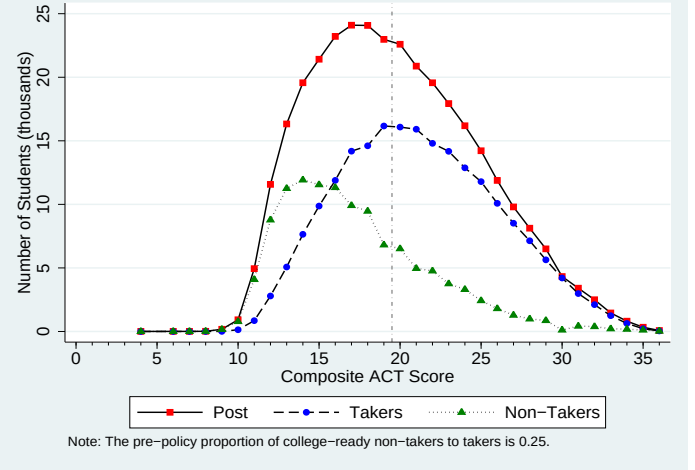
(a) Illinois and Kentucky



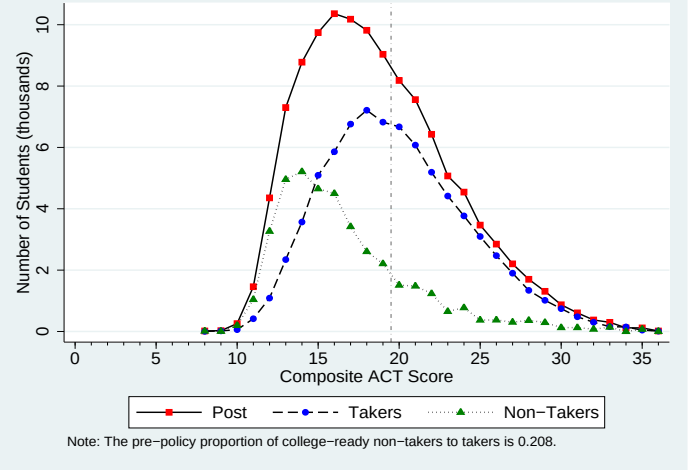
(b) Michigan



(c) Illinois



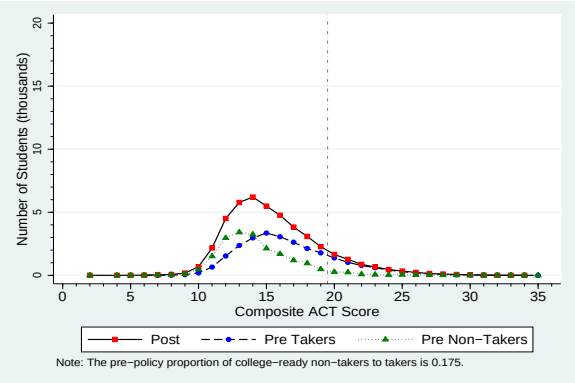
(d) Kentucky



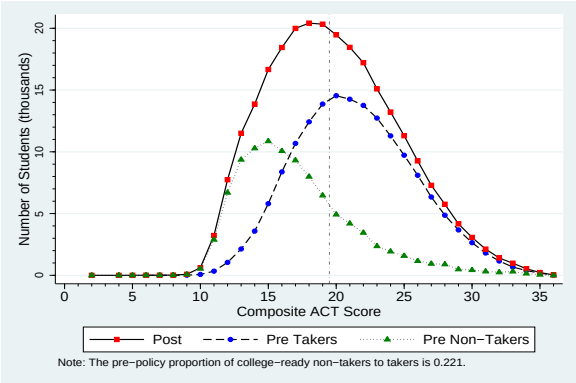
Notes: These figures use a one-in-four sample of all ACT-takers three years before and after adoption of mandatory ACT exams in Michigan and two other mandatory ACT states. Panel (a) shows the distribution of observed ACT scores pre- and post-policy, and of latent scores pre-policy for Illinois and Kentucky combined. Panel (b) shows the same figure for Michigan. Panels (c) and (d) show Illinois and Kentucky separately. Post-policy cohorts are weighted to resemble pre-policy cohorts according to observed characteristics. Weights are normed so the pre- and post-policy samples are of equal size.

Figure 1. 3: Observed and Latent ACT Scores by Subgroup

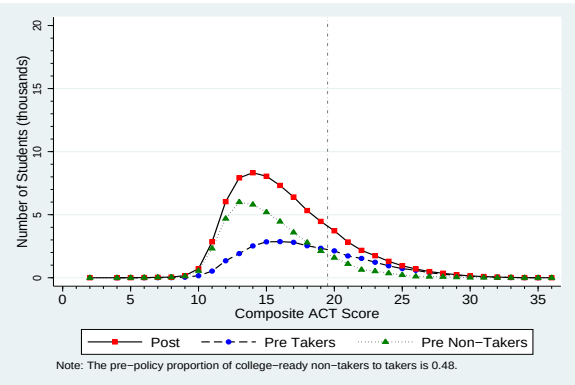
(a) Black



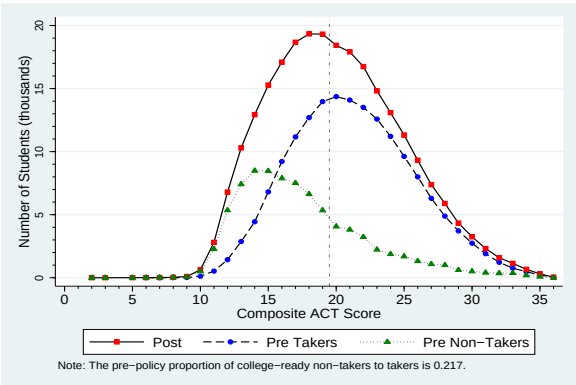
(b) White



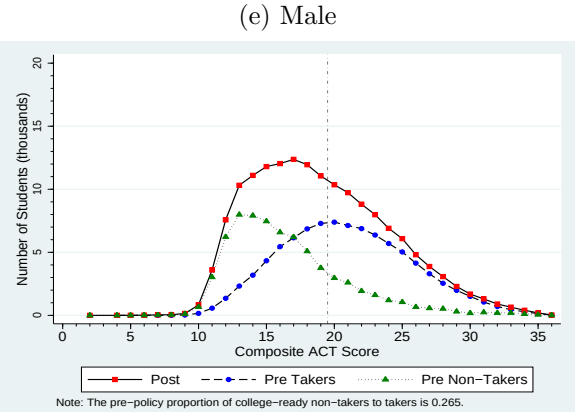
(c) Poor



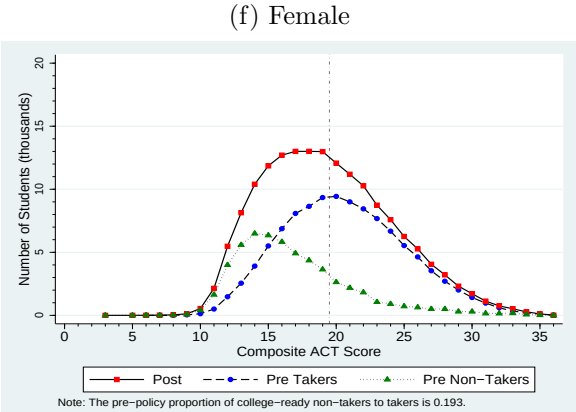
(d) Non-Poor



(e) Male



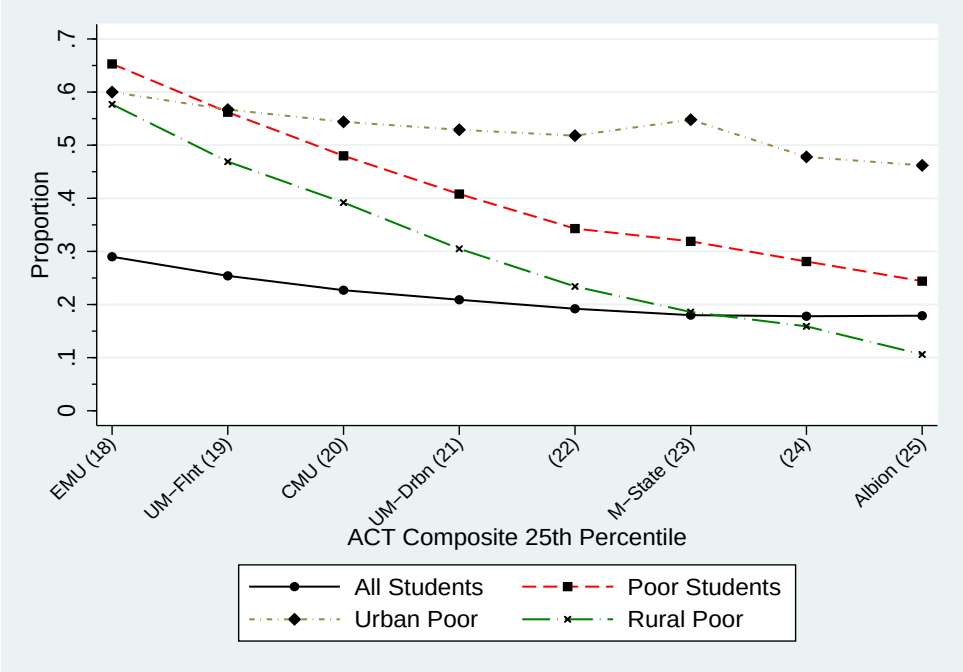
(f) Female



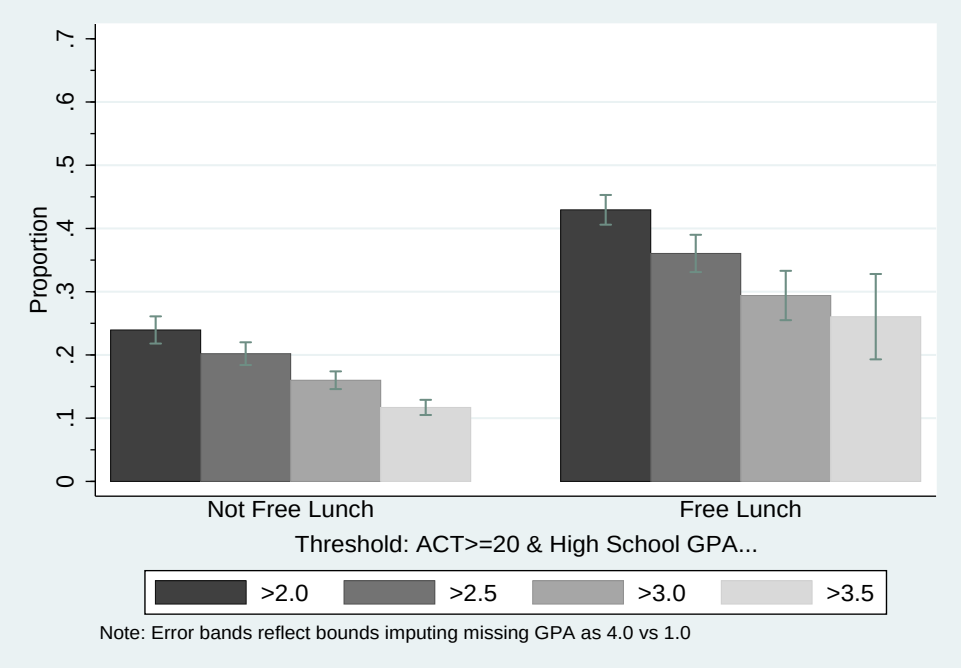
Notes: Figures show the distribution of ACT scores post-policy, pre-policy, and the difference, which is the latent score distribution among non-takers in the pre-period. All figures reweight the post-policy sample to resemble the pre-period sample according to observed characteristics following DiNardo, Fortin, and Lemieux (1996). The weights are normed so the pre- and post-policy samples are of equal size.

Figure 1. 4: Proportion of College-Ready Non-Takers to Takers, by ACT Score and GPA Threshold

(a) By ACT Score College-Readiness Threshold



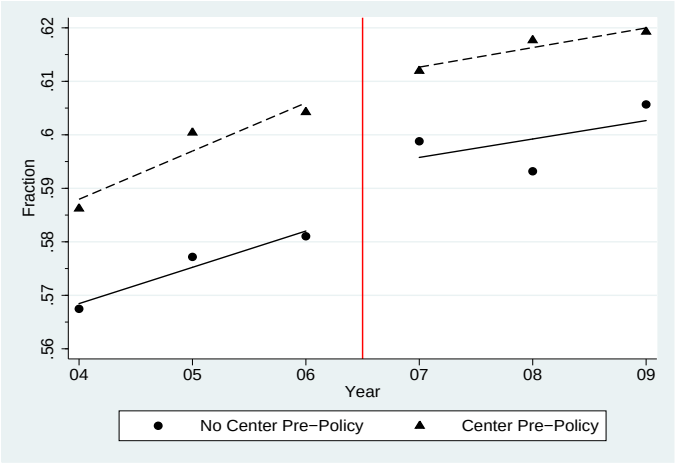
(b) By High School GPA College-Readiness Threshold



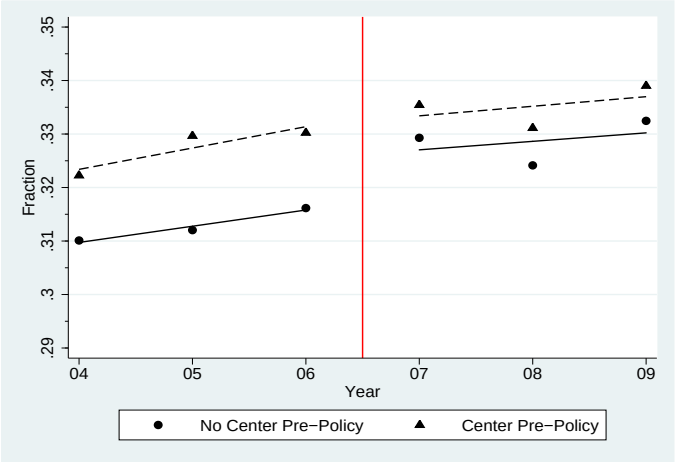
Notes: Panel (a) plots the pre-reform proportion of college-ready students not taking the ACT to those who take the ACT by the ACT score threshold used to calculate college-readiness. The 0.23 among all students at an ACT score of 20 corresponds to the earlier finding that for every 100 students scoring at or above a 20 on the ACT, there are another 23 students who would score as such but do not take the test. Panel (b) uses an ACT score of 20 threshold, and shows the proportion after further restricting the college-readiness threshold to include a high school GPA cutoff.

Figure 1. 5: College Enrollment by Cohort and Pre-Policy Test Center Status

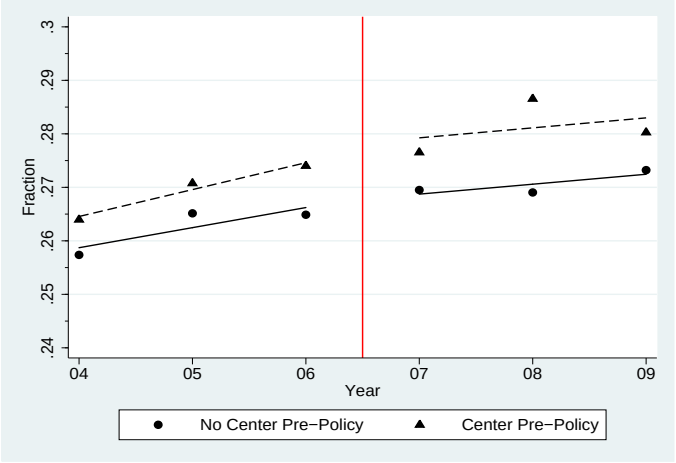
(a) Any College



(b) Four-Year College



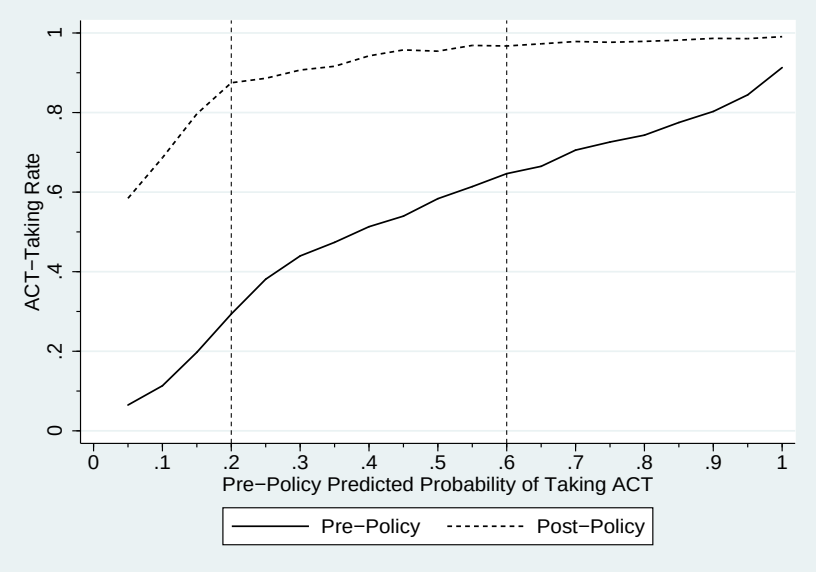
(c) Two-Year College



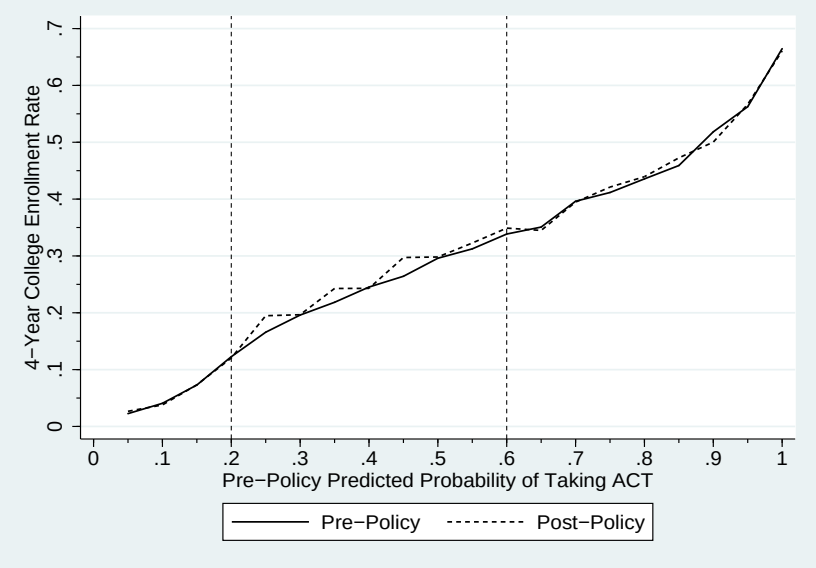
Notes: Figures show college enrollment pre- and post-mandatory ACT by whether a not a student attends a school with an ACT test center pre-mandatory ACT. The sample is restricted to the propensity score matched sample of high schools. Figure (a) includes any college enrollment, figure (b) includes four-year enrollment only, and figure (c) includes two-year enrollment.

Figure 1. 6: ACT-Taking and College Enrollment by Predicted Probability of ACT-Taking

(a) ACT-Taking

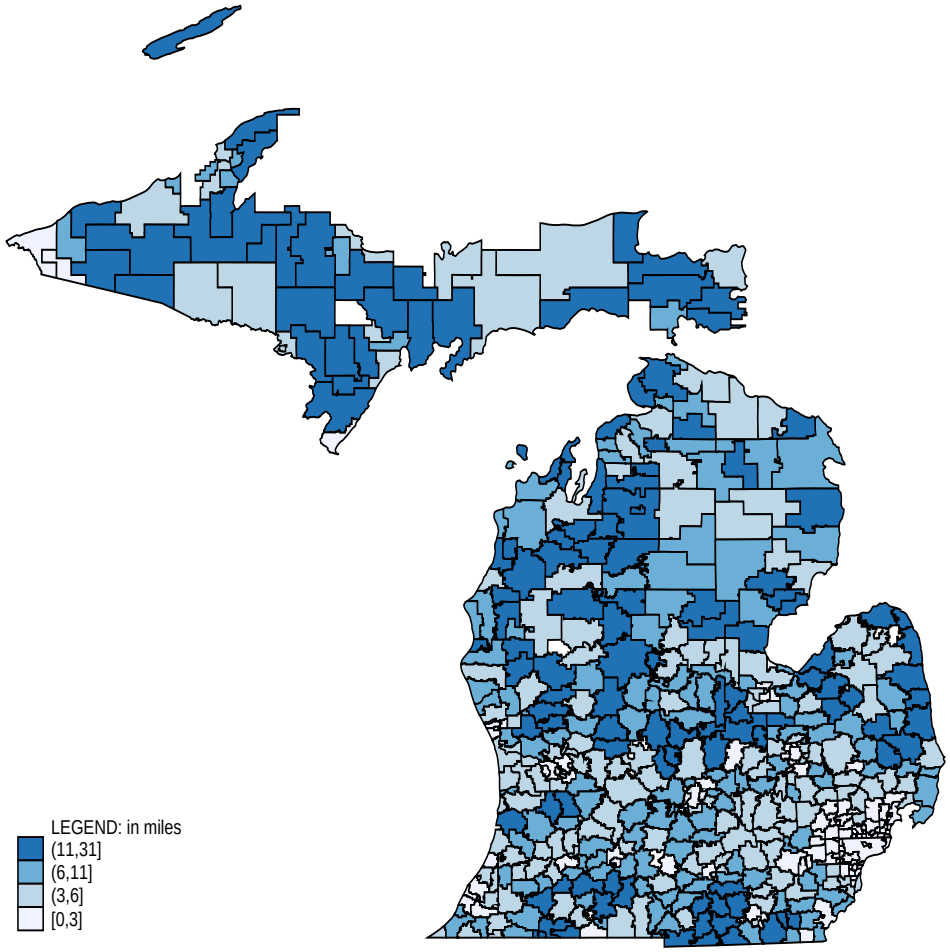


(b) 4-Year College Enrollment (Regression Adjusted)



Notes: Figure (a) plots the ACT-taking rate pre and post mandatory ACT at twenty quantiles of the predicted probability that a student would take the ACT based on pre-reform observed characteristics and ACT-taking. Figure (b) plots the raw, pre-policy four-year enrollment rate among students in the matched sample of high schools (solid line) at these same twenty quantiles. It then adds to this line the difference-in-difference four-year enrollment effect of the policy (dashed line). Note the smaller scale of the Y-axis in figure (b) designed to more clearly show the difference between the two lines.

Figure 1.A. 1: Distance to Nearest ACT Center by District, Pre-Policy



Notes: Figure plots school-district averages of the student-level driving distance from home to the nearest ACT test center pre-policy for students attending high school in the district.

CHAPTER II

Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion

Abstract

This paper examines the effect of early childhood investments on college enrollment and degree completion. We use the random assignment in the Project STAR experiment to estimate the effect of smaller classes in primary school on college entry, college choice, and degree completion. We improve on existing work in this area with unusually detailed data on college enrollment spells and the previously unexplored outcome of college degree completion. We find that assignment to a small class increases the probability of attending college by 2.7 percentage points, with effects more than twice as large among blacks. Among students enrolled in the poorest third of schools, the effect is 7.3 percentage points. Smaller classes increase the likelihood of earning a college degree by 1.6 percentage points and shift students towards high-earning fields such as STEM (science, technology, engineering and mathematics), business and economics. We find that test score effects at the time of the experiment are an excellent predictor of long-term improvements in postsecondary outcomes.

2.1. Introduction

Education is intended to pay off over a lifetime. Economists conceive of education as a form of "human capital," requiring costly investments in the present but promising a stream of returns in the future. Looking backward at a number of education interventions (e.g., Head Start, compulsory schooling), researchers have identified causal links between these policies and long-term outcomes such as adult educational attainment, employment, earnings, health and civic engagement (Ludwig & Miller, 2007; Deming, 2009; Angrist & Krueger, 1991; Dee, 2004; Lleras-Muney, 2005). But decision-makers attempting to gauge the effectiveness of current education inputs, policies and practices in the present cannot wait decades for these long-term effects to emerge. They therefore rely upon short-term outcomes – primarily standardized test scores – as their yardstick of success.

A critical question is the extent to which short-term improvements in test scores

translate into long-term improvements in well-being. Puzzling results from several evaluations make this a salient question. Three small-scale, intensive preschool experiments produced large effects on contemporaneous test scores that quickly faded (Schweinhart, et al., 2005; Anderson, 2008). Quasi-experimental evaluations of Head Start, a preschool program for poor children, reveal a similar pattern, with test score effects gone by middle school. In each of these studies, treatment effects have re-emerged in adulthood as increased educational attainment, enhanced labor market attachment, and reduced crime (Deming, 2009; Garces, Thomas, & Currie, 2002; Ludwig & Miller, 2007). Further, several recent papers have shown large impacts of charter schools on test scores of disadvantaged children (Abdulkadiroglu, et al., 2011; Angrist, et al., 2012; Dobbie & Fryer, 2011). A critical question is whether these effects on test scores will persist in the form of long-term enhancements to human capital and well-being.

We examine the effect of smaller classes on educational attainment in adulthood, including college attendance, degree completion and field of study. We exploit random variation in class size in the early grades of elementary school created by the Tennessee Student/Teacher Achievement Ratio (STAR) Experiment. Participants in the STAR experiment are now in their thirties, an age at which it is plausible to measure completed education. Our postsecondary outcome data is obtained from the National Student Clearinghouse (NSC), a national database that covers approximately 90 percent of students enrolled in colleges in the U.S.

We find that being assigned to a small class increases the rate of postsecondary attendance by 2.7 percentage points. The effects are considerably higher among populations with traditionally low rates of postsecondary attainment. For Black students and students eligible for free lunch the effects are 5.8 and 4.4 percentage points, respectively. At elementary schools with the greatest concentration of poverty, measured using the fraction of students receiving a subsidized lunch, smaller classes increase the rate of postsecondary attendance by 7.3 percentage points. We further find that being assigned to a small class increases the probability of earning a college degree by 1.6 percentage points. Smaller classes shift students toward earning degrees in high-earning fields such as science, technology, engineering and mathematics (STEM), business and economics.

Our results shed light on the relationship between the short-and long-term effects

of educational interventions. The short-term effect of small classes on test scores, it turns out, is an excellent predictor of its long-term effect on adult outcomes. We show this by adding K-3 test scores to our identifying equation; the coefficient on the class size dummy drops to zero. The coefficient on the interaction of class size and test scores is also zero, indicating that the scores of children in small classes are no less (or more) predictive of adult educational attainment than those of children in the regular classes.

Our analysis identifies the effect of manipulating a single policy-relevant educational input on adult educational attainment. By contrast, the early-childhood interventions for which researchers have identified lifetime effects (e.g., Head Start, Abecederian) are multi-pronged, including home visits, parental coaching and vaccinations in addition to time in a preschool classroom. We cannot distinguish which dimensions of these treatments generate short-term effects on test scores, and whether they differ from the dimensions that generate long-term effects on adult well-being. The effective dimensions of the treatment are also ambiguous in the recent literature on classroom and teacher effects. For example, Chetty et al. (2011) show very large effects of kindergarten classroom assignment on adult well-being. In those estimates, the variation in classroom quality that produces significant variation in adult outcomes excludes class size but includes anything else that varies at the classroom level, including teacher quality and peer quality, both of which are extremely difficult to manipulate with policy. By contrast, the effects we measure in this paper, both short-term and long-term, can be attributed to a well-defined and replicable intervention: reduced class size.

2.2 The Tennessee STAR Experiment

The Tennessee Student/Teacher Achievement Ratio (STAR) Experiment randomly assigned class sizes to children in kindergarten through third grade. The experiment was initiated in the 1985-86 school year, when participants were in kindergarten. A total of 79 schools in 42 school districts participated, with over-sampling of urban schools. An eventual 11,571 students were involved in the experiment. The sample is 60 percent white and the balance African American. About 60 percent of students were eligible for subsidized lunch during the experiment. The experiment is described in greater detail elsewhere (Word, et al., 1990; Folger & Breda, 1989; Finn & Achilles, 1990; Krueger, 1999; and Achilles, 1999.)

Children in the STAR experiment were assigned to either a small class (target size of 13 to 17 students) or regular class (22 to 25 students).[1] Students who entered a participating school after kindergarten were randomly assigned during those entry waves to a regular or small class. Teachers were also randomly assigned to small or regular classes. All randomization occurred within schools.

Documentation of initial random assignment in STAR is incomplete (Krueger, 1999). Krueger (1999) examines records from 18 STAR schools for which assignment records are available. He finds that, as of entry into STAR, 99.7 percent of students were enrolled in the experimental arm to which they were initially assigned. Krueger's approach, and that of the subsequent literature, is to assume that the class type in which a student is first enrolled is the class type to which she was assigned. We follow that convention in our analysis.

Numerous papers have tested, and generally validated, the randomization in STAR (Krueger, 1999). There are no baseline outcome data (e.g., a pre-test) available for the STAR sample. On the handful of covariates available in the STAR data (free lunch eligibility, race, sex), the arms of the experiment appear balanced at baseline (see Table 2.1 for a replication of these results). Recent work by Chetty et al. (2011) shows that the STAR entry waves were balanced at baseline on a detailed set of characteristics (e.g., family income, home ownership) contained in the income tax returns of the STAR subjects' parents.

2.3 Previous Research on the Long-Term Effects of Small Classes

A substantial body of research has examined the effect of Project STAR on short-and medium-run outcomes. We do not comprehensively discuss this literature but instead summarize the pattern of findings. These papers show that students assigned to a small class experience contemporaneous test score gains of about a fifth of a standard deviation. These test score results diminish after the experiment ends in third grade.[2]

[1] A third arm of the experiment assigned a full-time teacher's aide to regular classes. Previous research has shown no difference in outcomes between the regular-sized classes with and without an aide. We follow the previous literature in pooling students from both types of regular classes into a single control group. The results are substantively unchanged if we include an indicator variable for the presence of a full-time teacher's aide.

[2] Cascio and Staiger (2012) show that fade-out of test-score effects is, at least in some settings, a statistical artifact of methods used by analysts to normalize scores within and across grades. However, they

There is evidence of lasting effects on other dimensions. Krueger and Whitmore (2001) show that students assigned to small classes are more likely to take the ACT and SAT, required for admission to most four-year colleges. Schanzenbach (2006) reports that smaller classes reduce the rate of teen pregnancy among female participants by about a third. In addition, Fredriksson, Ockert, and Oosterbeek (2013) find positive long-term impacts of reduced class size in grades 4-6 in Sweden on educational attainment and wages.

The paper most closely related to our own examined the impact of Project STAR on adult outcomes using the income tax records of STAR participants and their parents (Chetty et al., 2011). That paper emphasizes the differential long-term impacts of being randomly assigned to classrooms of different "quality" levels stemming from higher-quality teachers and/or classmates, after accounting for class size. Chetty et al. (2011) document the sizeable long-term payoff to having a high quality classroom, though recognize that this cannot be directly manipulated by public policy. By contrast, we focus on the long-term impacts of randomly assigned class size, which is an easily measured input that can be manipulated by policy.

2.4. Empirical Strategy

The experimental nature of Project STAR motivates the use of a straightforward empirical specification. We compare outcomes of students randomly assigned to small and regular classes by estimating the following equation using Ordinary Least Squares:

$$y_{isg} = \beta_0 + \beta_1 SMALL_{is} + \beta_2 X_{is} + \beta_{sg} + \varepsilon_{isg}\,, \qquad (1)$$

where y_{isg} represents a postsecondary schooling outcome of student i, who entered the STAR experiment in school s and in grade g. X is a vector of covariates including sex, race and free lunch status (an indicator for whether the student ever received free or reduced price lunch during the experiment), included to increase precision. β_{sg} is a set of school-by-entry-grade fixed effects. We include these because students who entered STAR schools after kindergarten were randomly assigned at that time to small or regular classes. The variable of interest is $SMALL_{is}$, an indicator set to one if student i was assigned to a small class upon entering the experiment. The omitted group to which small classes are compared is regular classes (with or without a teacher's aide). We cluster

specifically note that the sharp drop in estimated effects that occurs after the end of the STAR experiment cannot be explained in this way.

standard errors by school, the most conservative approach. Standard errors are about ten percent smaller if we cluster at the level of school-by-wave.

2.5. Data

We use the original data from the STAR experiment, which includes information on the type of class in which a student is enrolled, basic demographics (race, poverty status, sex), school identifiers, and standardized test scores. These data also include the name and date of birth of the student, which we use to match to data on postsecondary attainment and completion.

Data on postsecondary outcomes for the STAR sample come from the National Student Clearinghouse (NSC). NSC is a non-profit organization that was founded to assist student loan companies in validating students' college enrollment. Borrowers can defer payments on most student loans while in college, which makes lenders quite interested in tracking enrollment. Colleges submit enrollment data to NSC several times each academic year, reporting whether a student is enrolled, at what school, and at what intensity (e.g., part-time or full-time). NSC also records degree completion and the field in which the degree is earned. States and school districts use NSC data to track the educational attainment of their high school graduates (Roderick, Nagaoka, & Allensworth, 2006). Recent academic papers making use of NSC data include Deming et al. (2011) and Bettinger et al. (2012).

With the permission of the Project STAR researchers and the state of Tennessee, we submitted the STAR sample to the NSC in 2006 and again in 2010.[3] The STAR sample was scheduled to graduate high school in 1998. We therefore capture college enrollment and degree completion for twelve years after on-time high-school graduation, when the STAR sample is about 30 years old.

The NSC matches individuals to its data using name and date of birth. If birth date is missing, the NSC attempts to match on name alone. Some students in the STAR sample are missing identifying information used in the NSC match: 12 percent have incomplete name or birthdate. In our data, a student that attends college but fails to produce a match in the NSC database is indistinguishable from a student who did not attend college. If the absence of these identifiers is correlated with the treatment, then our

[3] In 2006, the NSC used social security number as well as name and date of birth in its matches. As of 2010, NSC had ceased to use social security number for its matches.

estimates may be biased. To determine whether identifiers are missing at a differential rate across treatment groups, we estimate equation (1) replacing y_{isg} with an indicator variable equaling one if a student has a missing name or date of birth. We find a precisely estimated zero for β_1 (=-0.008, SE=0.008) indicating that the probability of missing identifying information is uncorrelated with initial assignment. In the concluding section of the paper, we present the results of a second test exploring the possible bias in our main result associated with missing identifiers.

Not all schools participate in NSC; the company estimates they currently capture about 93 percent of undergraduate enrollment nationwide. During the late 1990s, when the STAR subjects would have been graduating from high school, the NSC included colleges enrolling about 80% of undergraduates in Tennessee (Dynarski, Hemelt, & Hyman, 2012).[4] Since we miss about 20% of undergraduate enrollment using the NSC data, we expect that we will underestimate the college attendance rate of the STAR sample by about a fifth. The NSC data indicate that 39.4 percent of the STAR sample had attended college by age 30. Among those born in Tennessee in the same years as the STAR sample, the attendance rate is 52.8 percent in the 2005 American Community Survey (Ruggles, et al., 2010).[5] Our NSC estimate of college attendance is therefore, as expected, about four-fifths of the magnitude of the ACS estimate.

In the NSC, we find that 15.1 percent of the STAR sample has earned a college degree. This is substantially lower than the corresponding rate we calculate from the 2005 American Community Survey (29.3 percent). Not all of the colleges that report enrollment to the NSC report degree receipt, and this explains at least part of the discrepancy.[6]

The exclusion of some colleges from NSC will induce measurement error in the dependent variable. If this error is not correlated with treatment (i.e., classical measurement error) then the true effect of class size on college enrollment will be larger

[4] Dynarski et al. (2012) calculate this rate by dividing undergraduate enrollment at Tennessee colleges included in NSC as of 1998 by enrollment at all Tennessee colleges in 1998. The list of colleges participating in the NSC and the year that they joined is accessible on the NSC website. Enrollment data are from the Integrated Postsecondary Education Data System (IPEDS), a federally-generated database that lists every college, university and technical or vocational school that participates in the federal financial aid programs (about 6,700 institutions nationwide) (National Center For Education Statistics, 2010).

[5] We re-weight the Tennessee-born in the ACS data to match the racial composition of the STAR sample, which was disproportionately black.

[6] Using IPEDS, we calculate that 70% of undergraduate degrees are conferred by institutions that, according to the NSC website, report degrees to NSC. Dynarski et al. (2012) also find lower degree coverage in the NSC relative to enrollment coverage.

than our observed effect by the proportion of enrollment that is missed (approximately 20 percent).[7] This is because the true treatment effect is the sum of the observed treatment effect and the treatment effect of the unobserved college attenders (Bound, Brown, & Mathiowetz, 2001). However, if the measurement error in college attendance is correlated with assignment to treatment then our effect could be either downward or upward biased. This would be the case, for example, if colleges attended by marginal students are disproportionately undercounted by NSC.

To determine whether the NSC systematically misses certain types of schools, we compare the schools that participate in NSC with those in IPEDS. Along all measures we examined (i.e., sector, racial composition, selectivity), the NSC colleges are similar to the universe of IPEDS colleges, with a single exception: NSC tends to exclude for-profit institutions.[8] These are primarily trade schools such as automotive, technology, business, nursing, culinary arts and beauty schools. If small classes tend to induce into such schools those students who would not otherwise attend college, we will underestimate the effect of small classes on college attendance. If on the other hand small classes induce students out of such schools into colleges that we tend to observe, such as community colleges, then our estimates will be upward biased. In the concluding section of our paper, we conduct a back-of-the-envelope exercise to bound the possible upward bias that could be due to this phenomenon.

2.6. Results

In this section, we examine the effect of assignment to a small class on a set of postsecondary outcomes: college entry, the timing of college entry, college choice, degree receipt and field of degree.

2.6.1. College Entry

In Table 2.2, we estimate the effect of assignment to a small class on the probability of college entry by age 30. The effect is close to three percentage points (Column 1, 2.8 percentage points), which is an impact of approximately 7 percent relative to the control

[7] This is true in terms of percentage points. The percent increase in college attendance would remain unchanged.

[8] The conclusion is the same when we weight coverage by the number of degrees conferred rather than by undergraduate enrollment.

mean of 38.5 percent (control means are italicized in the tables). This estimate is statistically significant, with a standard error of about one percentage point. Including covariates does not alter the estimate, as is expected with random assignment. For the balance of the paper we report results that include covariates, since they are slightly more precise.

Splitting the sample by race reveals that the effects are concentrated among Blacks (5.8 points relative to a mean of 30.8 percent) and those eligible for free or reduced-price lunch (4.4 points relative to a mean of 27.2 percent). The effects are twice as large for boys (3.2 points relative to a mean of 32.4 percent) than girls (1.6 points relative to a mean of 45.5 percent). Breaking the effects down yet more finely shows that the effects are largest for Black females (7.2 points, standard error of 3.5), with no effect on white females (1.3 points, standard error of 2.3). The effects for Black and white males are indistinguishable (3.1 and 4.4 points, respectively; standard error of 1.8 and 2.4 points).

One caveat to consider when examining results by race and gender is that the probability of enrolling in a college not in the NSC could be correlated with race-gender, which could cause bias in the estimates. Dynarski et al. (2012) show that NSC coverage is similar by sex, but is lower for Black students than white students. To examine this issue for a population similar to the STAR sample of students, we examine the share of first-time college students in Tennessee in 1998 in IPEDS by race and sex attending for-profits (which tend not to appear in NSC) and attending any type of college. We find that black and female students tend to enroll in higher proportions in for-profit colleges. This suggests that part of the large treatment effect for black females could be due to these students being induced from non-NSC colleges to those that participate in NSC.

Our results by student demographics indicate that there is substantial heterogeneity by race and income in the effect of class size. However, policy decisions regarding staffing levels and class size tend to be set at the school level rather than the student level. School-level characteristics, rather than student-level characteristics, may therefore be the more policy-relevant dimension along which to measure heterogeneity in effects. In order to capture this policy-relevant variation in effects, we divide the STAR schools into three groups: those with low, medium and high levels of poverty, which we proxy with the share of children eligible for a subsidized lunch. We sort students by this

share, and construct the groups such that the number of students in each group is nearly identical (see Table 2.A.1). Note that the STAR sample was disproportionately poor and urban, so even the schools with the lowest levels of poverty are relatively disadvantaged.

When we estimate Equation (1) separately for these three groups of schools, we find that the treatment effect is concentrated in the poorest schools. At schools with low to medium concentrations of poverty, the estimated effect of class size on postsecondary attainment is indistinguishable from zero (Table 2.2, Columns 7 and 8). But the estimated effect is 7.3 percentage points in the poorest schools. This is a 28 percent increase relative to the control mean in these schools. A test of the equality of the coefficients for the poorest schools versus the combined bottom two terciles is strongly rejected (p-value of 0.008, Column 11).

Inequality in postsecondary education has increased in recent decades, with the gap in attendance between those born into lower-income and higher-income families expanding (Belley & Lochner, 2007; Bailey & Dynarski, 2011). The pattern of effects described above will tend to decrease gaps in postsecondary attainment. Figure 2.1 shows this graphically. On the top is depicted the gap in college attendance between blacks and whites in regular classes (left) and in small classes (right). The black-white gap is about half as large in small classes (7.7 percentage points) as it is in regular classes (12.4 percentage points). The drastic reduction in the race gap in college attendance is driven by females, for whom the race gap virtually disappears in small classes (results not shown).

In the control group, students who were eligible for free or reduced-price lunch are 29.1 percentage points less likely to attend college than their higher-income classmates. The gap is slightly smaller in the treatment group (25.7 percentage points). Finally, we compare the effect of small classes on the gap in postsecondary outcomes between schools with high and moderate levels of poverty. Among students in regular-sized classes, the gap in postsecondary attendance is 18.1 percentage points. Among students in small classes, the gap is nearly halved, to 9.8 percentage points.

Class size could plausibly affect the intensity with which a student enrolls in college, in addition to the decision to enroll at all. The overall impact on the intensity of enrollment is theoretically ambiguous: students induced into college by smaller classes may be more likely to enroll part-time than other students, while treatment could induce

those who would have otherwise enrolled part-time to instead enroll full-time. In the control group, about three-quarters of college entrants (ever) attend college full-time, while a quarter never do (Table 2.2, second row). When we re-estimate Equation (1) with these two variables as dependent variables, we find that the effect on entry is evenly divided between part-time and full-time enrollment. While the standard errors preclude any firm conclusions, these results suggest that the marginal college student is more likely than the inframarginal student to attend college exclusively on a part-time basis.

2.6.2. Timing of College Attendance

Class size could plausibly affect the timing of postsecondary attendance. The net effect is theoretically ambiguous. Smaller classes may lead students who would otherwise have attended college to advance through high school more rapidly, enter college sooner after graduation, and move through college more quickly. On the other hand, students induced into college by smaller classes may enter and move through college at a slower pace than their inframarginal peers.

We first estimate the effect of class size upon "on-time enrollment," which we define as entering college by fall of 1999, or about 18 months after the STAR cohort is scheduled to have graduated high school. This variable captures the pace at which students complete high school, how quickly they enter college, and whether they attend college at all. By this measure, 27.4 percent of the control group has enrolled on-time, or about three-quarters of the 38.5 percent who ever attend college (Table 2.2). Assignment to a small class increases the likelihood of entering college on time by 2.4 percentage points. Among those students enrolled in the poorest third of schools, the effect is 4.7 points, a 29 percent increase relative to this group's control mean of 16 percent. These results suggest that students in smaller classes are no less likely to start college on time than control students: 72 percent of the treatment-group students who attend college do so on time, while among the control group the share of attendance that is on-time is 71 percent.

We next look at the year-by-year evolution of the effect of class size on postsecondary attainment. For each year, we plot the share of students who have ever attended college, separately for the treatment and control group (Figure 2.2, top panel). We also plot the treatment-control difference, along with its 95% confidence interval

(Figure 2.2, bottom panel). The fraction of the sample that has ever attended college rises from under 5 percent in 1997 to over 20 percent in 1998 (when students are 18). The rate rises slowly through age 30, when the share of the sample with any college experience reaches nearly 40 percent. The difference between the two groups reaches about three points by age 19 and remains at that level through age 30.[9] When we examine the shares of students who are currently enrolled in college (Figure 2.3) we see that the treatment group is more likely to be enrolled in college at every point in time, peaking at around 25 percent in 1999. Plausibly, smaller classes could have sped up college enrollment and completion, and the control group could eventually have caught up with the treatment group in its rate of college attendance. This is not what we see, however. The effect is always positive, and is largest right after high school, when the sample is 18 to 19 years old.[10]

2.6.3. College Choice

By boosting academic preparation, smaller classes in primary school may induce students to alter their college choices. For example, those who would have otherwise attended a two-year community college may instead choose to attend a four-year institution. Bowen, Chingos, and McPherson (2009) suggest that attending higher quality colleges (which provide more inputs, including better peers) is a mechanism through which students could increase their rate of degree completion.

In Table 2.3, we examine the effect of class size on college choice. Across the entire sample, we find little evidence that exposure to smaller classes shifts students toward higher-quality schools. The treatment effect is concentrated on attendance at two-year institutions. While 22 percent of the control group starts college at a two-year school, the rate is 2.5 percentage points higher in the treatment group (with a standard error of 0.9 percentage points). The effect is 6.3 percentage points among students in the poorest third of schools. We find positive but imprecise effects on the probability of ever attending a four-year college, attending college outside Tennessee, or attending a

[9] To obtain the figures, we replace the small-class indicator variable in our identifying equation with a full set of its interactions with year fixed effects. The coefficients on these interactions and their confidence intervals are plotted in the bottom panel. In the top panel, we add these interactions to the year-specific control means.

[10] This pattern of findings sheds light on the difference between our findings and those of Chetty et al. (2011). We can reconcile our findings with Chetty et al. (2011) if we censor the NSC data so that they exclude the same enrollment spells that are unobserved in their data, see Table 2.A.2.

selective college.[11]

2.6.4. Persistence and Degree Completion

While college entry has been on the rise in recent decades, the share of college entrants completing a degree is flat or declining (Bound, Lovenheim, & Turner, 2010). About half of college entrants never earn a degree. A key concern is that marginal students attending college may drop out quickly, in which case the attendance effects discussed above would overestimate the effect of class size on social welfare.

We explore this issue by examining the effect of small classes on the number of semesters that students attend college, as well as on the probability that they complete a college degree. Overall, the number of semesters attempted (including zeroes) is quite low: the control group attempts an average of three semesters by age 30. Among those in the control group with any college experience, the average number of semesters attempted is eight.

The treatment group spends 0.22 more semesters in college than the control group (Figure 2.4, top; Table 2.4). The effects are somewhat larger among students in the poorest schools (coefficient of 0.32), though the effect is imprecisely estimated and the difference across terciles is less stark than with the college entry effects. The size of these effects is comparable to treatment effects found in the Opening Doors demonstration, which gave short-term rewards to community college students for achieving certain enrollment and grade thresholds (Barrow, et al., 2009).

Assignment to a small class increases the likelihood of completing a college degree by 1.6 percentage points (Table 2.4); the result is statistically significant at the 10 percent level. When we examine effects separately by highest degree earned, we find that the 1.6 percentage point effect is driven evenly by increases in 2-year (associates) and 4--year (bachelors) degree receipt (0.7 and 0.9 percentage points, respectively). When we turn to the timing of degree completion, we see that there is a positive treatment effect at every age. The difference is largest between age 22 and 23 (Figure 2.4, Panel C). Students assigned to small classes during childhood continue to outpace their peers in their rate of degree completion well into their late twenties. This may explain why Chetty

[11] We measure selectivity using Barron's quality categories. Using an index that includes multiple proxies for quality such as the acceptance rate, tuition, and the average ACT/SAT score of entering students provides similar results.

et al. (2011) do not find an effect of small classes on earnings, which they observe at age 27. Members of the treatment group are still attending and completing college at this age, and so have likely not yet spent enough time in the labor market for their increased education to offset experience forgone while in college.

2.6.5. Field of Degree

The earnings of college graduates vary considerably by field. In particular, those who study science, technology, engineering and mathematics (STEM), as well as business and economics, enjoy higher returns than other college graduates (Arcidiacono, 2004; Hamermesh & Donald, 2008). In this section we examine whether class size affects the field in which a student completes a degree.[12]

We divide degrees into three categories: 1) STEM fields; business and economics concentrations; and all others.[13] Students can earn more than one degree (e.g., an AA and a BA); we code them as having a STEM degree if any degree falls in this category, and as having a business or economics degree if any degree falls in this category and they have not earned a STEM degree. In practice, very few students earn both a STEM and business or economics degree.

Assignment to a small class shifts the composition of degrees toward STEM, business and economics. While 1.9 (2.6) percent of the control group earns a degree in a STEM (business or economics) field, the rate is 2.4 (3.3) in the treatment group (Table 2.4). However, these estimates are imprecisely estimated. In order to increase precision and to group fields by whether or not they are high-paying, we combine the STEM, business and economics fields into one category. Assignment to a small class increases degree receipt in these high-paying fields by 1.3 percentage points. This difference is statistically significant at the 5 percent level, with a standard error of 0.6 percentage points. There is no difference in the rate at which students receive degrees in other fields.

These results are consistent with two scenarios: (1) those induced into completing a degree tend to concentrate in STEM, business and economics or (2) inframarginal

[12] Field of study is available only for students who complete a degree; we are therefore unable to examine the field of study for non-completers.

[13] We follow a degree-coding scheme defined by the National Science Foundation (National Science Foundation, 2011). We apply this scheme to two text fields included in NSC: degree title (e.g., "associates" or "bachelor of science") and college major (e.g., "biology"). A small number of students who receive a degree are missing both degree title and college major, and are excluded from this analysis.

degree completers are shifted toward STEM, business and economics. While we cannot conclusively identify those who are and are not on the margin of completing a degree, our analysis by school-level poverty tercile (Table 2.4, Columns 2 and 3) suggests that the second scenario is at work. The effect of small classes on graduating in a STEM, business or economics degree is 1.9 percentage points (standard error of 0.8 points) among the less poor schools where students are more likely to be inframarginal degree completers. The effect is zero among the poorest third of schools, where students are more likely to be induced into completing a degree. These effects are statistically different from one another at the 10 percent level.

2.6.6. Testing for Sources of Heterogeneity in Effects

One interpretation of these results is that the groups with the lowest control means are most sensitive to class size. An alternative interpretation, however, is that the groups that display the largest response are actually exposed to a more intense dosage of the treatment. All of our estimates so far have been of the effect of the intention to treat (ITT), which is attenuated toward zero when there is crossover and noncompliance. The groups that show the largest ITT effects may have received larger dosages of the treatment, in the form of particularly small classes or more years spent in a small class. Krueger and Whitmore (2002) show that disadvantaged students in the treatment group are not systematically assigned to the smallest of the small classes. Here we examine whether they are exposed to more years in a small class.

We generate subgroup estimates of the effect of assignment to a small class on years spent in a small class. To do so, we instrument for years actually spent in a small class with years potentially spent in a small class. Potential years in a small class is the product of assignment to a small class and the number of years the student could be enrolled in a small class, based on year of entry into STAR. For example, a student who entered STAR in kindergarten could spend as many as four years in a small class, while a child who entered in third grade could spend only one.[14]

We estimate the following equations:

$$YEARS_{is} = \delta_0 + \delta_1 Z_{is} + \delta_{sg} + \psi_{isg} \qquad (2)$$

[14] Abdulkadiroglu et al. (2011) and Hoxby and Murarka (2009) use a similar approach when they instrument for years spent in a charter school with potential years spent in a charter school, where potential years is a function of winning a charter lottery and the grade of application.

$$COLL_{isg} = \alpha_0 + \alpha_1 YEARS_{is} + \alpha_{sg} + \varepsilon_{isg}\,, \qquad (3)$$

where $COLL_{isg}$ is an indicator variable for whether student i, who entered the STAR experiment in school s and in grade g, ever enrolls in college. YEARS is the number of years the student spends in a small class. Z is the potential number of years a student could attend a small class multiplied by an indicator for whether the student was assigned to a small class. School-by-entry-grade fixed effects are included in each equation. We estimate these equations separately by subgroup.

Table 2.5 reports the estimates of the first stage equation, the reduced-form intention-to-treat model (ITT) and the two-stage least squares model (2SLS). The first-stage results in column (1) measures compliance, reporting the number of years actually spent in a small class for each year assigned to a small class. Overall, for each year of potential small-class attendance, students on average attend 0.64 years in a small class. The compliance rate is consistently smaller for the groups for whom we have estimated the largest effects of ITT. This is likely driven by higher mobility among black and poor students. The 2SLS estimates (Column 3) indicate that each year spent in a small class increases college attendance rates by one percentage point for the entire sample, but by 2.8 points for students attending the poorest schools, 2.4 points for black students, and 1.6 points for poor students. These results indicate that students who are black, poor, or attend high-poverty schools benefit more from a year spent in a small class than do their peers.

2.6.7. Do Short-Term Effects Predict Long-Term Effects?

We have shown that random assignment to small classes increases college entry and degree completion and shifts students toward high-paying majors. Could these effects have been predicted by the short-term effects of STAR on test scores? That is, are the effects measured at the time of the experiment predictive of the program's long-term effects?

A back-of-the-envelope prediction would combine the experiment's effect on scores with information from some other data source on the relationship between scores and postsecondary attainment. We now make such an informed guess about the long-term effects of STAR, then compare our guess with the paper's findings.

The guess requires information about the relationship between standardized

scores in childhood and adult educational attainment, ideally for a cohort born around the same time as the STAR subjects. The NLSY79 Mother-Child Supplement contains longitudinal data on the children of the women of the National Longitudinal Survey of Youth. These children were born at roughly the same time as the STAR cohort. The children of the NLSY (CNLSY) were tested every other year, including between the ages of six and nine (the ages of the STAR subjects while the experiment was underway). Postsecondary attainment is also recorded in CNLSY.

In CNLSY a standard deviation increase in childhood test scores is associated with a 16 percentage-point increase in the probability of attending college.[15] Assignment to a small class in STAR increases the average of K-3 scores by 0.17 standard deviations. Under the assumption that the relationship between scores and attainment is the same for the STAR and NLSY79 children, a reasonable prediction of the effect of STAR on the probability of college attendance is 2.72 percentage points (=0.17*16). This back-of-the-envelope calculation is nearly identical to the 2.7 point estimate we obtained in our regression analysis, indicating that the contemporaneous effect of STAR on scores is an excellent predictor of its effect on adult educational attainment.

Another way to approach this question is to examine whether the estimated effect of small classes on postsecondary attainment disappears when we control for K-3 test scores. This is an informal test of whether class size affects postsecondary attainment through any channel other than test scores. This sort of informal test is often used when checking whether an instrument (e.g., assigned class size) affects the outcome of interest (e.g., postsecondary attainment) through any channel other than the endogenous regressor (e.g., test scores). We first re-estimate Equation (1) and report the main result in column 1 of Table 2.6. We then add to this regression a student's test scores and the interaction of the test scores and assignment to a small class. The interaction allows the relationship between test scores and postsecondary attainment to differ between small and regular classes:

$$Coll_{isg} = \beta_0 + \beta_1 SMALL_{is} + \beta_2 TEST_{is} + \beta_3 SMALL_{is} * TEST_{is} + \beta_4 X_{is} + \beta_{sg} + \varepsilon_{isg} \quad (4)$$

[15] We regress an indicator for college attendance against the average scores in multiple standardized tests administered when the subjects were between ages six and nine. Scores are normalized (within age) to mean zero and standard deviation one. We measure college attendance by 2006, when the children were 25 to 29 years old.

Here, $Coll_{isg}$ is a dummy that equals one if student i who entered the STAR experiment in school s and grade g ever attended college. $TEST_{is}$ is the average of student i's non-missing kindergarten through third grade math and reading test scores, normalized to mean zero and standard deviation of one. Results are in Table 2.6 (Column 2).

First looking to the coefficient on test scores, in STAR a one-standard deviation increase in K-3 scores is associated with a 17 percentage-point increase in the probability of attending college.[16] This is very similar to the relationship estimated among the children of the NLSY. The estimated coefficient on the interaction term between small class assignment and average test score is zero, indicating that scores have no differential predictive power for postsecondary attendance across students in small and regular classes. Similarly, the estimated coefficient on the small class indicator variable is also zero, suggesting that there is no additional boost to the likelihood a student attends postsecondary school from small class assignment after accounting for contemporaneous test scores (which are boosted by smaller classes). The pattern is similar if we replace college attendance with degree receipt (Columns 3-4). These findings indicate that short-term gains in cognitive test scores are indeed predictive of long-term benefits.

By contrast, we find that scores from tests administered after children left STAR are not nearly so predictive of its long-term effect. We estimate the equation just described, replacing contemporaneous scores with those obtained from tests administered in grades six through eight, three to five years after the experiment had ended. Now we see that, even after controlling for test scores, small-class assignment raises the likelihood of attending college by a statistically significant 2 percentage points. Further, the negative coefficient on the interaction term indicates that these subsequent test scores have less predictive power in small than regular classes. We conclude that scores recorded several years after the experiment do a significantly poorer job than contemporaneous scores in predicting the effect of the experiment on adult outcomes. One caveat to this analysis is that there could be omitted variables that are correlated both with assignment to a small class, test scores, and college attendance. If this were the case, then it might not be the contemporaneous test scores that are mediating the effect of small class assignment, but rather the omitted variables.

[16] Results are unchanged if we exclude the school-by-wave fixed effects and demographics.

2.7. Conclusion

We estimate the effect of class size in early elementary school on postsecondary attainment. Assignment to a small class increases college attendance by 2.7 percentage points. Enrollment effects are largest among black students, students from low-income families, and high-poverty schools, indicating that class-size reductions during early childhood can help to close income and racial gaps in postsecondary attainment. Assignment to a small class also increases degree completion by 1.6 percentage points, with the effects concentrated in high-earning fields such as business, economics, and STEM.

As a final check on the sensitivity of our main result to possible sources of bias, we conduct two exercises. First, we examine the extent that students missing name and date of birth could influence the results, given that the NSC uses these identifiers to match students to college enrollment data. We assign all students with a missing name or date of birth first as having enrolled in college and then as having not enrolled in college regardless of their observed enrollment status. After each of these imputations we re-estimate Equation (1). Imputing students with missing identifiers as enrolled (not enrolled) yields a point estimate of 0.017 (0.025) and standard error of 0.009 (0.011). These coefficients are somewhat attenuated relative to our main result of 0.027 (SE=0.011). However, this check shows that even if we impute the most extreme cases of possible bias due to missing identifiers, our result remains positive, statistically significant, and similar in magnitude to our main result.

Our final check is a back-of-the-envelope exercise to bound the possible upward bias that could be due to small class assignment inducing students out of colleges not participating in the NSC (e.g., for-profits) and into colleges that do participate (e.g., community colleges). Using the NSC participant list and IPEDS enrollment data, we calculate that 8.7 percent of first-time enrollment in Tennessee during 1998 is in for-profit colleges. If small classes induce all of these students out of for-profit institutions and into colleges that we observed in the NSC (an extreme assumption), then our estimated effect on college enrollment would be biased upward by 3.7 percentage points.[17] This upper

[17] In other words, if we assume that none of the treatment group attends for-profit colleges but 8.7 percent of the control group does, the implied total college enrollment rate among the control group would be 0.422. This rate is 3.7 percentage points higher than the observed attendance rate among the control group (excluding for-profit colleges) of 0.385.

bound on the upwards bias is larger than our observed treatment effect. However, a somewhat more realistic estimate based on past studies of STAR would be to assume that the treatment induces 10 percent of students out of for-profit institutions and into colleges that we observe (Krueger and Whitmore, 2001). This would cause our estimates to be biased upwards by 0.4 percentage points. This excludes any possible attenuation bias due to classical measurement error in the unobserved nonprofit college attendance, and any possible downward bias due to small classes inducing non-college attenders into for-profit institutions. This is thus a source of potential upward bias that under a somewhat plausible worst case scenario would explain only a small fraction of our treatment effect.

Is the nearly three percentage-point increase due to reduced class size that we estimate a large effect? To put this effect in context, we compare the estimate to those of other interventions that boost postsecondary attainment. We focus on the results of randomized trials when possible, turning to plausibly-identified quasi-experiments where no controlled experiment has been conducted. Deming and Dynarski (2010) provide a review of this literature, from which much of this information is drawn. We focus on evaluations of discrete, replicable interventions. We deliberately ignore several excellent papers that demonstrate that schools or teachers "matter" for postsecondary attainment, since they do not identify the effect of a manipulable parameter of the education production function (e.g., Deming et al., 2011, Chetty et al., 2011).

Two small experiments have tested the effect of intensive preschool on long-term outcomes. Abecedarian produced a 22 percentage-point increase in the share of children who eventually attended college. The Perry Preschool Program had no statistically significant effect on postsecondary outcomes (Anderson, 2008). The subjects in these experiments were almost exclusively poor and black. Head Start, a less intensive preschool program, increases college attendance by 6 percentage points (Deming, 2009), with larger effects for blacks and females (14 and 9 percentage points, respectively). Upward Bound provides at-risk high-school students with increased instruction, tutoring and counseling. The program had no detectable effect on the full sample of treated students, but it did increase college attendance among students with low educational aspirations by 6 percentage points (Seftor, Mamun, & Schirm, 2009).

There are no experimental estimates of the effect of financial aid on college entry. However, there are several well identified quasi-experimental studies showing that

student aid can boost postsecondary enrollment by several percentage points depending on how much aid is provided (Deming & Dynarski, 2010). Another way of increasing college enrollment is by assisting students with the administrative requirements of enrolling in college. Bettinger et al. (2012) randomly assign families to a low-cost treatment that consists of helping them to complete the FAFSA, the lengthy and complicated form required to obtain financial aid for college. Their intervention increases enrollment by eight percentage points.

The costs of the above interventions vary dramatically. We create an index of cost effectiveness for increasing college enrollment by dividing each program's costs by the proportion of treated students it induces into college.[18] Head Start costs $8,000 per child. Given the 6 percentage-point effect noted above, the amount spent by Head Start to induce a single child into college is therefore $133,333 (=$8,000/0.06). For Abecedarian, the figure is $410,000 (=$90,000/0.22). The cost of reduced class size is $12,000 per student, larger than that of Head Start but considerably smaller than that of Abcedarian. The amount spent in STAR to induce a single child into college is $400,000 (=$12,000/0.03). If the program could be focused on students in the poorest third of schools (the subpopulation that most closely matches that of the preschool interventions) then the cost drops to $171,000 per student induced into college.

Upward Bound costs $5,620 per student. If the program could be targeted to students with low educational aspirations, the implied cost of inducing a single student into college is $93,667 (=$5,620/0.06). Dynarski (2003) examines the effect of the elimination of the Social Security Student Benefit Program, which paid college scholarships to the dependents of deceased, disabled and retired Social Security beneficiaries. Eligible students were disproportionately black and low-income. The estimates from that paper indicate that about two-thirds of the treated students who attended college were inframarginal, while the other third was induced into the college by the $7,000 scholarship. These estimates imply that three students are paid a scholarship in order to induce one into college. The cost per student induced into college is therefore $21,000. Finally, the cost per treated subject in the FAFSA experiment (Bettinger et al., 2012) was $88 for an implied cost per student induced into college of $1,100

[18] All costs in this section are in 2007 dollars and come from Deming and Dynarski (2010) unless otherwise indicated. The costs for the early childhood programs and STAR have been discounted back to age zero using a 3 percent discount rate. Costs of the high school and college interventions have not been discounted.

(=$88/0.08).

A fair conclusion from this analysis is that the effects we find in this paper of class size on college enrollment alone are not particularly large given the costs of the program. If focused on students in the poorest third of schools, then the cost-effectiveness of class size reduction is within the range of other interventions. There is no systematic evidence that early interventions pay off more than later ones when the outcome is limited to increased college attendance.

In addition to estimating the effects of reduced class size during childhood on educational attainment, the results in our paper shed light on the relationship between the short-and long-term effects of an educational intervention. We find that the short-term effect of small class assignment on test scores is an excellent predictor of its effect on adult educational attainment. In fact, under the assumption that there are no omitted variables correlated with small class assignment, test scores, and college enrollment, the effect of small classes on college attendance is completely "explained" by their positive effect on contemporaneous test scores. Further, the relationship between scores and postsecondary attainment is the same in small and regular classes; that is, the scores of children in the small classes are no less (or more) predictive of adult educational attainment than those of children in the regular classes. This is an important and policy-relevant finding, given the necessity to evaluate educational interventions based on contemporaneous outcomes.

A further contribution of this paper is to identify the effect of manipulating a single educational input on adult educational attainment. The early-childhood interventions for which researchers have identified lifetime effects (e.g., Head Start, Abecederian) are intensive and multi-pronged, including home visits, parental coaching and vaccinations. We cannot distinguish which dimensions of these treatments generate short-term effects on test scores, and whether they differ from the dimensions that generate long-term effects on adult wellbeing. By contrast, the effects we measure in this paper, both short-term and long-term, can be attributed to a well-defined and replicable intervention: reduced class size.

References

Abdulkadiroglu, A., Angrist, J.D., Dynarski, S.M., Kane, T.J., & Pathak, P.A. (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. Quarterly Journal of Economics, 126 (2), 699–748.

Achilles, C.M. (1999). Let's put kids first, finally: Getting class size right. Thousand Oaks, CA: Corwin Press.

Anderson, M.L. (2008). Multiple inference and gender differences in the effects of early Intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. Journal of the American Statistical Association, 103 (484), 1481–1495.

Angrist, J.D. & Krueger, A.B. (1991). Does compulsory school attendance affect schooling and earnings? Quarterly Journal of Economics, 106 (4), 979–1014.

Angrist, J.D., Dynarski, S.M., Kane, T.J., Pathak, P.A., & Walters, C.R. (2012) Who benefits from KIPP? Journal of Policy Analysis and Management, 31 (4), 837–860.

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

Bailey, M.J. & Dynarski, S.M. (2011). Gains and gaps: A historical perspective on inequality in college entry and completion. In G. Duncan & R. Murnane (Eds.), Wither opportunity: Rising inequality, schools, and children's life chances. New York: Russell Sage.

Barrow, L., Richburg-Hayes, L., Rouse, C.E., & Brock, T. (2009). Paying for performance: The education impacts of a community college scholarship program for low-income adults. Working Paper No. 2009-13. Chicago: Federal Reserve Bank of Chicago.

Belley, P. & Lochner, L. (2007). The changing role of family income and ability indetermining educational achievement. Journal of Human Capital, 1 (1), 37–89.

Bettinger, E.P., Long, B.T., Oreopoulos, P. & Sanbonmatsu, L. (2012) The role of application assistance and information in college decisions: Results from the H&R Block FAFSA experiment. Quarterly Journal of Economics, 127 (3), 1205–1242.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In E. Leamer & J.J. Heckman (Eds.) Handbook of Econometrics. XXcity: etc.

Bound, J., Lovenheim, M. & Turner, S.E. (2010). Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. American Economic Journal: Applied Economics, 2 (3), 1–31.

Bowen, W.G., Chingos, M.M., & McPherson, M.S. (2009). Crossing the finish line: Completing college at America's public universities. Princeton, N.J.: Princeton University Press.

Cascio, E.U. & Staiger, D. (2012). Knowledge, test, and fadeout in educational interventions. NBER Working Paper No 18038. Cambridge, MA: National Bureau of Economic Research.

Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D.W. & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence From Project Star. Quarterly Journal of Economics, 126 (4), 1593–1660.

Dee, T.S. (2004). Are there civic returns to education? Journal of Public Economics, 88, 1697–1720.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. American Economic Journal: Applied Economics, 1 (3), 111–134.

Deming, D. & Dynarski, S.M. (2010). Into college, out of poverty? Policies to increase the postsecondary attainment of the poor. In P. Levine & D. Zimmerman (Eds.), Targeting investments in children: Fighting poverty when resources are limited. Chicago: University of Chicago Press.

Deming, D., Hastings, J., Kane, T., & Staiger, D. (2011). School choice, school quality and postsecondary attainment. NBER Working Paper No 17438. Cambridge, MA: National Bureau of Economic Research.

Dobbie, W. & Fryer, R.G. (2011). Are high quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. American Economic Journal: Applied Economics, 3 (3), 158–187.

Dynarski, S.M. (2003). Does aid matter? Measuring the effect of student aid on college attendance and completion. American Economic Review, 93 (1), 279–288.

Dynarski, S.M., Hemelt, S.W. & Hyman, J.M. (2012). Data watch: Using National Student Clearinghouse data to track postsecondary outcomes. Working Paper, University of Michigan.

Finn, J.D. & Achilles, C.M. (1990). Answers and questions about class size: A statewide experiment. American Educational Research Journal, 27, 557–577.

Folger, J. & Breda, C. (1989). Evidence from Project STAR about class size and student achievement. Peabody Journal of Education, 67, 17–33.

Fredriksson, P., Ockert, B., & Oosterbeek, H. (2013). Long-term effects of class size. Quarterly Journal of Economics, 128 (1), 249–285.

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. American Economic Review, 92 (4), 999–1012.

Hamermesh, D.S. & Donald, S.G. (2008). The effect of college curriculum on earnings: An affinity identifier for non-ignorable non-response bias. Journal of Econometrics, 144, 479–491.

Hoxby, C.M. & Murarka, S. (2009). Charter schools in New York City: Who enrolls and how they affect student achievement. NBER Working Paper No 14852. Cambridge, MA: National Bureau of Economic Research.

Krueger, A.B. (1999). Experimental estimates of education production functions. Quarterly Journal of Economics, 114, 497–532.

Krueger, A.B. & Whitmore, D.M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. Economic Journal, 111, 1–28.

Krueger, A.B. & Whitmore, D.M. (2002). Would smaller classes help close the black-white achievement gap? In J.E. Chubb & T. Loveless (Eds.) Bridging the Achievement Gap. Washington: Brookings Institution Press.

Lleras-Muney, A. (2005). The relationship between education and adult mortality in the United States. Review of Economic Studies, 72, 189–221.

Ludwig, J. & Miller, D.L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. Quarterly Journal of Economics, 122 (1), 159–208.

National Center For Education Statistics (2010). Integrated Postsecondary Education Data System (IPEDS). Washington, D.C.: U.S. Department of Education. National Science Foundation (2011). Science and Engineering Degrees: 1966-2008. Detailed Statistical Tables NSF 11-316. Arlington, VA: National Center for Science and Engineering Statistic.

Roderick, M., Nagaoka, J. & Allensworth, E. (2006). From high school to the future: A first look at Chicago Public School graduates' college enrollment, college preparation, and graduation from 4-year colleges. Chicago, IL: Consortium on Chicago School Research at the University of Chicago.

Ruggles, S., Alexander, J.T., Genadek, K., Goeken, R., Schroeder, MB., & Sobek, M. (2010). Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database], Minneapolis: University of Minnesota.

Schanzenbach, D.W. (2006). What have researchers learned from Project STAR? Brookings Papers on Education Policy, 2006(1), 205-228.

Schweinhart, L.J., Montie, J., Xiang, Z., Barnett, W.S., Belfield, C.R., & Nores, M. (2005). Lifetime effects: The High/Scope Perry Preschool study through age 40. Ypsilanti, MI: High/Scope Press.

Seftor, N.S., Mamun, A. & Schirm, A. (2009). The impacts of regular upward bound on postsecondary outcomes 7-9 years after scheduled high school graduation: Final report. Princeton, NJ: Mathematica Policy Research.

Word, E., Johnston, J., Bain, H.P., Fulton, B.D., Zaharias, J.B., Achilles, C.M., Lintz, M.N., Folger, J. & Breda, C. (1990). The state of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985-1990. Nashville: Tennessee State Department of Education.

Table 2.1. Means of Demographics and Outcome Variables by Class Size

	Regular Class	Small Class	Regression Adjusted Difference	
	(1)	(2)	(3)	
Demographics				
White	0.620	0.660	-0.003	(0.005)
Female	0.471	0.473	-0.000	(0.011)
Free Lunch	0.557	0.521	-0.015	(0.011)
College attendance				
Ever attend	0.385	0.420	0.027	(0.011)
Ever attend full-time	0.278	0.300	0.013	(0.011)
Enrolled On-Time	0.274	0.308	0.024	(0.011)
Number of Semesters				
Attempted	3.07	3.39	0.219	(0.133)
Attempted, conditional on attending	7.98	8.08	0.132	(0.209)
Degree Receipt				
Any degree	0.151	0.174	0.016	(0.009)
Associates	0.027	0.034	0.007	(0.004)
Bachelors or higher	0.124	0.141	0.009	(0.008)
Degree Type				
STEM, business or economics field	0.044	0.060	0.013	(0.006)
All other fields	0.085	0.094	0.003	(0.006)
First Attended				
2-year	0.215	0.245	0.025	(0.009)
Public 4-year	0.127	0.132	0.005	(0.007)
Private 4-year	0.042	0.043	-0.003	(0.004)
Number of Schools	79			
Number of Students	8,316	2,953		

Notes: Column (3) controls for school-by-wave fixed effects and demographics. Standard errors, in parentheses, are clustered by school.

Table 2.2. The Effect of Class Size on College Attendance - Linear Probability Models

	Total		White	Black	No Free Lunch	Free Lunch	Tercile of Poverty Share				P-value: High vs. Middle/Low
							High	Middle	Low	Middle & Low	
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
College Attendance											
Ever attend	0.028	0.027	0.011	0.058	0.010	0.044	0.073	-0.010	0.022	0.006	0.008
	(0.012)	(0.011)	(0.013)	(0.022)	(0.017)	(0.015)	(0.021)	(0.017)	(0.018)	(0.012)	
	0.385		*0.432*	*0.308*	*0.563*	*0.272*	*0.262*	*0.417*	*0.476*	*0.446*	
Ever attend full-time	0.014	0.013	-0.000	0.037	0.000	0.025	0.048	-0.012	0.008	-0.003	0.048
	(0.011)	(0.011)	(0.013)	(0.021)	(0.016)	(0.014)	(0.022)	(0.015)	(0.018)	(0.012)	
	0.278		*0.317*	*0.212*	*0.440*	*0.175*	*0.173*	*0.297*	*0.363*	*0.330*	
Enrolled On-Time	0.025	0.024	0.018	0.036	0.025	0.024	0.047	0.007	0.023	0.015	0.228
	(0.012)	(0.011)	(0.013)	(0.021)	(0.017)	(0.014)	(0.023)	(0.017)	(0.018)	(0.013)	
	0.274		*0.321*	*0.197*	*0.449*	*0.163*	*0.163*	*0.296*	*0.363*	*0.329*	
Demographics	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Number of Schools	79		79		79		24	29	26	55	
Number of Students	11,269	11,269	7,160	4,109	4,454	6,815	3,681	3,784	3,804	7,588	

Notes: All regressions control for school-by-entry-wave fixed effects. Demographics include race, sex and free lunch status. Standard errors, in parentheses, are clustered by school. Control means are in italics below standard errors.

Table 2.3. The Effect of Class Size on College Choice - Linear Probability Models

		Tercile of Poverty Share		P-value: High vs. Middle/Low
	Total	High	Middle & Low	
Dependent variable	(1)	(2)	(3)	(4)
College attendance	0.027	0.073	0.006	0.008
	(0.011)	(0.021)	(0.012)	
	0.385	*0.262*	*0.446*	
First Attended:				
2-year	0.025	0.063	0.007	0.009
	(0.009)	(0.019)	(0.010)	
	0.215	*0.162*	*0.242*	
Public 4-year	0.005	0.009	0.003	0.690
	(0.007)	(0.011)	(0.010)	
	0.127	*0.070*	*0.156*	
Private 4-year	-0.003	0.001	-0.004	0.491
	(0.004)	(0.004)	(0.005)	
	0.042	*0.030*	*0.049*	
Ever Attended:				
Out of state	0.013	0.029	0.006	0.197
	(0.009)	(0.013)	(0.012)	
	0.138	*0.100*	*0.157*	
Selective	0.009	0.007	0.011	0.839
	(0.009)	(0.016)	(0.011)	
	0.184	*0.090*	*0.231*	
Number of Schools	79	24	55	
Number of Students	11,269	3,681	7,588	

Notes: All regressions control for school-by-entry-wave fixed effects and demographics including race, sex, and free lunch status. Standard errors, in parentheses, are clustered by school. Control means are in italics below standard errors.

Table 2.4. The Effect of Class Size on Persistence and Degree Receipt - Linear Probability Models

		Tercile of Poverty Share		P-value: High vs. Middle/Low
	Total	High	Middle & Low	
Dependent variable	(1)	(2)	(3)	(4)
Number of Semesters	0.22	0.32	0.19	0.651
Attempted	(0.13)	(0.26)	(0.15)	
	3.07	*1.91*	*3.65*	
Receive Any Degree	0.016	0.011	0.019	0.624
	(0.009)	(0.012)	(0.012)	
	0.151	*0.071*	*0.191*	
Highest Degree				
Associates	0.007	0.007	0.007	0.918
	(0.004)	(0.006)	(0.006)	
	0.027	*0.013*	*0.034*	
Bachelors or higher	0.009	0.003	0.012	0.532
	(0.008)	(0.011)	(0.010)	
	0.124	*0.058*	*0.157*	
Degree Type				
STEM field	0.005	0.000	0.008	0.194
	(0.003)	(0.004)	(0.004)	
	0.019	*0.008*	*0.024*	
Business or economics field	0.007	0.001	0.011	0.189
	(0.005)	(0.004)	(0.006)	
	0.026	*0.012*	*0.033*	
All other fields	0.003	0.013	-0.000	0.279
	(0.006)	(0.008)	(0.008)	
	0.085	*0.039*	*0.108*	
STEM, business or economics field	0.013	0.001	0.019	0.092
	(0.006)	(0.006)	(0.008)	
	0.044	*0.020*	*0.057*	
Number of Schools	79	24	55	
Number of Students	11,269	3,681	7,588	

Notes: All regressions control for school-by-entry-wave fixed effects and demographics including race, sex, and free lunch status. Standard errors, in parentheses, are clustered by school. Control means are in italics below standard errors.

Table 2.5. Examining Whether Heterogeneity is in Treatment Effects or Dosage

	First Stage	Reduced Form	Two-Stage-Least-Squares	Control Mean
	(1)	(2)	(3)	(4)
Everyone	0.643	0.006	0.009	0.385
(n=11,269)	(0.016)	(0.003)	(0.005)	
High Poverty Share	0.602	0.017	0.028	0.262
(n=3,681)	(0.025)	(0.006)	(0.010)	
Middle/Low Poverty Share	0.662	0.001	0.002	0.446
(n=7,588)	(0.019)	(0.004)	(0.005)	
Black	0.589	0.014	0.024	0.308
(n=4,109)	(0.019)	(0.006)	(0.010)	
White	0.669	0.003	0.004	0.432
(n=7,160)	(0.019)	(0.004)	(0.006)	
Free Lunch	0.628	0.010	0.016	0.272
(n=6,815)	(0.015)	(0.004)	(0.007)	
Non-Free Lunch	0.665	0.002	0.003	0.563
(n=4,454)	(0.024)	(0.005)	(0.008)	

Notes: All regressions control for school-by-entry-wave fixed effects and demographics including race, sex, and free lunch status. Standard errors, in parentheses, are clustered by school.

Table 2.6. Examining Whether Short-Term Gains Predict Long-Term Gains - Linear Probability Models

	College Enrollment		Degree Receipt	
	(1)	(2)	(3)	(4)
Mean K-3 Test Score				
Small class	0.027	0.002	0.016	0.001
	(0.011)	(0.009)	(0.009)	(0.009)
Test score		0.169		0.096
		(0.006)		(0.007)
Small class * test score		-0.008		0.000
		(0.010)		(0.008)
Mean 6-8 Test Score				
Small class	0.027	0.020	0.016	0.010
	(0.011)	(0.010)	(0.009)	(0.008)
Test score		0.230		0.141
		(0.005)		(0.006)
Small class * test score		-0.014		0.009
		(0.008)		(0.008)
Control Mean	0.385	0.385	0.151	0.151
Number of Students	11,269	11,269	11,269	11,269

Notes: All regressions control for school-by-entry-wave fixed effects and demographics including race, sex, and free lunch status. Missing test-score indicators included for students with no test scores in grade range. Standard errors, in parentheses, are clustered by school.

Table 2.A.1. Student Demographics by School Poverty Share

	High Poverty	Middle Poverty	Low Poverty	Middle/Low Poverty
	(1)	(2)	(3)	(4)
White	0.253	0.746	0.881	0.814
Female	0.471	0.475	0.469	0.472
Free Lunch	0.855	0.504	0.292	0.398
Number of Schools	24	29	26	55
Number of Students	3,681	3,784	3,804	7,588

Notes: School poverty share is measured as the fraction of the school that is eligible for a subsidized lunch.

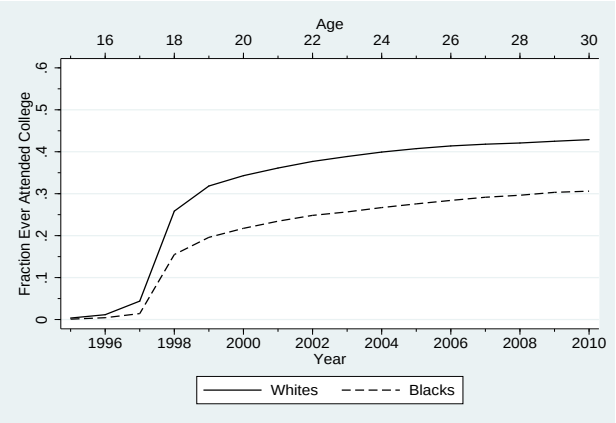
Table 2.A.2. The Effect of Class Size Censoring to Match IRS Data Span - Linear Probability Models

	Baseline - All Years of Enrollment	Exclude Pre-1999 Enrollment	Exclude Post-2007 Enrollment	Include 1999-2007 Enrollment Only
Dependent variable	(1)	(2)	(3)	(4)
Ever attend	0.027	0.018	0.023	0.015
	(0.011)	(0.011)	(0.011)	(0.011)
	0.385	*0.369*	*0.372*	*0.357*
Number of Students	11,269	11,269	11,269	11,269

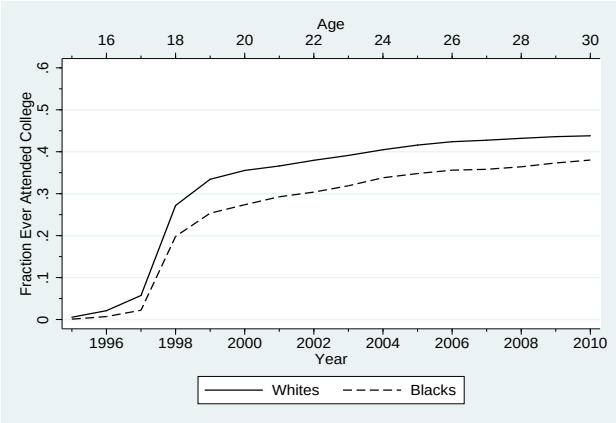
Notes: All regressions control for school-by-entry-wave fixed effects and demographics including race, sex, and free lunch status. Standard errors, in parentheses, are clustered by school. Control means are in italics below standard errors.

Figure 2. 1: The Effect of Class Size on Racial and Income Gaps in Postsecondary Attainment

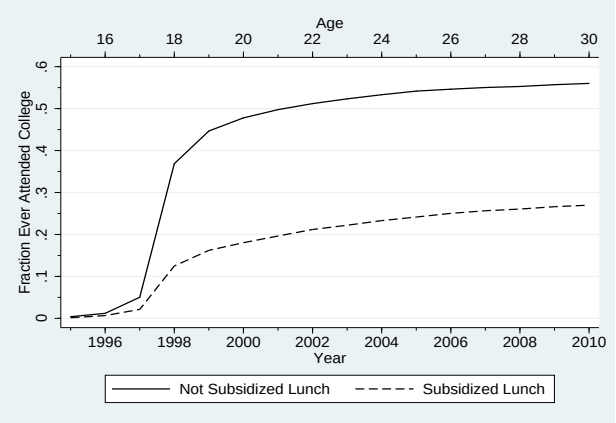
(a) Regular Class



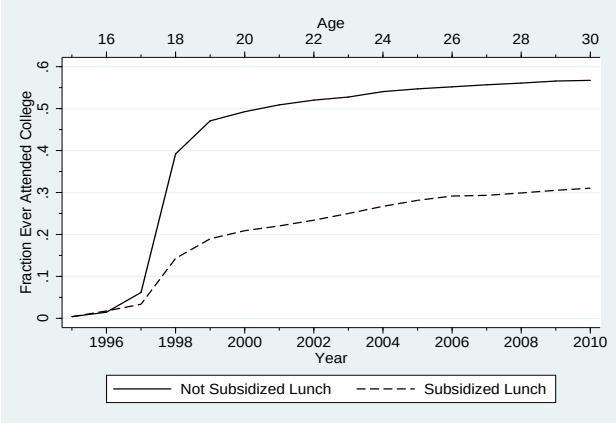
(b) Small Class



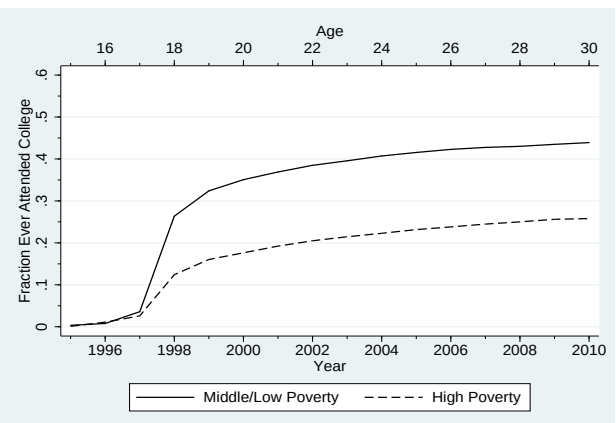
(c) Regular Class



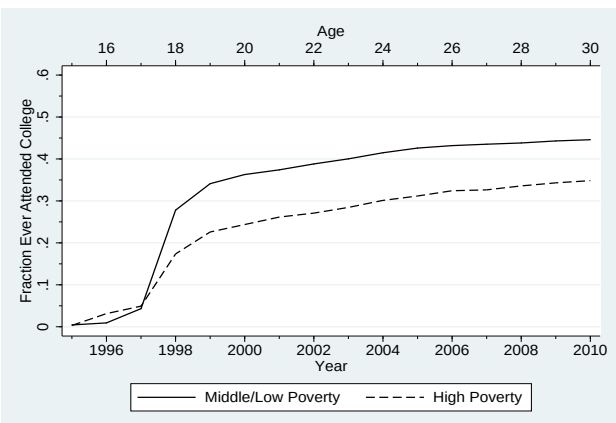
(d) Small Class



(e) Regular Class



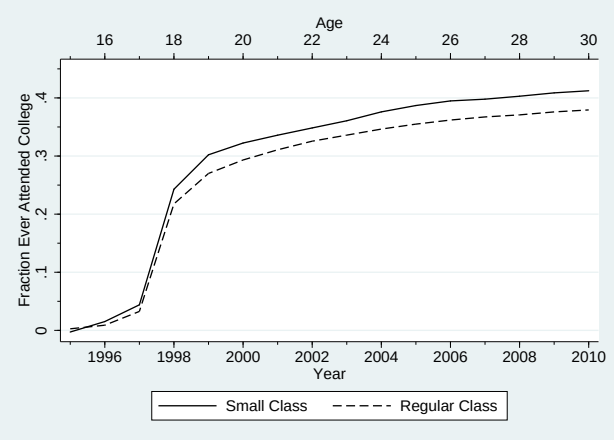
(f) Small Class



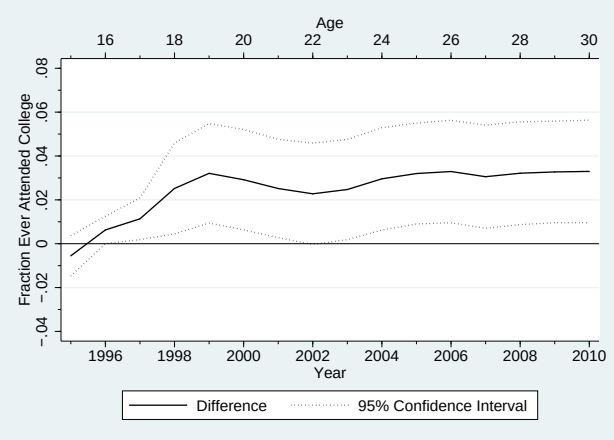
Notes: Figures (a), (c), and (e) plot the fraction ever attended college by year for STAR students assigned to regular size classes, and figures (b), (d), and (f) for STAR students assigned to small classes. Figures (a) and (b) compare college attendance by race, figures (c) and (d) by subsidized lunch status, and figures (e) and (f) by school poverty share.

Figure 2. 2: College Attendance Over Time, by Class Size

(a) Fraction Ever Attended College



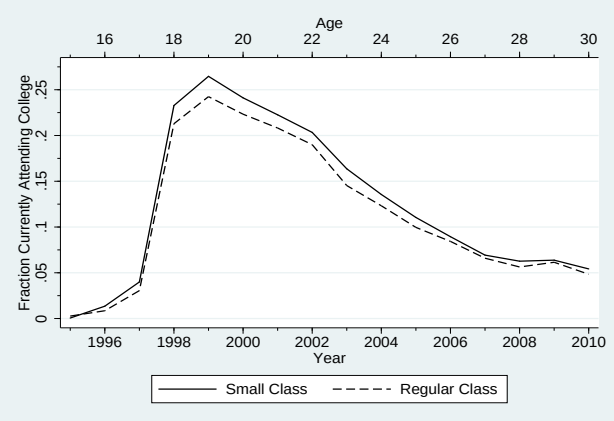
(b) Difference Between Small and Regular



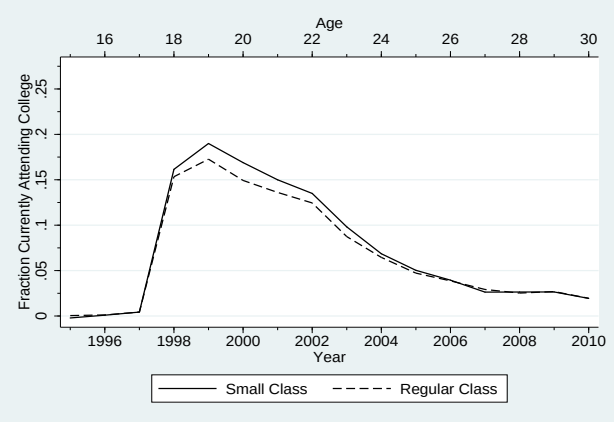
Notes: Figure (a) plots the mean fraction ever attended college by year for students assigned to small vs. regular size classes. It controls for both school-by-wave fixed effects and demographics, including race, sex and subsidized lunch status. Figure (b) plots the difference and its 95% confidence interval by year. Standard errors are clustered by school.

Figure 2. 3: Fraction Currently Enrolled in College Over Time, by Class Size and Enrollment Status

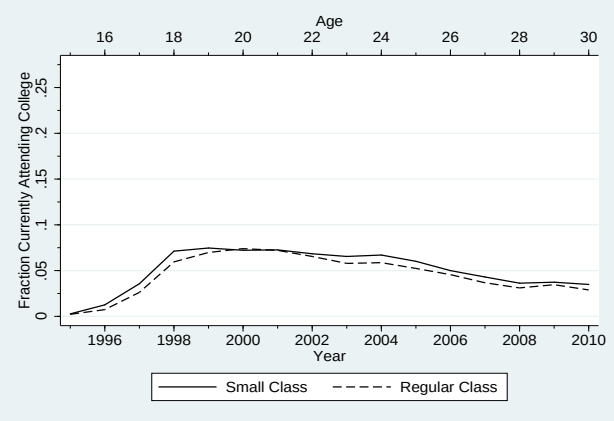
(a) Any Enrollment Status



(b) Full-Time Status



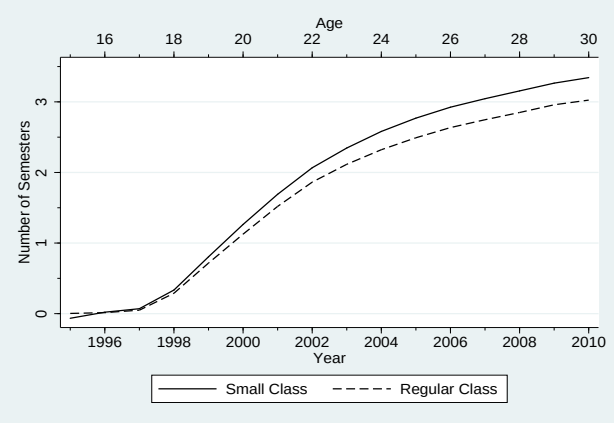
(c) Part-Time Status



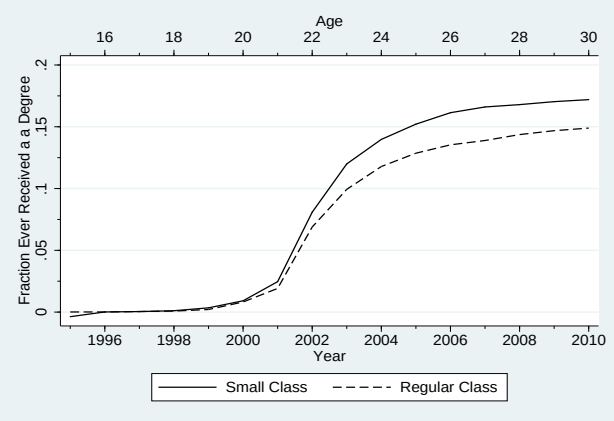
Notes: Figures plot the fraction currently attending college by year for STAR students assigned to small vs. regular size classes. All figures control for both school-by-wave fixed effects and demographics, including race, sex and subsidized lunch status.

Figure 2. 4: Postsecondary Persistence and Degree Receipt Over Time, by Class Size

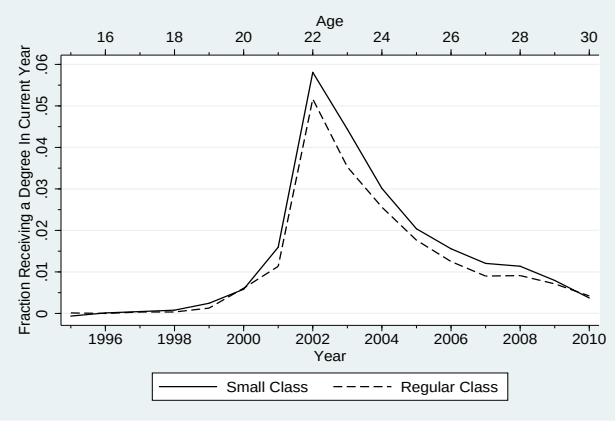
(a) Cumulative Number of Semesters Attended



(b) Fraction Ever Received a Degree



(c) Fraction Receiving Degree in Current Year



Notes: Figure (a) plots the mean cumulative number of semesters attended by year for STAR students assigned to small vs. regular size classes. Figure (b) plots the mean fraction ever receiving any postsecondary degree (associate's or higher). Figure (c) plots the mean fraction receiving any postsecondary degree in the current year. All figures control for both school-by-wave fixed effects and demographics, including race, sex and subsidized lunch status.

CHAPTER III

College Entrance Exams, Sample Selection, and the Distribution of Student Achievement

Abstract

Student scores on college entrance exams such as the ACT and SAT can provide policy-makers, administrators, and researchers with a powerful measure of college-readiness. However, fewer than half of public school students take one of these exams, raising concerns about the representativeness of observed scores. In order to examine the degree of selection into college entrance exam-taking, I exploit a policy in Michigan requiring all students to take the ACT. Using pre-policy data, I compare the performance of several parametric and semiparametric sample selection correction methods at approximating the true distribution of post-policy scores. I use the driving distance from a student's home to the nearest ACT test center as an exclusion restriction in the selection corrections. I find that when using basic student demographics or school-level information, all of the corrections I employ perform similarly poorly. However, when adding students' eighth and eleventh grade test scores to the prediction, simple OLS does a good job at predicting the true distribution of latent scores of all students. This suggests that the level of detail of data is much more important than the choice of correction technique in correcting for selection bias in college entrance exam scores.

3.1. Introduction

In the post-No Child Left Behind educational landscape of assessment-based accountability, measures of state, district, or school performance are increasingly important and have high stakes. An appealing measure used to evaluate the college-readiness of these groups is their average ACT or SAT score. However, given that less than half of public high school students in the U.S. take a college entrance exam, it is unclear to what extent this measure accurately reflects the true college-readiness level of the overall student population. Similarly, because test-taking rates vary across key subgroups of students, comparisons by administrators or policy-makers of the college-readiness of these subgroups may not be representative of the true differences that would be observed if scores were available for all students. Understanding the degree of selection into college entrance exam-taking, and the extent to which existing

econometric techniques can correct for this selection, is important for estimating the true levels and gaps in college-readiness.

Understanding the extent of selection into college entrance exam taking would also make an important methodological contribution to the line of literature that examines the impact of an education policy or intervention on student achievement, using student scores on college entrance exams as a proxy for achievement (Krueger and Whitmore, 2001; Card and Payne, 2002; Angrist, Bettinger and Kremer, 2006; Rothstein, 2006). These studies use parametric correction methods to correct for the selection bias in college entrance exam scores that arises from only a select group of students taking the exams. These techniques rely on assumptions whose validity is difficult to verify without understanding the full distribution of latent test scores, but if incorrect, can result in misleading findings. Unfortunately, gaining insight into the true distribution of latent test scores by using naturally occurring geographic or temporal variation in test scores and participation rates is unreliable, due to the correlation between test-taking rates and unobserved determinants of achievement.

In this paper, I uncover and describe the full distribution of typically latent college entrance exam scores by exploiting the recent statewide adoption of mandatory college entrance exams in Michigan. Nearly a dozen states in recent years have incorporated a college entrance exam into their eleventh grade statewide test that all public school students must take to fulfill the requirements of No Child Left Behind (NCLB). The result is that almost all public school students in these states now take a college entrance exam, while far fewer students took one before the implementation of the policy. My paper diagnoses whether states that are considering the implementation of these reforms can predict their new average scores using current econometric methods.[1]

I use student-level enrollment, demographics, ACT and SAT scores, and state assessment data for two recent cohorts of eleventh graders in Michigan that straddle the implementation of the reform. Using the complete post-policy distribution of scores as the full distribution pre-policy, I compute the latent scores of non-test-takers pre-policy, and compare this distribution to

[1] Anecdotally and in the media, state and district administrators who have recently adopted mandatory test-taking policies (or are considering doing so) have voiced concerns that falling average scores have or will cause public relations problems. Some have expressed frustration at their inability to predict by how much test scores will fall when they implement the reform. See articles: "Testing way for more to take ACT" (Dejka, 2012); "ACT scores have fallen, Are area students ready for next step?" (Dunlap Tribune, 2011); "Low test scores worry districts" (Hetzner, 2010); "Everyone's into ACT, and it shows" (Banchero, 2002).

the observed scores of test-takers pre-policy. I show that while non-takers have lower scores on average, a substantial fraction score above the pre-policy mean, calling into question some of the assumptions used in previous studies.

I then consider the performance of methods previously used to correct for selection bias in college entrance exam scores, by examining how closely the pre-policy corrections can predict the post-policy distribution of scores. I additionally employ a more flexible parametric correction method, as well as a semiparametric correction method that makes minimal distributional assumptions. I show that using basic student demographics and school- and district-level characteristics, none of the methods that I employ with the pre-policy data come near to replicating the non-selected distribution of post-policy ACT scores. However, when including measures of students' past and contemporaneous achievement (eighth and eleventh grade scores), both simple OLS and all of the correction methods accurately predict the correct mean ACT score.

The rest of the paper is outlined as follows: In Section 3.2.1, I describe the classic sample selection model and correction procedure introduced by Gronau (1974) and Heckman (1974, 1976, and 1979). Section 3.2.2 discusses studies that have used corrections for sample selection in college entrance exam scores. I also briefly discuss one recent study that, like my paper, explicitly examines the degree of selection in ACT- and SAT-taking (Clark, Rothstein, and Schanzenbach, 2009). I discuss my data in Section 3.3. Section 3.4.1 compares the latent score distribution of ACT-takers and non-takers pre-policy. Section 3.4.2 and 3.4.3 consider the performance of several student-level and group-level correction methods, respectively. Section 3.5 concludes.

3.2. Sample Selection Bias

Before describing a formal model, I briefly discuss the intuition behind the issue of sample selection bias. While this issue first arose in studying the determinants of wages among females (Gronau, 1974; Heckman, 1974), I frame the model in the context of estimating the determinants of student achievement, where I proxy for achievement using student scores on a college entrance exam such as the ACT. Consider a population of students where only a subset of students takes the ACT. Suppose I am interested in estimating the determinants of achievement among the entire population of students, but I only have ACT scores for the subset

of students who take the test. If it is random whether a student takes the ACT, then selection bias will not be an issue, because (in expectation) the population of test-takers will be identical to the population of non-takers across observable and unobservable characteristics. The determinants of achievement that we identify among test-takers would be the same as for non-takers.

The problem of sample selection bias arises when the decision to take the ACT is non-random. In this situation, the characteristics of test-takers and non-takers will likely be different. Sample selection occurs when some of the determinants of the decision to take the ACT also influence the ACT score itself. If these determinants are fully observed, then we can control for them, and selection bias will again not be an issue. However, if there are unobserved characteristics affecting test-taking that are correlated with observed characteristics affecting achievement, failure to control for these unobserved characteristics will lead to selection bias in the estimates of the effect of the observed characteristics on achievement.

3.2.1. The Model

Consider a model of ACT-taking and ACT scores of the form:

$$ACT_i^* = \beta X_i + \varepsilon_i \quad (1)$$

$$TAKE_i^* = \gamma Z_i + u_i \quad (2)$$

$$TAKE_i = 1 \text{ if } TAKE_i^* > 0; \quad TAKE_i = 0 \text{ otherwise} \quad (3)$$

$$ACT_i = ACT_i^* * TAKE_i, \quad (4)$$

where ACT_i^* is the latent ACT score of student i, with observed score ACT_i. $TAKE_i^*$ is a latent variable with associated indicator $TAKE_i$, reflecting whether a student takes the ACT. Equations (3) and (4) show the relationships between latent and observed ACT participation and scores. Equation (1) is the equation of primary interest, and Equation (2) models the sample selection. X_i and Z_i are vectors of observed, exogenous variables, β and γ are vectors of unknown parameters, and ε_i and u_i are mean zero error terms. X_i is contained in Z_i, and later I discuss the implications of whether there are additional variables in Z_i that are not in X_i.

Selection bias arises when the conditional expectation of the error term is not equal to zero: $E[ACT_i|X_i] = \beta X_i + E[\varepsilon_i|u_i > -\gamma Z_i]$. The error term, ε_i, is conditional on $TAKE_i^* > 0$, or on $u_i > -\gamma Z_i$. Thus, it is likely to not be mean zero and to be correlated with X_i, yielding an inconsistent estimate of β. A key insight in Heckman (1976) is that this is an omitted variables problem. If we can control for the conditional expectation of ε_i, $E[\varepsilon_i|u_i > -\gamma Z_i]$, then β can be consistently estimated. A large literature beginning with Gronau and Heckman's work in the

mid-1970's, and extending for over two decades, develops various parametric and semiparametric corrections for selection bias in the context of the above sample selection model.

Gronau (1974), Heckman (1974), and later Heckman (1976, 1979) proposed the first methods to correct for this sort of sample selection bias. Their model requires some additional assumptions beyond Equations (1) – (4). Assume that the error terms in Equations (1) and (2) are distributed jointly normal, or $(\varepsilon, u) \sim N(0,0,\sigma_\varepsilon^2, \sigma_u^2, \rho_{\varepsilon u})$. Furthermore, assume that ε and u are independent of *X* and *Z*. Also, σ_u^2 is typically normalized to equal 1, though this is innocuous. The properties of the joint normal distribution, also known as the bivariate normal distribution, allow for consistent estimation of β, either by limited information maximum likelihood (LIML), or by a two-step method in which $TAKE_i = \gamma Z_i + u_i$ (the selection equation) is estimated using probit, and then $ACT_i = \beta X_i + \lambda(\hat{\gamma} Z_i) + \varepsilon_i$ (the outcome equation) is estimated using OLS. λ, the inverse Mills ratio (IMR), is defined as $\frac{\varphi(.)}{\phi(.)}$, where $\varphi(.)$ and $\phi(.)$ represent the probability density and cumulative distribution functions of the standard normal distribution.

There are several important weaknesses to the above two-step correction procedure. First, it is sensitive to the bivariate normality assumption of the error terms. If this assumption is incorrect, then the estimates of β will be inconsistent. Also, when *X*=*Z*, the model is identified only through the non-linearity of the IMR. Given that the IMR is quasi-linear for most of the range of its argument, substantial collinearity between the estimated IMR and *X* in the outcome equation tends to decrease precision and bias results when *X*=*Z* (Puhani, 2002). This is particularly true when the values of $\widehat{TAKE}$ tend not to be in the tails, i.e., near zero or one. In Monte Carlo simulations, this method performs poorly without an exclusion restriction, that is, without a variable in *Z* that does not belong in *X*. Performance also suffers the greater the degree of censoring in the sample.

The maximum likelihood version of the bivariate normal selection model correction method is fully efficient under the assumptions of the model, but it is less robust to departures from the bivariate normal assumption than the two-step estimator. This is because it relies more heavily on the functional form assumption than the two-part estimator, which relies also on the non-linearity of the IMR for identification. Soon after Gronau and Heckman developed their corrections, other authors developed similar parametric correction methods that somewhat loosen the bivariate normality assumption (Olsen, 1980; Lee, 1982 and 1983). Other authors developed semiparametric correction methods that more fully relax the distributional assumptions about the

error terms, such that the resulting corrections have limited reliance on parametric assumptions (Heckman and Robb, 1985; Powell, 1987; Newey, 1988; Robinson, 1988; Ahn and Powell, 1993; and Choi, 1993). In Section 3.4 I describe these methods in further detail.

One important restriction to note is that the above parametric and semiparametric extensions require the use of an exclusion restriction, because there is no non-linear inverse Mills ratio to help eliminate the collinearity between X and Z. I use student-level distance to the nearest ACT test center as an exclusion restriction in my analysis. I discuss the details of this instrument and test its validity in Section 3.4.

3.2.2. Selection Bias in College Entrance Exam Scores

Several studies estimate the effect of an educational intervention on the college entrance exam scores of students, but must account for the increased probability of test-taking among the treated group. These studies use parametric methods to correct for selection bias at the student-level. Krueger and Whitmore (2001) use the two-step Heckman correction without an exclusion restriction. Angrist, Bettinger, and Kremer (2006) artificially censor the score distribution at various quantiles and estimate Tobit regressions, relying on different (and arguably stronger) assumptions than the Heckman correction.[2] Both papers also adjust for selection by trimming the treated students with the lowest scores in the spirit of Lee (2009), creating a nonparametric upper bound of the treatment effect. In this context, the lower bound is simply the comparison of mean scores across groups, without correcting for selection. In both Krueger and Whitmore (2001) and Angrist, Bettinger, and Kremer (2006), the strong assumptions of the parametric methods, and the wide range of nonparametric bounds leave room for question as to the true treatment effects.

A second set of papers controls for selection bias in group-level estimates where the groups (for example, states or schools) have different test-taking rates (Card and Payne, 2002; Rothstein, 2006). These studies use ACT/SAT micro-data, but do not have student-level information on non-takers, so they cannot model the selection process at the individual level as in the Heckman correction. They control for selection bias using a group-level control function approach that follows Gronau's (1974) original group-level correction procedure. Assuming bivariate normality of the error terms, the control function term is the IMR evaluated at the

[2] Artificially censoring the sample of test-takers assumes that all non-takers would score below the censoring point. Furthermore, Tobit assumes that errors in the selection and outcome equations are identical, rather than the weaker assumption of being jointly normally distributed.

group-level test participation rate. This group-level method assumes that the only variables in *Z* are the variables by which the data are grouped. For example, if the analysis is at the school level, the assumption is that the only factor affecting ACT-taking is a student's school, and that within a school, variation in test-taking is due only to the disturbance term.

In both Card and Payne (2002) and Rothstein (2006), the authors test the sensitivity of their estimates to using other functions besides the IMR as the control function in their analysis. Their results do not change substantially, which reassures the authors that they have adequately corrected for selection bias. However, without a measure of "truth" with which to compare their estimates, it is possible that all of the corrections perform similarly poorly.

Most similar to my paper is a recent study by Clark, Rothstein, and Schanzenbach (CRS, 2009) that attempts to explicitly examine the extent of selection bias in college-entrance scores.[3] They model selection following the group-level bivariate selection model (Gronau, 1974). The bulk of the paper consists of regressing observed mean state- or school-level ACT and SAT scores on the IMR evaluated at the state or school's test participation rate. The coefficient on the IMR term demonstrates the degree of selection in the sample. However, without an exogenous shock to the participation rate, the coefficient on this term will be biased, if there is correlation between test-taking rates and unobserved factors affecting achievement. The main contribution of the paper is that the authors use ACT data from Illinois, and exploit the change to a mandatory ACT-taking policy, in order to generate a plausibly exogenous change in the IMR term. The authors estimate the parameters of the bivariate normal selection model, and conclude that the correlation between latent and observed school-level mean scores is so high that using observed mean scores as a proxy for latent mean scores is unlikely to substantially alter the main results of research, not correcting for selection bias.

My paper improves on Clark, Rothstein, and Schanzenbach (2009) in two key ways. First, their paper is estimated entirely within a framework of the bivariate normal selection model, requiring strong functional form assumptions. I test the performance of different parametric correction methods that make different functional form assumptions, as well as a

[3] A number of older papers examine whether state-level mean SAT scores suffer from selection bias by controlling for state and year fixed effects, state-year level characteristics and polynomials of the state-year level test-participation rate (Dynarski 1987; Dynarski and Gleason 1993). These studies find substantial selection bias in state-level averages and suggest correction methods that are heavily dependent on functional form assumptions. None of these studies exploit exogenous changes in test participation rates and scores.

semiparametric correction that largely relaxes the functional form assumptions. Second, because CRS (2009) do not have data on non-test takers, they are forced to use the grouped data selection model, and assume that their group variable (school or state) is the only factor affecting test-taking. I model selection at the individual level and use measures of past and contemporaneous student achievement (eighth and eleventh grade test scores) of test-takers and non-takers to help predict latent college entrance exam scores.

3.3. Data

For the majority of the analyses in my paper, I use a new student-level data set containing two recent cohorts (2004–05 and 2007–08) of all first-time eleventh graders attending Michigan public high schools.[4] Because students who drop out before graduating, or who take the special education version of the eleventh grade test, are likely to not take the ACT in the 2008 cohort, I restrict my sample to high school graduates not taking the special education version of the eleventh grade test. To abstract away from changing demographics over time, I would ideally have liked to use the last pre-policy cohort (2006) and the first post-policy cohort (2007). However, several thousand students in the 2006 cohort were required to take the ACT as part of a pilot program the year before the statewide launch. Also, non-compliance was higher in the first post-policy cohort, as districts struggled to adapt to the reform. Thus, I use the 2005 and 2008 cohorts.

The data contains time-invariant demographic information such as sex, race, and date of birth, as well as time-varying characteristics such as free and reduced-price lunch status, limited-English-proficiency (LEP) status, special education (SPED) status, and student home addresses. The data also contains eighth grade and eleventh grade state-assessment results. For the cohorts of students after the implementation of the mandatory ACT exam, the eleventh grade results include ACT scores.

I have acquired and merged on several other key pieces of information. First, using student name, date of birth, sex, race, and eleventh grade home zip code, I matched the student-level Michigan data to micro-data from ACT Inc. and the College Board on every ACT-taker and SAT-taker in Michigan over the sample period. I also acquired from ACT Inc. a list of all ACT

[4] These data were provided by the Michigan Department of Education, Center for Educational Performance and Information, and Michigan Consortium for Educational Research.

test centers in Michigan over the sample period, including their addresses and their open and close dates. I geocoded student home addresses during the eleventh grade, and the addresses of these test centers, to calculate the driving distance from the student's eleventh grade home to the nearest ACT test center.[5]

Table 3.1 shows sample means for the combined sample and for the two cohorts separately, before and after implementation of the mandatory ACT. In the three years between these two cohorts of students, the demographic composition of Michigan students changed substantially. The percentage of eleventh graders who are black increased from 13% to 16%, while the percentage eligible for free lunch jumped from 20% to 28%. The local unemployment rate rose slightly from 7.3% to 7.7% during the sample period.[6] Prior to the mandatory policy, 64% of high school graduates took the ACT. The percentage who took it in the post-policy cohort is 98.5%. While 95% of students in each school are required to take the eleventh grade assessment for NCLB purposes, it is not technically a graduation requirement; hence this small remaining gap. Prior to the implementation of the mandatory ACT policy, 7.6% of high school graduates in Michigan took the SAT. After, 3.9% took it.

ACT-taking rates increased more for those groups of students who had lower rates prior to the policy. This is particularly pronounced among students eligible for free or reduced-price lunch, whose rate of ACT-taking more than doubled from 43% to 97%. These same groups tend to experience larger drops in their mean scores. The exception is black students who had only a slightly larger change in test-taking than white students, and whose mean score decreased by slightly less than white students.

3.4. Selection in College Entrance Exam Taking

I begin the exploration of selection into college entrance exam-taking by using the post-policy distribution to back out the predicted distribution of latent ACT scores of non-takers before the policy change. This allows me to describe where in the observed pre-policy score distribution these non-takers come from. I then proceed to compare the effectiveness of a variety of sample selection methods which correct for selection bias at the student and group-level.

[5] In the rare case when the eleventh grade home address is missing, I use the home address during the surrounding grades. Two percent of the sample has no non-missing addresses during any grade in high school, and are dropped from the analysis.

[6] Unemployment rates at the city (when available) or county level are from the Bureau of Labor Statistics.

3.4.1. Latent Scores of Non-Takers Pre-Policy

It is reasonable to expect that students not taking a college-entrance exam are lower achieving on average than those taking the test. Various analyses undertaken in papers correcting for selection in college-entrance exam scores make stronger assumptions about how negatively selected these students are. The Tobit analysis used in Angrist, Bettinger, and Kremer (2006), in which the authors artificially censor the scores of test-takers below various percentiles, assumes that all non-takers would score at or below that point in the distribution. Similarly the bounding exercises undertaken by that paper and in Krueger and Whitmore (2001) create upper bounds assuming that all non-takers would score below test-takers.

I test whether these assumptions are valid by predicting the ACT score distribution that we would see among non-takers during the pre-policy period if they were required to take the ACT. I do this by subtracting the number of test-takers scoring at each ACT score in the pre-period from the number scoring at each score in the post-period. Students who take the SAT but not the ACT pre-policy are included in all of my analyses, which prevents inaccurate categorization of students in the pre-policy period who took the SAT instead of the ACT as non-takers.[7]

Because the composition in Michigan shifts toward higher minority and more disadvantaged students from the pre- to the post-policy cohort, I reweight the post-policy cohort following DiNardo, Fortin, and Lemieux (DFL, 1996) to resemble the pre-policy students according to their observed characteristics. Using OLS, I estimate:

$$PRE_{is} = \alpha + \beta_1 X_{it} + \beta_2 S_{st} + \varepsilon_{ist}, \quad (5)$$

where $PRE_{is}$is an indicator for student *i* in school *s* being in the pre-period. *X* is a vector of individual level covariates, and *S* is a vector of school-level covariates.[8] I predict $\hat{PRE}_{is}$, which is the propensity score of being in the pre-policy period. The DFL weight equals: $\frac{\hat{PRE}_{is}}{(1-\hat{PRE}_{is})}$, which I censor at its first and ninety-ninth percentile and normalize so that the pre- and post-policy cohorts are of equal size. When summing the number of ACT-scorers in the pre-policy cohort,

[7] For students taking the ACT multiple times, I use their first score. For students taking the SAT but not the ACT, I include their SAT score scaled to the ACT metric. For students taking both tests, I use their first ACT score.
[8] *X* includes all interactions of LEP, SPED, free lunch status, race dummies, and a gender dummy. *S* includes fraction on free lunch, fraction black, number of eleventh graders, pupil-teacher ratio, student-guidance counselor ratio, and dummies for urban-rural status. The R-Squared from the regression is 0.149.

each student receives a weight of one, but when summing in the post-policy cohort, each student receives a weight equal to their censored DFL weight.

Assume that after the DFL-reweighting and cohort size adjustment the only difference between the pre- and post-period cohorts is that nearly everyone takes the ACT in the post-period. Then the difference in the number of students scoring at each ACT score bin should reflect the distribution of unobserved latent scores of students who did not take the exam before it was mandatory.

Figure 3.1a plots the frequency distribution of ACT scores pre-policy (circles), the reweighted post-policy distribution of scores (squares), and the difference, or the latent scores of non-takers pre-policy (triangles). While the latent scores of non-takers are shifted to the left relative to the test-takers, there is a long tail of students with reasonably high latent scores. Table 3.2 reports moments and percentiles of the three distributions. The patterns across the three distributions in the mean, variance, and skewness are all easily discernible from the figure: the non-takers have a lower mean, higher variance, and greater skew. Almost 60% of takers score at a college-ready level, while less than 30% of the non-takers would do so. I use a score of 20 to denote college-readiness, which ACT Inc. reports likely qualifies a student for admission to a "traditional" four-year institution (ACT Inc., 2002). Finally, Kolmogorov-Smirnov tests of equality between both the pre-policy non-taker and taker distributions, and the observed pre- and post-policy distributions are strongly rejected. This provides evidence that there is substantial selection into ACT-taking.

In Angrist, Bettinger, and Kremer (2006), the authors use Tobit analyses, censoring scores at the first and tenth (among other) percentiles. They suggest that while the assumption that non-takers would all score below the first percentile of observed scores is unlikely, it might be reasonable that they score below the tenth percentile. In Figure 3.1b, I plot the densities of pre-policy observed scores among takers and latent scores among non-takers. I mark with a vertical dotted line the first percentile of the observed score distribution of takers. As indicated at the top of the figure, 96% of non-takers have latent scores above the first percentile. Looking to the tenth percentile, we see that 67% of non-takers would still score above that mark; 23% would score above the median, 14% above the 75^{th} percentile, and 4% above the 95^{th} percentile. It is clear that while most non-takers would score in the lower half of the distribution, there are a non-trivial number of scores in the top quartile.

This suggests that the upper bounds of the exercises conducted in Krueger and Whitmore (2001) and in Angrist, Bettinger, and Kremer (2006) are quite far from the true treatment effects. In the latter paper, the authors present quantile-specific upper bounds that equal the treatment effect at that quantile under the assumption that no test-takers induced to take the test by the treatment would score above that quantile. The authors present these bounds at the 75th, 85th, and 95th percentiles of the control group distribution, with the magnitudes of the upper bound decreasing sharply as the quantiles increase. My analysis suggests that the reader should not interpret the upper bound as a good proxy for the true effect until perhaps the 95th percentile. Note that Angrist, Bettinger, and Kremer (2006) analyze scores on a college entrance exam in Colombia. This is a very different context than the analysis in my paper and may include a substantively different selection process. Thus, while my results likely extrapolate to Krueger and Whitmore (2001), this is less certain to be true for Angrist, Bettinger, and Kremer (2006).

3.4.2. Comparing Individual-Level Selection Bias Corrections

In this section, I test the performance of a number of sample selection correction methods at correcting for selection in college entrance exam scores. I examine whether researchers or state policy-makers in states without a mandatory college entrance exam could use available data to predict the mean score if all students were to take the test.[9]

In Table 3.3 (row 1, column 1), I report the mean ACT score in the post-policy period as 19.25. Since (nearly) all students take the test in my sample post-policy, I consider this the true mean latent score, or $E[ACT_i^*]$. I use this as a benchmark to see how well the selection corrections applied to the pre-policy can approximate this parameter. Before correcting the pre-period sample for selection, I first test how well OLS performs at simply predicting the ACT scores of non-takers using observed characteristics. Using the pre-policy students, I estimate the following equation using OLS:

$$ACT_{is} = \alpha + \beta X_{is} + \varepsilon_{is}, \quad (6)$$

where ACT_{is} is the ACT score of student *i* at school *s* in the 2005 cohort, X_{is} is a vector of basic student demographics including sex, race, and free lunch status, and ε_{is} is a mean zero error term clustered at the school level. Because researchers will have access to different levels of detail in

[9] In the current educational landscape of No Child Left Behind, the parameter of interest is often not a mean score but rather the fraction of students scoring above some threshold. In a future version of this paper, I will also gauge the performance of these corrections at predicting the fraction of students who score at a college-ready level.

their data, I start by estimating Equation (6) using only the most basic demographics in the prediction, and add more covariates later. I predict $\widehat{ACT_{is}}$ for all students in the pre-policy period whether they took the ACT or not, and compare this to the true $E[ACT_i^*]$ estimated using the post-policy data.[10]

The mean of the predicted values using OLS is 20.67 with a standard error of 0.10.[11] The raw mean of observed scores in the pre-policy period is 20.86. OLS using basic student demographics does very little to get closer to the true mean score of 19.25.

The next correction method that I test uses Tobit estimation on artificially censored test score data as in Angrist, Bettinger, and Kremer (2006). In that paper, the authors censor their data at the tenth percentile of observed scores, because their treatment group is approximately ten percentage points more likely to take the test than their control group. In the current paper, students after the ACT becomes mandatory are 34 percentage points (= 98.5 - 64.1) more likely to take the ACT. So, I estimate two Tobit regressions: In the first, I censor the data so that students not taking the ACT, and students who take the test but score below the tenth percentile of observed scores, are assigned the score associated with the tenth percentile; in the second Tobit regression, I censor the data at the thirty-fourth percentile. I estimate Equation (6) on this artificially censored data using Tobit and report the mean of the predicted ACT scores (Table 3.3, column 5 and 6). The Tobit estimates a mean score that is somewhat lower (20.48 using the tenth percentile, and 19.99 using the thirty-fourth), and closer to the truth of 19.25, than OLS, which estimates a predicted mean score of 20.67.

I next test the performance of the Heckman Two-Step correction procedure.[12] In the first step, I use probit to estimate:

$$TAKE_{is} = \delta + \gamma Z_{is} + u_{is}\ , \qquad (7)$$

where $TAKE_{is}$ is an indicator variable for ACT-taking, and Z_{is} equals X_{is}. In the second step, I estimate using OLS:

$$ACT_{is} = \alpha + \beta X_{is} + \lambda(\hat{\gamma} Z_{is}) + e_{is}, \qquad (8)$$

[10] Coefficients from this regression as well as from the regressions correcting for sample selection can be seen in Tables 3.A.2, 3.A.3, and 3.A.4.

[11] Unless otherwise noted, I calculate all standard errors throughout the paper by taking the standard deviation of the parameter across 200 bootstrap replications. I resample schools to maintain the clustered structure of the error term.

[12] I also attempt to estimate the correction using LIML instead of the 2-step estimator, but the likelihood function does not converge, foreshadowing the performance of the Heckman correction generally.

where $\lambda(.)$ is the IMR. The Heckman correction performs worse than Tobit and worse than OLS. The mean predicted score is 20.68, as compared to the true mean score of 19.25.

Because the performance of the Heckman correction without an exclusion restriction is usually poor (e.g., Puhani, 2002), I incorporate the use of an exclusion restriction. The variable that I add to *Z* but not to *X* is the student-level driving distance from a student's home to the nearest ACT test-center.[13] I hypothesize that some students have easier access to a test center than other students, and will thus have a slightly higher probability of taking the test. The identifying assumption is that conditional on *X*, these students are otherwise similar. The mean distance to a test center in the pre-policy period is 4.9 miles, and the median is 3.1 miles. This distance measure varies dramatically by urban/rural status, with the mean distance in urban areas equal to 2.3 miles and in rural areas 8.5 miles. For percentiles of the distribution pre- and post-policy by urban/rural status, see Table 3.A.1.

In Table 3.4, I examine the relationship between distance and test-taking. I estimate Equation (7) using OLS but add a quadratic in distance to the right hand side of the equation.[14] Without controlling for any other covariates, there is no statistically significant relationship between distance and ACT-taking pre-policy. A test that the two terms are jointly equal to zero has an F-statistic of 6.3. However, as I start to add covariates, a relationship emerges. When I control for basic student demographics, the coefficients on the distance and distance squared terms become statistically significant, and the F-statistic rises to above ten, the rule-of-thumb threshold for weak instruments (Staiger and Stock, 1997). Moving from the 25th to the 75th percentile of distance is associated with a decrease in the ACT-taking probability of 2.6 percentage points off a baseline rate of 64%. Further controlling for school and district covariates (including urban / rural status, and average eighth grade and eleventh grade test scores), and student-level eighth and eleventh grade test scores slightly attenuates this relationship, but the F-statistic is stronger at 13.2.[15]

In columns (6) through (10), I informally test the validity of the exclusion restriction. I replace the dummy for ACT-taking as the dependent variable with the average of the student's

[13] I use a student's home address during eleventh grade. In the rare case that a student has multiple addresses during eleventh grade, I use the one with the shortest distance to a center.
[14] Standard errors are clustered at the school level and calculated without the bootstrap.
[15] All test scores are standardized within cohort and grade to have mean of zero and standard deviation of one. For eighth grade, I use the average of a student's math and English scores, where both are standardized before taking the average. For eleventh grade, I use social studies scores since post-policy math and English scores are in part determined by a student's ACT score.

eleventh grade math and English test scores. Without controlling for any covariates, students living farther away from an ACT-test center tend to have higher scores. The point estimates reported in column (6) suggest that moving from the 25th to 75th percentile of distance is associated with a 0.11 standard deviation higher 11th grade score. This is likely due to the same reason that there is no relationship between distance and test-taking before controlling for covariates: disadvantaged, low-performing students in Detroit and other urban areas live very close to a center. However, as I add covariates, the relationship disappears, suggesting that the distance measure is affecting ACT-taking but not latent scores, and thus providing evidence that the exclusion restriction is valid.

I re-estimate Equations (7) and (8) after including the distance variable in Equation (7) and add the results to Table 3.3 (column 7). The instrument hardly improves the performance of the Heckman correction. The predicted mean ACT score barely budges from 20.68 to 20.67, and is still far from the true mean score of 19.25.

Soon after Gronau and Heckman's development of the bivariate normal selection correction, other authors developed similar parametric corrections that slightly loosened the distributional assumptions. Olsen (1980) showed that joint normality of the error terms is an overly strong assumption. He finds that the necessary assumption is that the distribution of *u*, *F(u)*, is known, and that *e* is a linear transformation of this distribution. If *F(u)* is normal, this amounts to joint normality. He derives a correction assuming *u* is distributed uniformly on the interval (0, 1). With this distribution of *u*, the first stage can be estimated using a linear probability model (LPM) rather than probit. Olsen derives that the term that controls for the conditional error term in the outcome equation is no longer $\lambda(\hat{\gamma}Z_i)$ but rather $\rho\sigma_e\sqrt{3}(Z_i * \hat{\gamma} - 1)$. This approach requires an exclusion restriction because there is no non-linear IMR to help with identification. I test the performance of this correction by estimating Equation (7) using OLS and including $(\hat{\gamma}Z_{isdt} - 1)$ in Equation (8) in place of $\lambda(\hat{\gamma}Z_{isdt})$. The Olsen correction does no better at predicting the post-policy mean ACT score. It is again 20.67, identical to that predicted using the Heckman correction with an exclusion restriction.[16]

[16] Two other ways of estimating a similar correction assume a χ^2 or Student-T error distribution for the error term rather than the uniform or normal distribution (Lee, 1982; Heckman, Tobias, and Vytlacil, 2003). I attempt this procedure, but like the Heckman correction using LIML, the maximum likelihood estimator does not converge.

The final method I test to correct for selection bias is semiparametric, in that it relaxes the distributional assumptions about the error terms in both the outcome and selection equations. I estimate the selection equation (7) using nonparametric multivariate kernel regression to estimate the propensity score, $P[TAKE_{ist} = 1|Z_{ist}]$. Using generic notation for ease of exposition, a kernel regression estimator calculates $E[Y|X_0]$ by taking the weighted average of Y_i for observations with X_i close to X_0, and weighting by how near X_i is to X_0. Formally, the kernel estimator, $\hat{g}(X_0)$, equals:

$$\frac{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{X_i - X_0}{h}\right) Y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{X_i - X_0}{h}\right)}, \tag{9}$$

where N is the sample size, h is a bandwidth parameter, K is the kernel weighting function, and Y and X are the dependent and independent variables of interest. Reverting back to the notation in this paper, the kernel estimator uses the data to predict $\widehat{TAKE}_{ist}$ without any assumptions about the distribution of u_{ist}, or the functional form of the relationship between $TAKE_{ist}$ and Z_{ist}. However, it does not provide a $\hat{\gamma}$, so if I am interested in understanding the factors that affect ACT-taking, then this strategy is less useful.[17,18]

In the second step of the semiparametric correction method, I estimate:

$$ACT_{ist} = \alpha + \beta X_{ist} + g(\hat{\gamma} Z_{ist}) + e_i, \tag{10}$$

approximating the control function term, $g(\hat{\gamma} Z_{ist})$ following Newey (1988) using a power series in $P[TAKE_{ist} = 1|Z_{ist}]$.[19] Newey suggests using either a series in the predicted value from the first step or a series in the inverse Mills ratio evaluated at the predicted value. I follow Newey, Powell, and Walker (1990) and use the series in the IMR evaluated at the propensity score.[20] I

[17] In estimating the multivariate kernel regression, I follow methods for discrete and continuous data developed in Racine and Li (2004) and Li and Racine (2004), using cross-validation to choose the bandwidth.

[18] There are also several techniques available to estimate the selection equation semiparametrically by assuming $E[e_i|Z_i, TAKE_i = 1] = g(\gamma Z_i)$, where *g(.)* is called a single-index function (Ichimura, 1993; Klein and Spady, 1993). This allows for estimation of $\hat{\gamma}$. However, I prefer to use the non-parametric kernel regression since the single-index assumption that γ and Z_i can be neatly aggregated into a scalar, where selection is assumed to depend only on this scalar, is a still a strong one.

[19] Newey developed the procedure to be used having estimated the single-index model in the first step. Choi (1993) derives the statistical properties of the Newey correction using the propensity score from the first step. There are alternative methods of semiparametrically estimating the second step besides the Newey correction (Powell, 1987; Ahn and Powell, 1993). I focus on the one Newey method for brevity.

[20] Newey, Powell, and Walker (1990) use just two terms for the series, which they choose using cross-validation. Given my larger sample size, and that according to Newey (1988) the number of terms should increase with the sample size, I use a quintic in the p-score. In a future version of this paper, I will use cross-validation to verify the correct choice. I have explored results using between two and ten terms, and results are similar regardless.

report the results using the Newey correction in column 9. It predicts a post-policy mean ACT score of 20.65. This is slightly closer to the true mean score of 19.25 than OLS, and the Olsen and Heckman corrections, which predict mean scores of 20.67 and 20.68. But it still performs worse than Tobit on the artificially censored score data, which predicts a mean score of 20.48.

I now examine whether a researcher who has access to school- and district-level covariates including demographics, urbanicity, and average eighth and eleventh grade test scores can do a better job at correction for selection in ACT scores. The pattern of results remains the same when adding these covariates to Equation (6) and re-estimating the equation. All of the specifications do a somewhat better job of attaining the post-policy mean score than using only the student demographics; but again, none of the corrections do much better than OLS.

Finally, I include student-level eighth and eleventh grade test scores in the prediction equation.[21] A state analyst would presumably have access to this data, though it is rare for a researcher to have such data matched to a student's college entrance exam score. The performance of all of the corrections, and of simple OLS, is much better using the student-level scores in the prediction. This is perhaps unsurprising, as we would expect a student's past and contemporaneous achievement to be an excellent predictor of their ACT score. The Tobit and Newey corrections both do considerably better than OLS, attaining or nearly attaining the post-policy mean.[22]

As another way to compare the performance of these estimators and the use of different levels of data, I plot kernel densities of the actual and predicted ACT scores. Figure 3.2a plots raw ACT scores pre- and post-policy, as well as the predicted ACT scores from Equation (6) including both student demographics and school- and district-level covariates. Also plotted are 95% confidence intervals of the predicted values. Note that the plots of the predicted values have a much smaller variance than the true scores, which makes sense because the variance of the actual ACT scores will equal the variance of the predicted values, plus the variance of the

[21] Note that for all analyses using student- or school-level eleventh grade scores, social studies scores are used, which during post-policy period are from a separate state test from the ACT and thus are not mechanically affected by a student's ACT scores.

[22] Estimating the specifications including eighth grade scores, but not eleventh grade scores, performs better than the specifications without any student-level measures of achievement, but worse than those that include eighth and eleventh grade scores. For example, the predicted mean score using OLS and eighth grade scores is 19.91. It is 20.48 without any student scores, 19.52 using both eighth and eleventh grade scores, and is 19.25 using the post-policy data. Complete results using only eighth grade scores are available from the author upon request.

residuals.[23] Figure 3.2b shows the same picture but plots the predictions adding the student-level test scores.[24] As was apparent in Table 3.3, the fit of the prediction is substantially better. Figure 3.3 zooms in on Figure 3.2b, drops the actual pre- and post-distributions, and compares the densities of $\widehat{ACT}_{is}$ estimated on the pre-policy data, using a few of the corrections to the densities of $\widehat{ACT}_{is}$ pre- and post-policy from OLS. As in Table 3.3, the Heckman correction (with the IV) is very similar to the OLS pre-policy regression, whereas the Tobit (censoring at the tenth percentile) and Newey corrections are between the pre- and post-policy OLS densities.

Although none of the sample selection estimators do a particularly good job at correcting for selection bias when applied to the restricted covariate set, Tobit estimation on artificially censored ACT score data often performs better than the other corrections. This is somewhat surprising given the strong assumptions of this model. In the previous analyses presented in Table 3.3, I censor the pre-policy ACT score distribution at the tenth percentile following Angrist, Bettinger, and Kremer (2006), and also at the thirty-fourth percentile. In Figure 3.4, I more closely examine how the performance varies by the percentile at which I censor the data. The horizontal dotted line gives the mean ACT score post-policy as a measure of the true, unselected mean score. I first include only basic student demographics in the Tobit estimation (circle-shaped markers). The predicted mean generated by Tobit gets closer to the true post-policy mean as the censoring point increases. Tobit matches the post-policy mean using approximately the seventy-fifth percentile as the censoring point, and over-corrects using higher percentiles.

The next plot (triangle-shaped markers) reports the predicted mean latent ACT score using student demographics and school characteristics. The line is very similar to the previous estimation, though reaches the true mean at a slightly lower censoring percentile of sixty. Additionally including eighth and eleventh grade student test scores (square shaped markers) shifts the line downward substantially, such that the optimal performance of Tobit occurs when the censoring point is near the tenth percentile. It appears that there is no clear censoring point that works best for all data detail levels.

[23] A more accurate way to plot the predicted scores would be to add pulls from the distribution of residuals to the predicted scores. I reserve this for a future draft. For now, while the predicted distributions from all of the estimators will be inaccurate, I can still compare the relative performance of each, as they are all inaccurate in a similar way.
[24] I omit confidence intervals on the densities of predicted values for readability.

In addition to the overall mean ACT score, administrators and policy-makers are interested in how this parameter varies across key student subgroups. While I have shown that using student-, school-, and district-level demographics and test scores allows for a highly accurate prediction of true latent scores for all students, the nature of selection into test-taking may differ across groups. The latent scores of some students may be easier or more difficult to predict using observed characteristics.

I explore this in Table 3.5 by estimating Equation (6), including the full set of covariates, separately by race and by free-lunch status. There is a large gap in mean observed scores pre-policy between black and white students, and between poor and non-poor students. As we saw in Table 3.1, the difference in the change in test-taking rates between poor and non-poor students due to the mandatory ACT policy is large (non-poor students experience a 30pp increase while poor students experience a 54pp increase). As expected, the gap in scores between these two groups increases by 0.4 points.

With the exception of the Newey correction, which performs quite poorly at predicting the post-policy mean ACT score among whites, OLS and the corrections tend to understate the post-policy gap in scores. This is due to overestimating mean scores more for black students than for white students. Apparently, black students are selecting into ACT-taking in a way that is more difficult to predict based on observed characteristics. The same phenomenon occurs for the poor and non-poor students. OLS and the correction methods seem to do a better job among non-poor students, though at times they overcorrect and predict lower than the true mean latent ACT score. For poor students, they tend not to correct for selection enough to attain the true mean. Both of these forces work in the direction of underpredicting the true gap in scores.

Looking across the subgroups, the relative performance of the corrections is a bit more varied than for the entire sample. Each of the corrections performs best for certain subgroups, but none seems to consistently best-estimate the post-policy mean ACT score for every group.

3.4.3. Comparing Group-Level Corrections

Many researchers using ACT/SAT micro-data as their dependent variable do not observe students who do not take the exam, and so cannot estimate an individual's probability of test-taking (Card and Payne, 2002; Rothstein, 2006). Following Gronau (1974), their strategy is to use group-level data where group mean scores are the dependent variable and the control function term is the IMR (or some other function) evaluated at the group-level participation rate.

This approach is equivalent to the two-step individual-level Heckman correction, where Z_i is comprised of the student-level variables used to aggregate to the group level.

This equivalence is rarely pointed out, and implies that the level of aggregation matters. If the data are grouped at the school level as in Rothstein (2006), then the control function approach assumes that the only factor affecting test participation (i.e., that is in Z_i) is the school that a student is in. This is equivalent to assuming that variation in test-taking within a school is due only to the disturbance term. Likewise in Card and Payne (2002), data are grouped at the state-year-parental education level. For their control function to fully correct for selection, test-taking must depend only on the state, year, and education of a student's parents.[25]

In this section, I estimate a variety of group-level parametric selection corrections with data aggregated at increasingly fine levels. I compare the results to the "truth" estimated in the mandatory ACT period. Based on these results, I can suggest a "best practice" method to researchers seeking to correct for selection at the group level in their own work.

I begin by estimating the following equation separately for both the pre- and post-policy period using weighted least squares, where the weights are equal to the number of eleventh graders in the school during that cohort:

$$ACT_s = \alpha + \beta X_s + e_s, \quad (11)$$

where ACT_s is the mean ACT score at school s, and X_s is a vector of school-level covariates, including the fraction of black students, the fraction on free lunch, the teacher-pupil ratio, the average eleventh grade social studies score (standardized across individuals at the grade-year level), and the average eighth grade math and English scores.[26]

I report the predicted mean ACT score in Table 3.6. The observed mean ACT score post-policy is 19.28, and pre-policy it is 20.63. Using the pre-policy data, OLS does little to correct for selection, producing a predicted post-policy mean ACT score of 20.59 (row 1, column 4). I then re-estimate Equation (11) inserting $H(\hat{p})$, where $\hat{p}$ is the school-level ACT participation rate, and $H(.)$ is a control function term. I first set $H(\hat{p}) = \hat{p}$. Then I set $H(\hat{p}) = \log(\hat{p})$, and

[25] Card and Payne (2002) discuss this in a working version of their paper: "This specification is consistent with a conventional joint-normal model of latent test scores and test participation in which the probability of writing the SAT is assumed to depend on a set of factors that are identical for individuals in the same family-background-state-year cell. More generally, it can be interpreted as an approximation to the selectivity adjustment implied by an arbitrary model of test score outcomes and test participation in which test participation depends on a single index that is 'fully absorbed' by family background x state x year dummies (Ahn and Powell, 1993)."

[26] I drop schools that do not appear in both 2005 and 2008 with at least one ACT-taker. 2% of students in my sample are dropped due to this restriction.

$H(\hat{p}) = \lambda(\hat{p})$, where $\lambda(.)$ is the inverse Mills ratio.[27] The control functions improve slightly on the uncorrected OLS regression, producing predicted mean scores of 20.39, 20.45, and 20.40, respectively. However, they remain far from attaining the unselected mean ACT score of 19.28. As found in Card and Payne (2002) and Rothstein (2006), the results are not sensitive to which control function is used; yet, it appears that this is because they equally miss the mark.

We might also think that the degree of selection in ACT-taking could vary with X_s. In other words, selection into test-taking might play out differently at schools serving different types of students. To address this possibility, I allow the control function term to vary by various school-year characteristics such as the fraction free lunch and the mean eleventh grade test scores. The results are typically no better than when using the control function terms without an interaction; and, in fact, the interaction with test scores (row 1, column 10) performs slightly worse, predicting a post-policy mean ACT score of 20.58.

Finally, I test the performance of the above set of group-level control functions, grouping the data at increasingly refined levels of aggregation. First, I create cells at the school-by-free lunch status-by-minority status level. The post-policy mean score, 19.28, is identical to before. The pre-policy predicted mean, 20.53, is slightly closer to the post-policy mean than the 20.59 estimated using the school-level data. Though the standard errors preclude firm conclusions, the corrections, which estimate predicted means between 20.19 and 20.52, do seem to get slightly closer to the truth when the cells are more refined.

I then group the data at the school-by-free lunch status-by-minority status-by-eleventh grade test score quartile level. First, note that the raw mean score stays the same in the post-period (19.28), but drops substantially in the pre-period from 20.59 to 19.96.[28] However, conditional on this drop, as with the individual-level selection model when I allow student scores to enter the model, OLS in the pre-period does a better job at approximating the true post-policy mean score. In the first two levels of aggregation, OLS predicts nearly the same mean ACT score that is observed in the pre-period data, doing little to correct for selection. For this most refined

[27] I also include a quadratic and cubic in p, but do not report results as they perform consistently worse than the linear term alone.

[28] The reason for the change is that the sample is changing slightly. Students with missing eleventh grade scores are now dropped since they do not fall into a test score quartile. Also, while the school-level sample was conditioned on only appearing if there was at least one ACT-taker in the school in both years, there are cells at the more refined level of aggregation, for example poor minority students with the lowest test scores, in which there are no ACT-takers. These students are not contributing to the average ACT scores in less refined cells but are included in the weighting.

cell level, OLS using pre-policy data predicts a post-policy mean score of 19.61, which is somewhat lower than the raw pre-policy mean of 19.96. The control functions slightly improve upon OLS, predicting mean latent scores between 19.45 and 19.51 (row 3, columns 5-10). These are all within the 95% confidence interval of the 19.38 mean score predicted by OLS using post-policy data (column 2).

In summary, none of the control functions stands out as the clear choice across each of the levels of aggregation, and a researcher's best bet to avoid selection bias is to obtain data at the most detailed level of aggregation possible. In particular, aggregating the data by performance levels on a measure of student achievement outside of the college entrance exam is particularly effective.

3.5. Conclusions

College entrance exam scores on the ACT and SAT provide a powerful measure of college-readiness. States, districts, and schools can use these scores to diagnose their performance at preparing their students for post-secondary education. Researchers have used these scores to evaluate the effects of educational interventions (Krueger and Whitmore, 2001; Card and Payne, 2002; Angrist, Bettinger and Kremer, 2006; Rothstein, 2006). The main drawback is that less than half of public high school students take the ACT or SAT. Without knowing the degree of selection into test-taking, an administrator or policy-maker's ability to harness these scores as a true measure of college-readiness that is representative of the overall population is limited. Researchers have attempted to control for the resulting selection bias using a variety of parametric methods.

I use the implementation of a policy in Michigan requiring all eleventh graders to take the ACT to compute the distribution of latent scores for students who were not taking the test prior to the policy. I show that the assumptions made by certain correction and bounding methods—such as, for example, that all students not taking the test have lower latent scores than all test-takers, or score below some percentile of the observed distribution—are far from true. I further use the near complete distribution of ACT scores post-policy as the true distribution of latent scores, to which I compare the performance of several parametric and semiparametric selection correction methods using pre-policy data. I show that none of the correction methods do a particularly good job at predicting the post-policy mean latent ACT score, when applied using

only basic demographics and with distance to the nearest test center as an exclusion restriction. However, with enough information about students—in particular, with student-level measures of achievement such as state-administrated standardized test scores—simple OLS does a good job at correcting for selection bias. I also show that group-level correction methods using a control function evaluated at the test-taking rate of the group perform poorly, regardless of the functional form of the control function. Aggregating the groups to increasingly refined cells seems to help correct for selection bias, in particular when including test score levels in the aggregation.

In summary, the richness of the data used to model selection into college entrance exam-taking matters far more than the econometric method used to correct for selection. This may come as a disappointment to researchers working with data that do not contain measures of student achievement in addition to their college entrance exam score. However, this is good news for states considering the implementation of a mandatory ACT or SAT policy and who hope to predict their future mean score. While they should not rely on being able to accurately forecast differences in scores across groups, simple OLS regression using data that contains demographic and test score information of students and their schools can come very close to providing the overall picture of college-readiness that they seek.

References

Ahn, Hyungtaik and James L. Powell. 1993. "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." *Journal of Econometrics* 58: 3–29.

Angrist, Joshua, Eric Bettinger and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review* 96(3): 847–862.

Banchero, Stephanie. 2002. "Everyone's Into ACT, and it Shows." *Chicago Tribune*, August 21.

Card, David and A. Abigail Payne. 2002. "School Finance Reform, the Distribution of School Spending, and the Distribution of Student Test Scores." *Journal of Public Economics* 83: 49–82.

Choi, Kyungsoo. 1993. "Semiparametric Estimation of Sample Selection Model Using Series Expansion and the Propensity Score." Unpublished dissertation, University of Chicago.

Clark, Melissa, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2009. "Selection Bias in College Admissions Test Scores." *Economics of Education Review* 28: 295–207.

Desjka, Joe. 2012. "Testing Way for More to Take ACT." *Omaha World-Herald*, February 13.

DiNardo, John, Nicole Fortin and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica* 64(5): 1001–1044.

Dunlap Tribune Staff Writer. 2011. "ACT Scores Have Fallen, Are Area Students Ready for Next Step?" *Dunlap Tribune*, Dunlap Tennessee, February 25.

Dynarski, Mark. 1987. "The Scholastic Aptitude Test: Participation and Performance." *Economics of Education Review* 6(3): 263–273.

Dynarski, Mark, and Philip Gleason. 1993. "Using Scholastic Aptitude Test Scores As Indicators of State Educational Performance." *Economics of Education Review* 12(3): 203–211.

Gronau, Reuben. 1974. "Wage Comparisons – A Selectivity Bias." *Journal of Political Economy* 82(6): 1119–1143.

Heckman, James J. 1974"Shadow Prices, Market Wages, and Labor Supply." *Econometrica* 42(4): 679–694.

Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5(4): 475–492.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(l): 153–161.

Heckman, James J. and Richard Robb Jr. 1985. "Alternative methods for evaluating the impact of interventions." In *Longitudinal Analysis of Labor Market Data*, Eds, James J. Heckman and Burton Singer, Cambridge University Press.

Heckman, James, Justin L. Tobias, and Edward Vytlacil. 2003. "Simple Estimators for Treatment Parameters in a Latent-Variable Framework." *Review of Economics and Statistics* 85(3): 748–755.

Hetzner, Amy. 2010. "Low Test Scores Worry Districts." *Milwaukee Journal Sentinel*, August 18.

Hyman, Joshua M. 2013. "ACT for All: The Effect of Mandatory College Entrance Exams on Postsecondary Attainment and Choice." Working Paper.

Ichimura, Hidehiko. 1993. "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models." *Journal of Econometrics* 58: 71–120.

Klein, Roger W. and Richard H. Spady. 1993. "An Efficient Semiparametric Estimator for Binary Response Models." *Econometrica* 61(2): 387–421.

Krueger, Alan B. and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *Economic Journal* 111: 1–28.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76: 1071–1102.

Lee, Fung-Lei. 1982. "Some Approaches to the Correction of Selectivity Bias." *Review of Economic Studies* 49: 355–372.

Lee, Fung-Lei. 1983. "Generalized Econometric Models With Selectivity." *Econometrica* 51(2): 507–512.

Li, Q. and J.S. Racine. 2004. "Cross-validated local linear nonparametric regression." *Statistica Sinica* 14: 485–512.

Newey, Whitney K. 1988. “Two Step Estimation of Sample Selection Models.” Unpublished manuscript.

Newey, Whitney K., James L. Powell, and James R. Walker. 1990. “Semiparametric Estimation of Selection Models: Some Empirical Results.” *American Economic Review Papers and Proceedings* 80(2): 324–328.

Olsen, Randall J. 1980. “A Least Squares Correction for Selectivity Bias.” *Econometrica* 48(7): 1815–1820.

Powell, James L. 1987. “Semiparametric Estimation of Bivariate Latent Variable Models.” Working paper no. 8704, *Social Systems Research Institute*, University of Wisconsin, Madison, WI.

Puhani, Patrick A. 2002. “The Heckman Correction for Sample Selection and its Critique.” *Journal of Economic Surveys* 14(1): 53–68.

Racine, J.S. and Q. Li. 2004. “Nonparametric estimation of regression functions with both categorical and continuous Data.” *Journal of Econometrics* 119: 99–130.

Robinson, Peter M. 1988. “Root-N-Consistent Semiparametric Regression.” *Econometrica* 56(4): 931–954.

Rothstein, Jesse. 2006. “Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions.” *American Economic Review* 96(4): 1333–1350.

Staiger, Douglas and James H. Stock. 1997. “Instrumental Variables Regression with Weak Instruments.” *Econometrica* 65(3): 557–586.

Table 3.1. Sample Means of Michigan 11th Grade Cohorts of 2005 and 2008

	Both Cohorts	2005 Cohort	2008 Cohort	Difference (3) - (2)	P-Value (4)=0
	(1)	(2)	(3)	(4)	(5)
Demographics					
Female	0.516	0.514	0.517	0.003	0.226
White	0.790	0.805	0.775	-0.030	0.000
Black	0.145	0.132	0.158	0.026	0.000
Hispanic	0.029	0.027	0.031	0.004	0.000
Other race	0.035	0.036	0.035	0.000	0.600
Free or reduced lunch	0.242	0.204	0.279	0.075	0.000
Local Unemployment	7.518	7.285	7.745	0.460	0.000
Driving miles to nearest ACT test center	3.71	4.87	2.58	-2.29	0.000
Took SAT	0.058	0.076	0.039	-0.037	0.000
SAT Score	25.2	24.8	25.9	1.0	0.000
Took SAT & ACT	0.054	0.070	0.039	-0.031	0.000
Took ACT or SAT					
All	0.815	0.641	0.985	0.345	0.000
Male	0.793	0.598	0.984	0.387	0.000
Female	0.836	0.681	0.986	0.305	0.000
Black	0.780	0.575	0.947	0.372	0.000
White	0.822	0.652	0.993	0.341	0.000
Free or reduced lunch	0.749	0.434	0.970	0.536	0.000
Not free or reduced lunch	0.838	0.693	0.991	0.299	0.000
First ACT or SAT Score					
All	19.9	20.9	19.3	-1.6	0.000
Male	19.9	21.0	19.2	-1.8	0.000
Female	19.9	20.7	19.3	-1.4	0.000
Black	16.0	16.8	15.6	-1.2	0.000
White	20.6	21.4	20.0	-1.5	0.000
Free or reduced lunch	17.1	18.3	16.8	-1.5	0.000
Not free or reduced lunch	20.7	21.3	20.2	-1.1	0.000
Number of Students	197,016	97,108	99,908		

Notes: The sample is first-time eleventh graders in Michigan public high schools during 2004-05 and 2007-08 who graduate high school, do not take the SPED eleventh grade test, and have a non-missing home address. Free lunch lunch status is measured as of eleventh grade.

Table 3.2. ACT Score Distributions Pre- and Post-Policy

	2005 Cohort		2008
	Takers	Non-Takers	Cohort
	(1)	(2)	(3)
Moments			
Mean	20.85	17.54	19.70
Variance	4.53	5.05	4.98
Skewness	0.31	1.04	0.42
Leptokurosis	2.72	3.66	2.66
Percentiles			
1st	12	10	11
5th	14	12	12
10th	15	12	13
25th	17	14	16
Median	21	16	19
75th	24	20	23
90th	27	25	27
95th	29	28	29
99th	32	33	32
Fraction Scoring>=20	0.588	0.276	0.479
K-S Test vs Column 1			
D-Stat		0.344	0.157
P-Value		0.000	0.000
Number of Students	62,185	33,486	95,671

Notes: The sample is as in Table 3.1. The reported number of students in the 2008 cohort is adjusted to match the size of the 2005 cohort and also includes only the 98.5% of the sample who take the ACT. Column (2) reports the distribution of latent ACT scores of students not taking the exam calculated using the methodology described in the text. The K-S Test is a Kolmorogov-Smirnov non-parametric test of the equality of the distributions.

Table 3.3. Mean Latent ACT Score by Correction Method and Control Variables

					Pre-Policy, by Correction Method					
	Post-Policy ("Truth")		Pre-Policy (Biased)		Tobit		Heckman			
	Raw	OLS	Raw	OLS	10th Pctl	34th Pctl	No IV	With IV	Olsen	Newey
Control Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Student Demographics	19.25	19.25	20.86	20.67	20.48	19.99	20.68	20.67	20.67	20.65
		(0.11)		(0.10)	(0.10)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)
…Plus School-Level Covs	19.25	19.25	20.86	20.48	20.26	19.72	20.49	20.49	20.48	20.47
		(0.11)		(0.09)	(0.10)	(0.11)	(0.09)	(0.09)	(0.09)	(0.09)
…Plus Student Test Scores	19.25	19.25	20.86	19.52	19.25	18.59	19.59	19.59	19.52	19.29
		(0.11)		(0.09)	(0.10)	(0.11)	(0.09)	(0.09)	(0.09)	(0.11)

Notes: The sample is as in Table 3.1. Columns (1) and (3) give raw mean ACT scores for each sample. Cells in columns (2) and (4) - (10) report the mean predicted ACT score from regressions of ACT scores on covariates. The predicted ACT score is calcuated for ACT-takers and non-takers. Standard errors calculated using 200 bootstrap replications resampling schools.

Table 3.4. Testing the Exclusion Restriction: the Relationship Between Test Center Proximity, Test-Taking, and Achievement

	Dependent Variable = Took the ACT					Dependent Variable = 11th Grade Test Score				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Distance to Center										
Distance (miles)	-0.003	-0.009***	-0.005***	-0.006***	-0.006***	0.030***	-0.003	0.005**	0.001	0.002
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.007)	(0.005)	(0.002)	(0.002)	(0.002)
Distance Squared (/ 10)	-0.000	0.003***	0.002***	0.003***	0.003***	-0.014***	-0.002	-0.002*	-0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)
Student-Level Demographics	N	Y	N	Y	Y	N	Y	N	Y	Y
School- & District-Level Covs	N	N	Y	Y	Y	N	N	Y	Y	Y
Student-Level Test Scores	N	N	N	N	Y	N	N	N	N	Y
R-Squared	0.001	0.062	0.078	0.112	0.261	0.003	0.111	0.165	0.203	0.644
F-Stat	6.27	11.29	6.69	10.62	13.22	15.89	9.90	2.20	0.05	0.41
Sample Size	97,108	97,108	97,108	97,108	97,108	86,680	86,680	86,680	86,680	86,680

Notes: The sample is as in Table 3.1 but includes only the 2005 eleventh grade cohort. Distance is driving distance in miles from the student's home address during eleventh grade to the nearest ACT test center. The distance-squared term is divided by ten for interpretability. The dependent variable in columns (1)-(5) is a dummy for taking the ACT (mean = 0.64), and in columns (6)-(10) is the average of eleventh grade math and English test scores standardized to have mean zero and SD one. The drop in sample size between columns (1)-(5) and (6)-(10) is due to missing eleventh grade test scores. Student-level test scores included as covariates are average math and English eighth grade score and eleventh grade social studies score. See text for the complete list of covariates. Standard errors clustered at the school-level. *** indicates statistical significance at the 0.01 level, ** at the 0.05 level, and * at the 0.01 level.

Table 3.5. Race and Poverty Gaps in Mean Latent ACT Scores by Correction Method

	Black	White	Gap	Poor	Non-Poor	Gap
	(1)	(2)	(3)	(4)	(5)	(6)
Post-Policy						
Raw	15.61	19.98	4.38	16.77	20.19	3.42
OLS	15.61	19.98	4.38	16.77	20.19	3.42
	(0.17)	(0.09)	(0.19)	(0.07)	(0.10)	(0.11)
Pre-Policy						
Raw	16.76	21.44	4.68	18.29	21.28	3.00
OLS	16.04	20.07	4.03	17.21	20.12	2.91
	(0.19)	(0.08)	(0.21)	(0.08)	(0.09)	(0.11)
Tobit (10th Pctl)	15.77	19.81	4.05	16.97	19.82	2.85
	(0.20)	(0.09)	(0.23)	(0.08)	(0.10)	(0.12)
Heckman	16.06	20.14	4.07	17.23	20.19	2.97
	(0.17)	(0.09)	(0.2)	(0.08)	(0.09)	(0.11)
Olsen	16.04	20.07	4.03	17.21	20.12	2.91
	(0.19)	(0.08)	(0.21)	(0.08)	(0.09)	(0.11)
Newey	15.99	21.06	5.07	16.88	19.95	3.07
	(0.20)	(1.01)	(1.02)	(1.26)	(0.10)	(1.26)

Notes: The sample is as in Table 3.1. Cells report the mean predicted ACT score from regressions of ACT scores on the full set of covariates, including student-level eighth and eleventh grade test scores. The predicted ACT score is calcuated for ACT-takers and non-takers. Poverty status is proxied for using free lunch receipt measured during eleventh grade. Standard errors calculated using 200 bootstrap replications resampling

Table 3.6. Group-Level Mean Latent ACT Score by Control Function and Level of Aggregation

	Post-Policy (Truth)		Pre-Policy (Biased)		Pre-Policy, By Control Function Term					
	Raw	OLS	Raw	OLS	p	ln(p)	IMR(p)	p*Lunch	IMR(p)* lunch	IMR(p)* Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
School	19.28	19.29	20.63	20.59	20.39	20.45	20.40	20.43	20.45	20.58
		(0.10)		(0.11)	(0.13)	(0.11)	(0.12)	(0.11)	(0.11)	(0.11)
Schl-Free Lunch-Minority	19.28	19.28	20.59	20.53	20.19	20.30	20.22	20.32	20.34	20.52
		(0.10)		(0.10)	(0.12)	(0.11)	(0.12)	(0.11)	(0.11)	(0.1)
Schl-Free Lunch-Minority-Test Score Quartile	19.28	19.38	19.96	19.61	19.49	19.51	19.48	19.50	19.50	19.45
		(0.10)		(0.10)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)

Notes: The sample is as in Table 3.1 but excludes the 2% of the sample who enroll in high schools that do not appear in both 2005 and 2008 with at least one ACT-taker. Cells report the mean predicted ACT score from group-level regressions of average ACT score on group-level covariates. Free lunch status measured as of eleventh grade. IMR=inverse Mills ratio. Standard errors calculated using 200 bootstrap replications resampling schools.

Table 3.A.1. Summary Statistics of Distance from Student Home to Nearest Test Center

	Overall			Urban		Rural	
	Total	Pre	Post	Pre	Post	Pre	Post
Mean	3.71	4.87	2.58	2.32	1.33	8.54	4.02
SD	3.89	4.67	2.47	1.79	0.90	5.90	3.29
Percentiles							
1st	0.2	0.3	0.2	0.3	0.2	0.4	0.2
5th	0.5	0.7	0.4	0.6	0.3	1.1	0.4
10th	0.7	1.0	0.6	0.7	0.4	1.8	0.7
25th	1.2	1.7	1.0	1.2	0.7	4.0	1.6
Median	2.4	3.1	1.8	1.9	1.1	7.5	3.3
75th	4.7	6.5	3.4	2.9	1.7	12.0	5.5
90th	8.6	11.5	5.7	4.2	2.4	16.6	8.1
95th	11.9	14.8	7.4	5.3	3.0	19.5	9.8
99th	18.7	21.1	11.2	9.7	4.6	26.7	15.1
Sample Size	197,016	97,108	99,908	20,433	20,859	25,195	25,858

Notes: The sample is as in Table 3.1. Distance, measured in miles, is the driving distance from the student's home address during eleventh grade to the nearest ACT-test center. In the post-policy period, the distance is the distance from a student's home to his or her high school. If a student has multiple addresses during eleventh grade, then the smallest distance is used.

Table 3.A.2. The Relationship Between ACT Scores and Student Demographics

			Pre-Policy, by Correction Method				
	Post-Policy	Pre-Policy	Tobit	Heckman			
	OLS	OLS	(10th Pctl)	No IV	With IV	Olsen	Newey
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Free Lunch	-2.551***	-1.841***	-2.060***	-6.492***	0.003	0.602	-0.691***
	(0.086)	(0.106)	(0.123)	(1.659)	(0.509)	(0.613)	(0.204)
Female	0.252***	-0.130***	-0.139***	1.393***	-0.729***	-0.969***	-0.653***
	(0.032)	(0.035)	(0.038)	(0.507)	(0.156)	(0.203)	(0.099)
Black	-3.444***	-4.102***	-4.736***	-4.116***	-4.091***	-4.096***	-3.922***
	(0.155)	(0.207)	(0.277)	(0.484)	(0.144)	(0.148)	(0.203)
Hispanic	-2.113***	-1.822***	-1.969***	-3.403***	-1.188***	-0.985***	-1.581***
	(0.097)	(0.200)	(0.223)	(0.669)	(0.310)	(0.366)	(0.208)
Other	0.997***	0.616**	0.621**	2.281***	-0.047	-0.339	0.254
	(0.330)	(0.321)	(0.330)	(0.851)	(0.302)	(0.309)	(0.287)
Sample Size	98,417	62,185	62,185	62,185	62,185	62,185	62,185

Notes: The sample is as in Table 3.1. The level of observation is the student. Each column is from a separate regression of ACT scores on the reported student-level demographics. Standard errors calculated using 200 bootstrap replications resampling schools. *** indicates statistical significance at the 0.01 level, ** at the 0.05 level, and * at the 0.01 level.

Table 3.A.3. The Relationship Between ACT Scores and Student and School Characteristics

			Pre-Policy, by Correction Method				
	Post-Policy	Pre-Policy	Tobit	Heckman			
	OLS	OLS	(10th Pctl)	No IV	With IV	Olsen	Newey
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student-Level							
Free Lunch	-1.703***	-1.078***	-1.232***	-2.539***	-2.445***	-1.713***	-1.074***
	(0.057)	(0.073)	(0.082)	(0.454)	(0.375)	(0.447)	(0.073)
Female	0.253***	-0.058*	-0.061*	0.561***	0.521***	0.233	-0.062*
	(0.030)	(0.035)	(0.038)	(0.179)	(0.147)	(0.203)	(0.036)
Black	-2.870***	-3.370***	-3.856***	-3.257***	-3.265***	-3.305***	-3.370***
	(0.107)	(0.124)	(0.139)	(0.157)	(0.152)	(0.146)	(0.124)
Hispanic	-1.783***	-1.568***	-1.682***	-2.056***	-2.028***	-1.779***	-1.570***
	(0.094)	(0.131)	(0.146)	(0.220)	(0.198)	(0.192)	(0.131)
Other	0.506***	0.157	0.136	0.569**	0.542**	0.367	0.152
	(0.198)	(0.210)	(0.216)	(0.263)	(0.247)	(0.250)	(0.211)
School-Level							
Pupil Teacher Ratio	0.012*	-0.002	-0.006	-0.006	-0.005	-0.004	-0.003
	(0.008)	(0.005)	(0.007)	(0.008)	(0.008)	(0.007)	(0.006)
Fraction Free Lunch	0.888**	-0.581**	-0.945***	-0.422	-0.431	-0.516*	-0.594**
	(0.377)	(0.283)	(0.336)	(0.415)	(0.391)	(0.315)	(0.284)
Fraction Black	1.808***	1.024	1.042	1.459	1.427	1.221*	1.019
	(0.396)	(0.775)	(0.977)	(0.933)	(0.888)	(0.776)	(0.779)
Number of Eleventh Graders	-0.000	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average Eighth Grade Score	1.949***	2.337***	2.520***	3.507***	3.437***	2.907***	2.327***
	(0.183)	(0.222)	(0.237)	(0.436)	(0.377)	(0.430)	(0.218)
Average Eleventh Grade Score	2.592***	1.224***	1.320***	2.200***	2.138***	1.683***	1.214***
	(0.170)	(0.194)	(0.204)	(0.381)	(0.347)	(0.401)	(0.196)
Sample Size	98,417	62,185	62,185	62,185	62,185	62,185	62,185

Notes: The sample is as in Table 3.1. The level of observation is the student. Each column is from a separate regression of ACT scores on the reported student- and school-level covariates. Missing value indicators and district-level versions of school covariates also included but coefficients not reported. Standard errors calculated using 200 bootstrap replications resampling schools. *** indicates statistical significance at the 0.01 level, ** at the 0.05 level, and * at the 0.01 level.

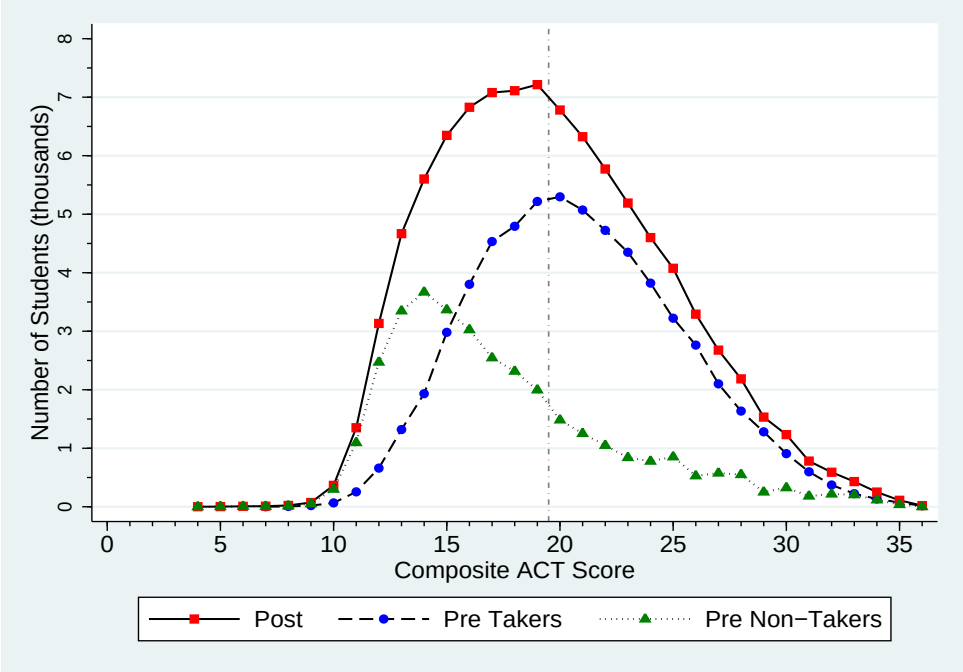
Table 3.A.4. The Relationship Between ACT Scores, Demographics, and Achievement

			Pre-Policy, by Correction Method				
	Post-Policy	Pre-Policy	Tobit	Heckman			
	OLS	OLS	(10th Pctl)	No IV	With IV	Olsen	Newey
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student-Level							
Free Lunch	-0.369***	-0.254***	-0.323***	-0.781***	-0.769***	-0.258	-0.279***
	(0.021)	(0.043)	(0.048)	(0.067)	(0.066)	(0.205)	(0.044)
Female	0.490***	0.027	0.051*	0.298***	0.292***	0.030	0.015
	(0.019)	(0.027)	(0.028)	(0.036)	(0.036)	(0.119)	(0.027)
Black	-0.648***	-1.295***	-1.566***	-0.904***	-0.914***	-1.291***	-1.358***
	(0.047)	(0.083)	(0.091)	(0.094)	(0.093)	(0.193)	(0.086)
Hispanic	-0.607***	-0.728***	-0.791***	-0.790***	-0.791***	-0.729***	-0.799***
	(0.058)	(0.102)	(0.109)	(0.112)	(0.112)	(0.103)	(0.107)
Other	0.383***	0.209*	0.170	0.402***	0.397***	0.211	0.129
	(0.092)	(0.124)	(0.123)	(0.044)	(0.109)	(0.160)	(0.129)
Eighth Grade Score	1.592***	1.833***	1.974***	2.206***	2.197***	1.837***	1.793***
	(0.029)	(0.034)	(0.037)	(0.057)	(0.043)	(0.170)	(0.033)
Eleventh Grade Score	3.036***	2.616***	2.810***	3.150***	3.138***	2.621***	2.537***
	(0.021)	(0.032)	(0.033)	(0.109)	(0.056)	(0.228)	(0.037)
School-Level							
Pupil Teacher Ratio	0.005	-0.003	-0.006	-0.004	-0.004	-0.003	-0.005
	(0.008)	(0.004)	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)
Fraction Free Lunch	-0.150	-0.448	-0.720**	-0.527*	-0.524*	-0.448	-0.556*
	(0.374)	(0.319)	(0.367)	(0.390)	(0.388)	(0.327)	(0.325)
Fraction Black	-0.008	-0.267	-0.307	-0.383	-0.383	-0.268	-0.099
	(0.396)	(0.558)	(0.718)	(0.645)	(0.641)	(0.554)	(0.522)
Number of Eleventh Graders	0.000	0.001**	0.001**	0.001**	0.001**	0.001**	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average Eighth Grade Score	0.943***	1.085***	1.126***	1.492***	1.486***	1.089***	0.992***
	(0.169)	(0.166)	(0.177)	(0.190)	(0.190)	(0.267)	(0.159)
Average Eleventh Grade Score	-0.336**	-0.206	-0.208	-0.160	-0.162	-0.205	-0.244*
	(0.165)	(0.154)	(0.164)	(0.177)	(0.176)	(0.157)	(0.147)
Sample Size	98,417	62,185	62,185	62,185	62,185	62,185	62,185

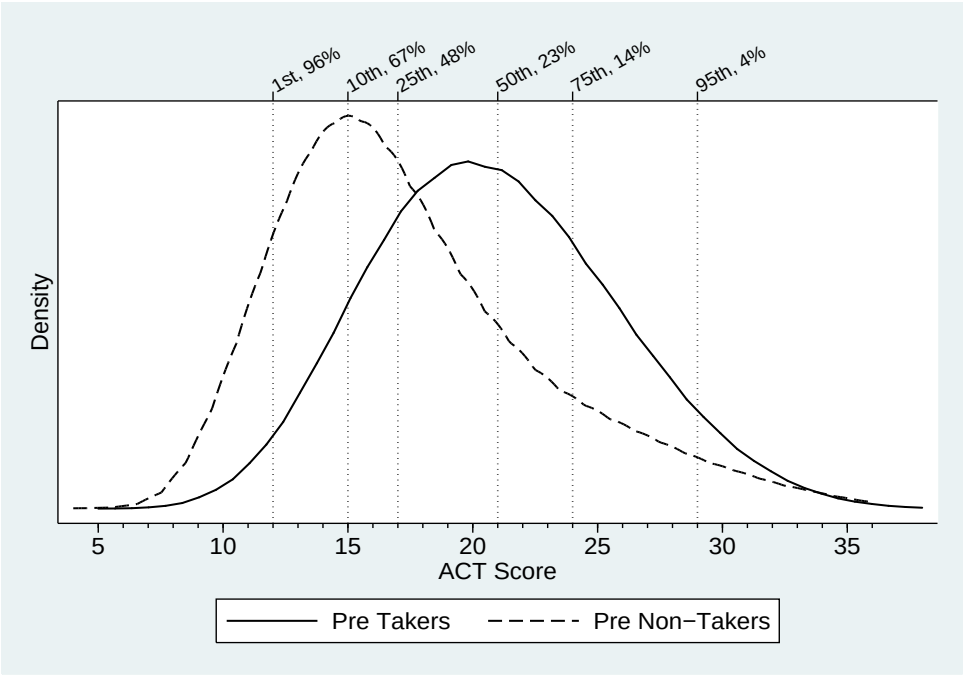
Notes: The sample is as in Table 3.1. The level of observation is the student. Each column is from a separate regression of ACT scores on the reported student- and school-level covariates. Missing value indicators and district-level versions of school covariates also included but coefficients not reported. Standard errors calculated using 200 bootstrap replications resampling schools. *** indicates statistical significance at the 0.01 level, ** at the 0.05 level, and * at the 0.01 level.

Figure 3. 1: Observed and Latent ACT ACT Scores Pre- and Post-Mandatory ACT

(a) Calculating Latent Scores Pre-Policy (Frequency Distributions)



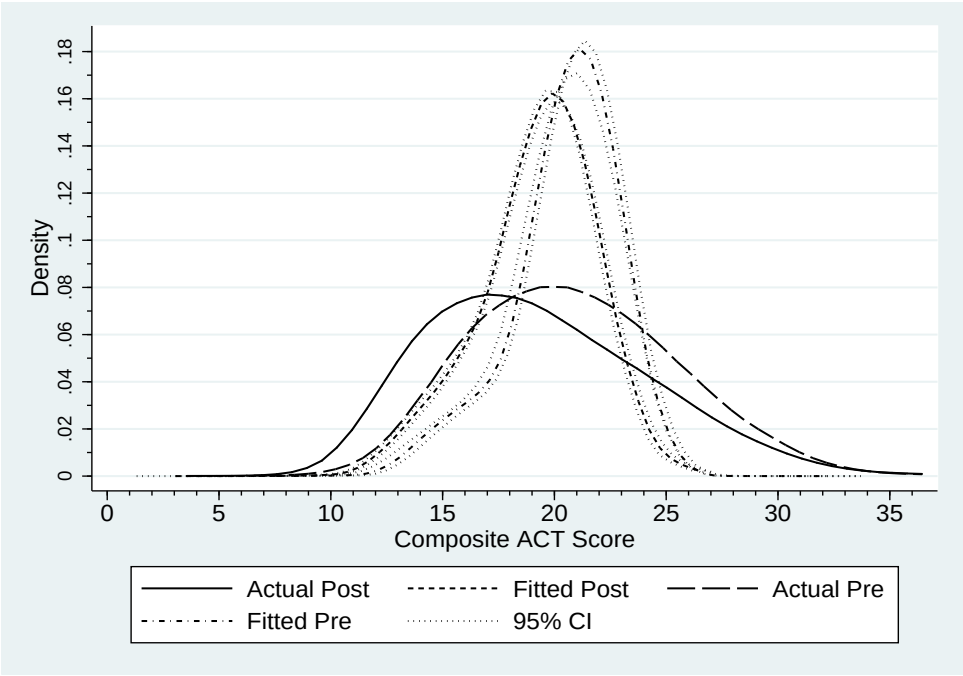
(b) Pre-Policy Observed and Latent Score Densities



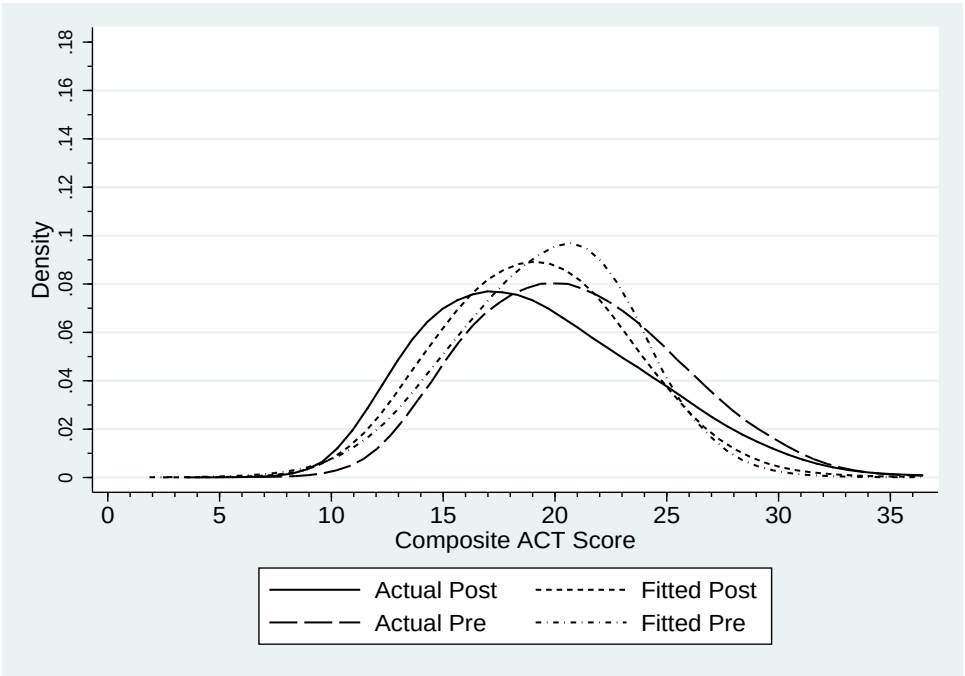
Notes: Figure (a) shows: 1) the distribution of ACT scores pre-policy, 2) the distribution post-policy reweighted following DiNardo, Fortin, and Lemieux (1996) to resemble th pre-policy cohort, and 3) the difference between (1) and (2), which is the latent score distribution among non-takers in the pre-period. Figure (b) plots kernel densities of (1) and (3). Along the top of the figure are percentiles of (1) followed by the fraction of (3) that has a latent score higher than that value.

Figure 3. 2: Observed and Predicted ACT Scores Pre- and Post-Policy

(a) Predicting ACT Scores Using Basic Student and School Characteristics

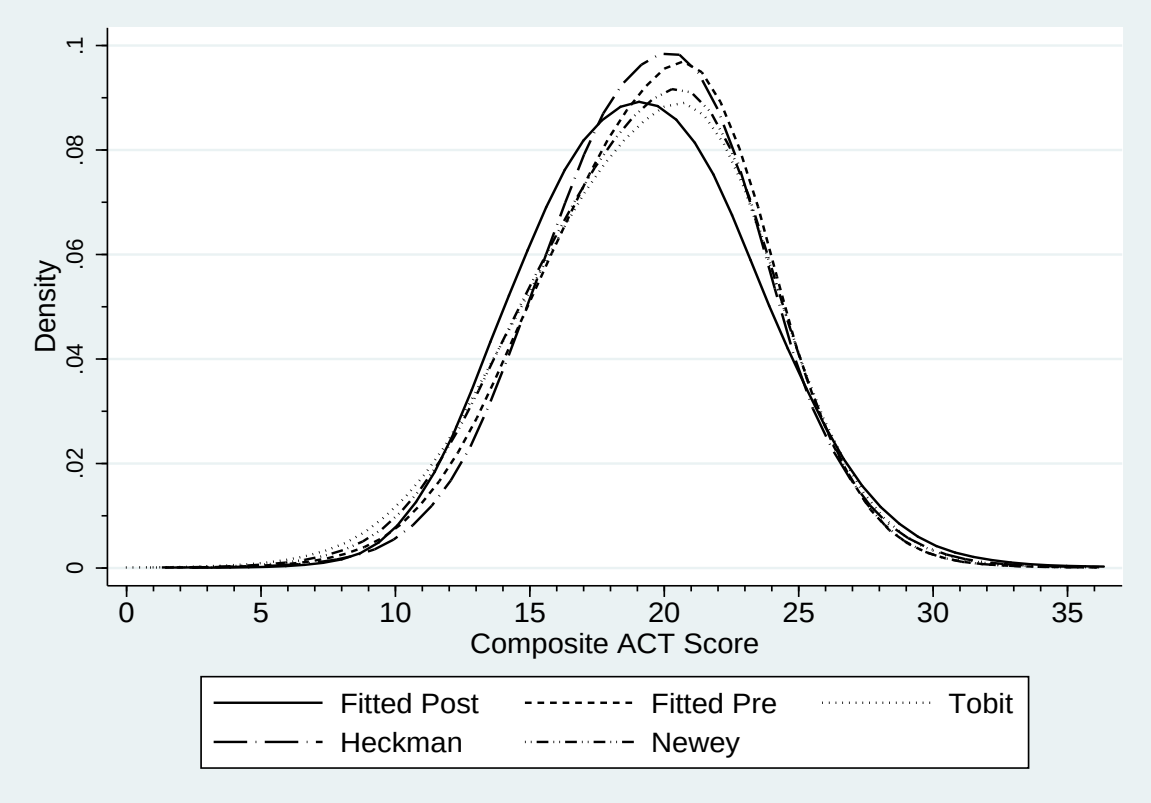


(b) Predicting ACT Scores Using Grade 8 and 11 Student Test Scores



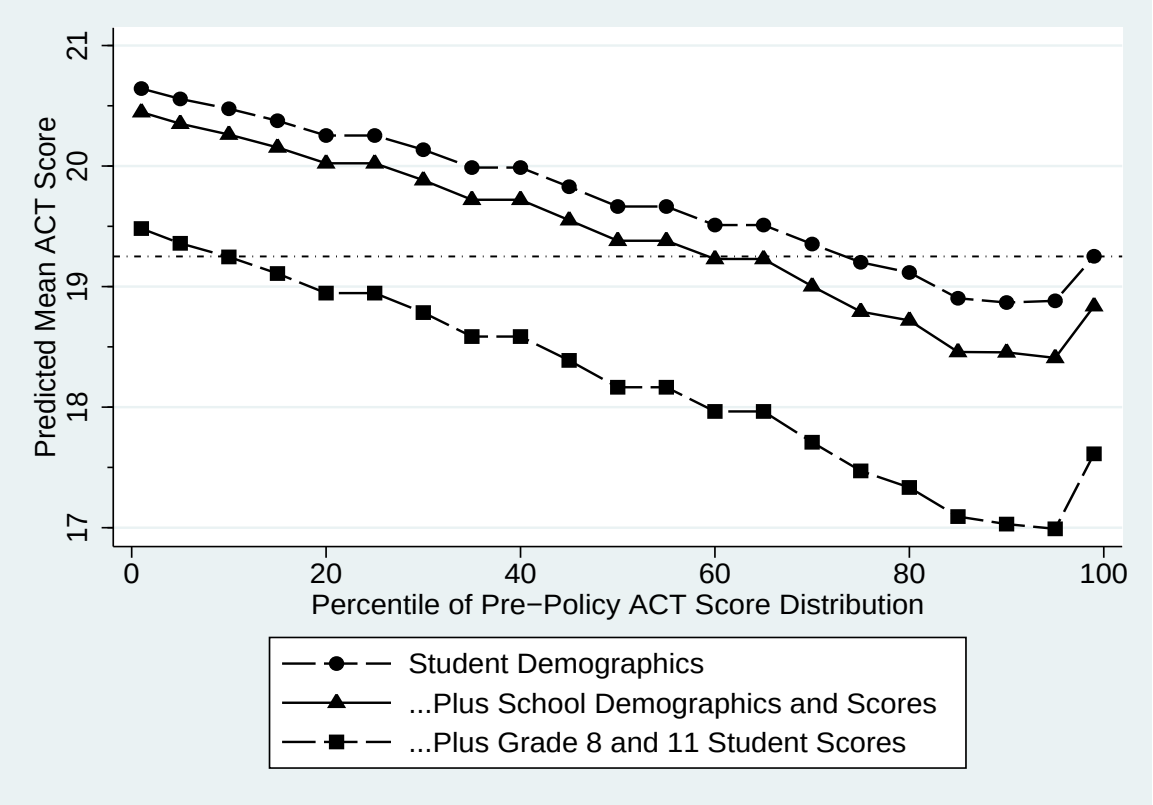
Notes: Figure (a) shows pre- and post-policy raw ACT scores and fitted values from regressions of ACT scores on student-level demographics and school-level demographics and test scores. The pre-policy fitted values are predicted out of sample to all students. Figure (b) shows the same picture but adds student-level eighth and eleventh grade test scores to the prediction equations. 95% confidence intervals are omitted from (b) for readibility.

Figure 3. 3: Comparing the Performance of Sample Selection Corrections



Notes: Figure shows pre- and post-policy fitted values from regressions of ACT scores on student-, school-, and district-level demographics, and eighth and eleventh grade test scores. The pre-policy fitted values are predicted out of sample to all students. Tobit, Heckman, and Newey are several selection corrections estimated using the pre-policy sample. 95% confidence intervals, which are quite tight, are omitted for readibility.

Figure 3. 4: Predicted ACT Score Mean From Tobit, by Censoring Point and Covariate Set



Notes: Figure shows predicted mean ACT scores estimated from Tobit regressions of ACT scores on covariates using pre-policy data. The data is censored so that ACT-takers scoring below a certain percentile and non-takers are assigned the score at the censoring point. The x-axis is the percentile at which the data is censored. The horizontal line gives the true post-policy mean score.