REPORT

to

DAVIS JOINT UNIFIED SCHOOL DISTRICT

How Does the AIM Program Affect Student Outcomes in the Davis Joint Unified School District?

May 2015

Scott Carrell, Ph.D.
Michal Kurlaender, Ed.D.
Marianne Page, Ph.D.
K. A. Kramer, M.A.
School of Education & Department of Economics
University of California, Davis

1. Introduction

In spring 2014, leaders in the DJUSD asked us to investigate the impact of special programs in the District, in particular the largest special program: AIM (formerly known as GATE).[1] The goal of this report is to summarize our efforts in doing so. Specifically, our analysis explores the following three questions:

(1) Who is GATE/AIM identified in the Davis Joint Unified School District and how has this changed over time?
(2) What is the *causal* impact of the AIM program on student academic achievement for students in the AIM program?
(3) What are the *causal* effects of AIM on students not in AIM?

This report is organized as follows: in Section 2 we provide some background on the difficulty of program evaluations of this type as well as review the evidence from similar studies looking at gifted and talented education programs. Section 3 describes the data we employ. Section 4 presents our methodological approach and results for each of these research questions, and Section 5 discusses conclusions from our analysis.

2. Background

Nearly all program evaluations face a similar challenge: it is very difficult to distinguish between true program effects and differences in outcomes that would naturally exist between program participants and non-participants. This occurs because program participants usually have specific characteristics that make them eligible for, or that cause them to select into, the program. For example, with respect to gifted and talented programs, we expect that even in the absence of the program those who currently participate would be likely to have higher test scores than those who do not participate. In fact, it is often differences in test scores that distinguish eligible from ineligible students in the first place. A simple comparison of test score differences between students who participate in the program and non-participants would not, therefore, provide meaningful information about the program's effects.

Some researchers begin to address this issue by controlling for demographic factors and individual student test scores from tests taken before the student enters the program. Unfortunately, this is not sufficient for isolating a "causal" program effect as students with similar demographics and test scores but differences in participation status are still likely to differ on other dimensions that influence outcomes. Examples of differences that might influence both program participation and student outcomes include the student's own level of motivation, parental support, or parent's access to information about the program.

[1] Given the more recent change in name, we will use AIM and GATE interchangeably throughout this report.

Under ideal research conditions, researchers would be able to overcome these challenges by running experiments in which similar students were randomly assigned either to a GATE classroom (the treatment group) or a regular classroom (the control group). The impact of the GATE program could then be estimated by comparing outcomes across students in the two groups, without risk of contamination by other factors that are correlated with both program participation and outcomes. In practice, experiments are expensive and rare. As a result, program evaluators must rely on alternative quasi-experimental methods to try to isolate a causal program effect. Despite a vast literature on gifted and talented education programs, our review of the literature focuses on studies that use methodological approaches that are identified by the U.S. Department of Education, Institute of Education Science's *What Works Clearinghouse*[2] standards; that is, studies that use real world settings to come as close as possible to mimicking an ideal experiment. There are very few studies that convincingly approach this ideal, but among those that do, the evidence that existing GATE programs positively (or negatively) affect those who participate is sparse.

Bui, Imberman, and Craig (2014) use a regression discontinuity approach (described in the next section) to compare seventh grade achievement test scores across students who barely qualify for a GATE program implemented in a school district in the Southwest vs. narrowly miss qualifying for the program. Seventh grade achievement test scores are recorded after a year and a half of participating in the GATE program. The authors find no effects of GATE on achievement scores in math, reading, language, science, and social studies. In the same study, the authors use an alternative estimation strategy that is based on lottery admissions to two oversubscribed GATE magnet programs in another school district. Again, they find no effects on achievement scores in math, reading, language, and social studies, but they do find some evidence that science scores increase for the eligible students who are randomly admitted to the program.

Card and Giuliano (2014) also use a regression discontinuity design to examine the effects of a gifted and talented program implemented in a large, anonymous school district. The district under study placed three distinct groups of students in self-contained, gifted classrooms. The first group consisted of non-disadvantaged students with very high IQ scores (above 130). The second group consisted of English language learners and free-and-reduced price lunch participants with IQ's over 116 points. The remaining seats in the gifted classrooms were filled by high achieving students, whose IQ scores did not meet the thresholds described above. Specifically, high achieving students were identified as those who scored highest among their school/grade cohort on statewide achievement tests in the previous year.

The authors find no evidence that GATE program participation improved standardized achievement scores in math or reading among the first two groups, although there were small positive effects on the second group's writing scores. In contrast, the authors do find evidence of positive effects on reading and math for the third group, who entered the GATE program due to high achievement, rather than high IQ. The effects are most pronounced among lower income and minority students. They conclude that a separate

[2] http://ies.ed.gov/ncee/wwc/default.aspx

classroom environment is more effective for students selected on past achievement, particularly disadvantaged students who are often excluded from gifted and talented programs.

Before we turn to our analysis it is worth noting why we might expect to find a positive effect of GATE on participating students. First, gifted students often have unique learning needs that may be better addressed by teachers and a classroom environment that is set up for those needs (i.e. differentiated instruction opportunities). Second, the quality of the peers in a classroom of high achievers would also likely raise the achievement for a participating child—a finding that is corroborated in the peer effects literature (Hanushek, Kain, Markman, & Rivken, 2003; Angrist & Lang, 2004; Carrell & Hoekstra, 2010). Finally, there are a host of other dimensions (often unobserved by the researcher) that may result in improved outcomes, such as higher motivation or parent participation in the child's schooling environment. In reality all of these forces are often inter-related.

3. Data

In order to examine the academic achievement effects of the AIM/GATE program, we employ individual-level student data provided to us by the Davis Joint Unified School District. The data were de-identified by personnel in the DJUSD district office to protect the identity of each student in the dataset. The dataset encompasses all students who were enrolled in DJUSD elementary schools in the fourth grade from academic years 2006-07 through 2012-13.[3] The information provided to us on each student includes the following: 1) the scale scores from any second, third, fourth, fifth, or sixth grade math or ELA California State Test (CST) the student may have taken; 2) Otis-Lennon School Ability Test (OLSAT) scores from universal AIM/GATE testing in third grade; 3) demographic data including age, race, gender, free and reduced lunch eligibility, English learner status, the primary language spoken in the home, and the highest level of parental education; 4) AIM/GATE qualification and placement information; and 5) school enrollment and attendance records.

4. Methods & Results

a. AIM Qualification and Program Participation

We begin with a descriptive analysis of the AIM program by examining which students qualify for the program and how. Table 1 shows that for the fourth grade cohorts from 2006-2013, 31% of students in DJUSD qualify to participate in AIM while 19% of 4th graders are placed into self-contained AIM classrooms. Figures 1 and 2, along with Table 1, show that there have been no major variations in the size of the AIM program over time. In the last seven years, the number of 4th grade students who qualify for AIM has fluctuated around 200, which is equivalent to 30-35% of 4th graders. The number of students in 4th grade AIM self-contained classrooms has remained stable at about 120 a year, which is approximately 20% of 4th grade students.

[3] We do not use prior years because the program was under a different master plan, and we do not use the most recent years because students did not take the California State Test (CST).

While the size of the AIM program has been stable over time, the way students qualify for the AIM program has changed. Table 2 shows that the number of students who qualify through universal testing has decreased, while the number of students who qualify through retesting and private testing has increased. These changes correspond to the different AIM (GATE) Master plans in place during the time the fourth grade students qualified for AIM. Students in 4^{th} grade in the 2007 – 2009 academic years qualified for AIM (GATE) under the 2005 Master Plan while the students in 4^{th} grade in the 2010 – 2013 academic years qualified for AIM under the 2008 Master Plan. Of the students who qualified under the 2005 Master Plan, 41% qualified though universal testing (primarily by taking the OLSAT), 30% qualified through retesting by DJUSD (primarily by taking the TONI), 23% qualified through private testing, and 7% qualified by other means. Of the students who qualified under the 2008 Master Plan, 29% qualified through universal testing (primarily by taking the OLSAT), 40% qualified through retesting by DJUSD (primarily by taking the TONI), 30% qualified though private testing, and 1% qualified through other means.

Figure 4 provides a more detailed examination of the AIM qualifiers by qualification type and OLSAT score. Each bar of a histogram represents the number of students who qualify for the AIM program by their corresponding OLSAT score. The first chart shows the OLSAT scores of all AIM qualified students, regardless of their qualification type. The next three charts show OLSAT scores broken out by qualification type. The chart titled "Qualified Though Universal Testing" shows the OLSAT scores of students who qualified via universal testing, the chart titled "Qualified Through Retesting" show the scores of students who qualified through retesting, and the chart titled "Qualified Through Private Testing" shows the scores of students who qualified through private testing. As expected, students who qualify through universal testing tend to have very high OLSAT scores with a mean at the 96^{th} percentile. The vast majority of these students score at or above the published test score threshold (~96) required for qualification. On the flipside, students who qualify through retesting or private testing score throughout the entire OLSAT distribution with scores ranging from the 3^{rd} percentile to the 99^{th} percentile. On average, students who qualify through retesting scored at the 75^{th} percentile while students who qualify through private testing score at the 77^{th} percentile of the OLSAT.

Finally, Figure 5 shows the yearly distributions of the OLSAT scores of AIM qualifiers. These figures show a striking pattern, with an increase from 2007 through 2013 in the number of students who qualify for the AIM program from the lower parts of the OLSAT distribution. As shown in Table 4, in 2007, the student with the lowest OLSAT score who qualified for AIM scored at the 27^{th} percentile and the mean OLSAT score of all AIM qualifiers was 87. By 2013, that average OLSAT score of AIM qualifiers had decreased to 78 while the AIM qualified student with the lowest OLSAT score scored in the 4^{th} percentile.

b. Academic Achievement Effects of AIM on Students in AIM

Comparing the average outcomes of students in AIM self-contained classrooms with students not in AIM self-contained classrooms does not provide a good estimate of the causal effect of AIM. To understand why, we need to think about why some students are in AIM self contained classrooms while others are not. All students in the AIM self-contained classrooms did well enough on a test to qualify for AIM. Although many factors go into how well a person scores on a test, in general, if someone scores well on one test of cognitive abilities or academic achievement we might then expect them to also score well on another test of cognitive abilities or academic achievement. Hence, if a student scored well enough on a test to be in an AIM classroom, is reasonable to anticipate that her score will be better than average on an achievement test like the CST relative to students whose test scores did not make them eligible. Therefore, even if the AIM program had no positive effects on achievement, we would still expect the students in AIM self-contained classrooms to have higher CST scores, on average, compared to students not in AIM.

In many areas of social science, researchers use randomized control trials to disentangle the actual effect of a program from the effect of different types of individuals choosing to participate or not participate in a program. In the context of AIM, one could construct a randomized control trial by randomly assigning students to the AIM program. Outcomes of students assigned to the AIM program and students assigned to traditional classrooms could then be compared. Doing so would allow one to estimate the true effect of the AIM program. To our knowledge, no researcher has yet implemented a randomized experiment to test the effect of a gifted and talented program.[4] However, because of how students qualify for AIM in the DJUSD, certain parts of the qualification process are effectively random. It is this randomization that we exploit to estimate the effect of the AIM program on students in the program. This approach is called a regression discontinuity research design (RDD). This common approach to program evaluation is similar to that used to examine the impact of Head Start (Ludwig and Miller, 2007), and more recently the effects of Gifted and Talented Education (GATE) programs (Bui et al., 2014; Card and Giuliano, 2014).

In order to implement an RDD, a researcher must identify a score threshold where, for individuals scoring close to the threshold, making or missing the threshold is essentially random. One main pathway into qualifying for the AIM program is for a student to obtain a 96th percentile on their OLSAT total score and a 96th percentile on either their verbal or non-verbal score. As such, a 96th percentile on the OLSAT total score is such a threshold. While we would not expect students to score at the top of the distribution due to random chance, we would expect that whether or not a student scored in the 95th percentile (a non-qualifying score) vs. the 96th percentile (a qualifying score) to be largely random.

Another main pathway into qualifying for the AIM program for students is to score within +/- 5 standard errors of measurement of a 96 on their total score. This qualifies the student for retesting (and is equivalent to scoring at about the 90th percentile depending on the year). Again, we would expect whether a student scored within +/- 6 standard errors of measurement (a score that would not qualify the student for retesting) versus +/- 5 standard errors of measurement (a score that would qualify the student for retesting) to be largely random.

[4] There are many reasons that school districts might want to avoid randomized control trials.

Because we expect whether a student scores on either side of the threshold(s) to be effectively random, we also would not expect systematic differences between those same students on either side of the cut offs, except in their likelihood of being eligible for the AIM program. Thus, through this quasi-experimental method, we can compare the students who just miss the threshold(s) with students who just make the threshold(s) to gauge the causal effect of AIM.

Figure 6 illustrates the expected results of a typical regression discontinuity design under three hypothetical alternative scenarios. In the top chart, we can see that a student's expected CST score rises as they score higher on the OLSAT. However, there is no jump at the qualification threshold to get into AIM. In this case, we would conclude that AIM had no effect on a student's CST scores. This contrasts with the second chart. Here again, expected CST scores rise with a student's OLSAT score, but once the qualification threshold is reached, CST scores of students just to the right of the qualification threshold jump by 15 points. In this case we could conclude that the AIM program had a positive effect on students around the threshold of 15 points on the CST. In the third chart, students just to the right of the threshold have CST scores 15 points lower than students to the left of the threshold. Here we would conclude that the AIM program has a negative effect on students around the threshold of 15 points on the CST.

Figure 7 shows the actual estimated effects of the AIM program on CST scores.[5] Students just to the right of the qualification threshold score, on average, 3 points higher on their ELA CST and 6 points lower on their math CST than students just to the left of the qualification threshold. These effects are small and statistically indistinguishable from zero. As such, there is no evidence that the AIM program has an effect on students in the AIM program.

c. Academic Achievement Effects of AIM on Students not in AIM

In order to estimate the effects of AIM on students not in AIM, we compared the outcomes across cohorts of students within a school in years when more vs. less of their peers scored above an AIM qualification threshold.[6] Within a school, the exact number of students who score above an AIM qualification threshold should be partially random in a given year. This quasi-experimental method gives us an accurate estimate of the causal effect of the AIM program on students not in the AIM program. We do this instead of comparing the average outcomes of students in classrooms where many students transferred to the AIM program in 4th grade to the average outcomes of students in classrooms where many students did not transfer to the AIM program in 4th grade. Such a comparison would be likely to produce misleading results. Students in these two types of classrooms are likely to be systematically different in ways that a researcher cannot easily observe but would affect their academic achievement.

Results for this analysis are shown in Table 5. Since, on average, 20% of a third grade cohort at AIM schools scores above the threshold, these estimates should be scaled by one-fifth. For math scores, the coefficient of -18.8 suggests that non-AIM students' test

[5] For more details about our methodology, please see the technical appendix.
[6] For more details about our methodology, please see the technical appendix.

scores fall, on average, approximately 3.9 points in 4th-6th grade after their peers leave for self-contained AIM classrooms. Likewise, the coefficient of -8.3 for ELA scores suggests that non-AIM student's test scores fall, on average, approximately 1.7 points in 4th-6th grade. These effects are small and are not statistically different from zero. As such, there is no evidence that the AIM program has an effect on students not in the AIM program.

5. Conclusions

Our analyses lead to several conclusions about DJUSD's AIM program. Identification into the program has been both inconsistent and variable over time. The number of students who have qualified for AIM through universal testing has decreased while the number of students who have qualified through private testing and retesting has increased. At the same time, there has been a drop in the average and minimum OLSAT scores of the students who qualify for AIM. These changes have resulted in a lack of transparency about who is truly "fit" for the program. The wide distribution of OLSAT scores among AIM participants also calls into question whether the AIM program is primarily serving "gifted" students.

The data also indicate that, in addition to the District's efforts to retest students across different risk factors, many students currently enter the program through private testing. There is a wide range of OLSAT scores among both retested and privately tested students, which suggests that if some students who are not retested or privately tested were to receive those test alternatives they might also be eligible to enroll in AIM. Given the distribution of OLSAT scores among participating AIM students, the number of potentially eligible students who are missed because they do not receive private testing or district retesting may be quite large.

We also find no evidence that the program positively affects achievement scores of participating students and no evidence that the program negatively affects non-participating students. This "no benefit, no harm" finding should be considered within the context of the program's cost: the DJUSD spends considerable resources on universal testing and retesting (as do private citizens for additional testing) and given the financial costs and capacity constraints associated with this program, we should expect these costs to be balanced by some measurable benefit.

These analyses do not allow us to make explicit recommendations about what an ideal gifted and talented program in the DJUSD should look like. We can conclude, however, that clearer identification of "gifted" children will increase program transparency. Additionally, the current program, which represents a very large swath of Davis students, conflates giftedness with a host of other characteristics, including high achievement and parents' access to alternative routes of program entry. This likely affects the type of specialized services that can be offered.

6. References

Angrist, Joshua D., and Kevin Lang, 2004. "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program," American Economic Review, American Economic Association, vol. 94(5), pages 1613-1634, December.

Bui, Sa A., Steven G. Craig, and Scott A. Imberman. 2014. "Is Gifted Education a Bright Idea? Assessing the Impacts of Gifted and Talented Programs on Students." *American Economic Journal: Economic Policy* 6(3): 30-62.

Burke, Mary A. and Tim R. Sass. 2004. "Classroom Peer Effects and Student Achievement." In the *American Economic Association Annual Meetings*.

Card, Davis and Laura Giuliano. 2014. "Does Gifted Education Work? For Which Students?" NBER Working Paper 20453.

Carrell, Scott E and Mark L. Hoekstra, 2010. "Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone's Kids,"*American Economic Journal: Applied Economics*, American Economic Association, vol. 2(1), pages 211-28, January.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633-79.

Hanushek, Eric A., John F. Kain, Markman, Jacob M., and Steven G. Rivkin, 2003. "Does peer ability affect student achievement?," *Journal of Applied Econometrics*, John Wiley & Sons, Ltd., vol. 18(5), pages 527-544.

Hoxby, Caroline. 2000. "Peer effects in the classroom: Learning from gender and race variation." NBER working paper 7867.

Hoxby, Caroline M., and Gretchen Weingarth. 2006. "Taking race out of the equation: School reassignment and the structure of peer effects." Working paper, Harvard University

Lefgren, Lars. 2004. "Educational peer effects and the Chicago public schools." *Journal of Urban Economics* (2) : 169-191.

Ludwig, Jens and Douglas L. Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression-Discontinuity Design." *Quarterly Journal of Economics* 122(1): 159-208.

Vigdor, Jacob, and Thomas Nechyba. 2007. "Peer effects in North Carolina public schools." In *Schools and the equal opportunity problem*, ed. Ludger Woessmann and Paul E. Peterson, 73–102. Cambridge, MA: MIT Press.

Technical Appendix

Academic Achievement Effects of AIM on Students in AIM

To estimate the causal impact of the AIM program on student academic achievement for students in the AIM program, we implement a regression discontinuity research design (RDD). The RDD exploits the increase in the probability of participating in the AIM program at the OLSAT score cutoffs used for determining program eligibility (e.g., 96 vs. 95). This approach is similar to that used to examine the impact of Head Start (Ludwig and Miller, 2007), and more recently on the effects of Gifted and Talented Education (GATE) programs (Bui et al., 2014; Card and Giuliano, 2014).

The identifying assumption of the research design is that while other determinants of academic achievement such as motivation, ability, and parental support vary smoothly over OLSAT score thresholds, participation in the AIM program varies discontinuously. This smoothness assumption relies on the fact that third graders have little scope to change their OLSAT test scores during universal testing. In particular, no individual student has the ability to manipulate or change their score from just below to just above the qualifying cutoff score used for program eligibility.

To formally test for discontinuities in academic achievement at qualifying OLSAT score thresholds, we estimate the following equation:

$$CST\_Score_{igst} = \beta_0 + \beta_1 OLSAT\_Cutoff_i + \beta_2 OLSAT\_Score_i + l_{gst} + \varepsilon_{igst} \qquad (1)$$

where *CST_Score* is the California State Test (CST) score earned by student *i*, in grade *g*, on subject *s*, in year *t*. *OLSAT_Cutoff* is an indicator variable equal to one if the student scored above the predicted cutoff on the OLSAT test administered during universal testing during third grade. *OLSAT_Score* is the student *i*'s own OLSAT score. l_{gst} *is a grade by subject by year fixed effect.* ε_{igst} is the error term. Under the identifying assumption that other determinants of achievement are continuous at the OLSAT cutoff, β_1 will be an unbiased estimate of the effect of participation in AIM on academic achievement. Standard errors are clustered at the OLSAT score level.

To estimate the OLSAT cutoff scores, in each year we predict the probability of fourth grade students being enrolled in an AIM self-contained classroom as a function of an indicator variable at or above a cutoff score. We estimated separate cutoff scores ranging from 83 through 98 and use the cutoff score that statistically has the best fit to the data in each year. These "best fit" cutoff scores by year were found to be: 95 in 2006, 96 in 2007, 90 in 2008, 89 in 2009, 93 in 2010, 90 in 2011, 91 in 2012 with the years referring to the academic years in which the students took the OLSAT. In practice, these cutoff scores vary by year and generally decrease over time, which is likely due to expansion of the retesting program to students in the lower part of the OLSAT distribution. Said differently, over time, the probability of qualifying for the program discontinuously increased at lower OLSAT score thresholds in the later years (2008-2012) compared to earlier years (2006-2007).[7]

We then re-centered our data around the best fit OLSAT cutoff scores and estimated the probability of being in an AIM self-contained program as a function of each individual student's OLSAT score (i.e., the running variable) and the OLSAT cutoff score (i.e., the discontinuity). Results are presented graphically in Figure A1 and show that there is a large, positive, and statistically significant increase of 29 percentage points in the probability of a student being in a self-contained AIM classroom just above the (re-centered) OLSAT score threshold. Said differently, students just above the cutoff have a nearly 60% probability of being in an AIM self-contained classroom, while those just below the cutoff score have only a 30% probability of being in an AIM self-contained classroom.

We exploit this near 30-percentage point discontinuous jump in the probability of being an AIM self-contained classroom to estimate the causal impact of the AIM program on student achievement for students in the AIM program by estimating equation *(1)* using 4th, 5th and 6th grade test scores as our outcome variable of interest.[8] Again, our main identifying assumption is that the students just above the OLSAT cutoff are no different than students just below the cutoff.[9] Results for 4th, 5th and 6th grade CST scores[10] are shown graphically in Figure 7, with the corresponding point estimates presented in Table A1. Similar to the other recent causal research on gifted and talented programs, our results show no evidence of a positive effect on academic achievement for students in the AIM program (Bui et al., 2014; Card and Giuliano, 2014). The point estimate of -1.56 indicates, that, on average, students just above the AIM cutoff threshold score a statistically insignificant 1.56 points lower on their Math and ELA CST score compared students just below the threshold. Results shown separately for Math and ELA support this finding, with statistically insignificant point estimates of -6.36 for math and 3.23 for ELA.

In summary, results from this analysis show there are no statistically significant effects on academic achievement (positive or negative) of the AIM program on students in the DJUSD self-contained AIM classrooms.

---

[7] Note: For the RD analysis we were unable to you data from the 2013 cohort due to the fact that the ELA and math CSTs were largely not administered to students in California during the 2014 academic year.
[8] One potential concern with using CST scores is that the highest achieving students may be top-coded on their CST scores. That is, if a substantial fraction of high achieving students score perfect scores on the CST, our effects would under-estimate the true effect of the program. In examining the CST scores from 4th-6th grade, this is not a serious concern. For the students within 4 percentile points of the OLSAT cutoff, only 9.3% scored a perfect 600 on the Math CST and only 0.3% scored a perfect 600 on the ELA CST.
[9] To test this assumption we examined whether pre-AIM second grade CST scores were different for students across the OLSAT cutoff scores. Results show no statistically significant differences in scores.
[10] Recent studies in the education literature have shown a positive link to changes in primary school standardized test scores and long-term outcomes. For example, interventions like smaller class sizes and better teachers that raise test scores also increase the likelihood of college attendance, increase future earnings, and decrease teenage pregnancy (Chetty et al. 2014).

Academic Achievement Effects of AIM on Students not in AIM

Our approach to measuring the effects of AIM on students not in AIM relies on the exogenous change in peer group quality that occurs starting in fourth grade when AIM students "leave" their regular classrooms for the AIM program.[11] Similar to the RDD approach previously discussed, our identifying assumption relies on the fact that students who do not qualify for the AIM program have no scope to change their own OLSAT score or the OLSAT scores of their third grade peers who do qualify for the AIM program during universal testing.

Similar to recent studies in the peer effects literature, we exploit the natural variation in cohort composition across time within a given school.[12] Thus, our methodology relies on using the idiosyncratic changes in the fraction of peers who qualify for the AIM program during universal testing by scoring above the OLSAT qualification cutoff scores within a school, over time. Formally, we estimate the following equation separately for ELA and math scores using ordinary least squares:

$$CST\_Score_{iwgt} = g_0 + g_1 PeerAim_{wt} + g_2 X_{it} + l_{gw} + a_t + \varepsilon_{iwgt}$$

where CST_Score_{iwgt} is the California State Test (CST) score earned by student i, in school w, in grade g, on subject s, in year t. $PeerAim_{wt}$ is the fraction of peers in the school year cohort who score above the qualification threshold. X_{it} is a vector of individual i's specific (pre-treatment) characteristics, including own second and third grade CST test scores, OLSAT score, race, gender, and subsidized lunch status. l_{gw} and a_t are school by grade and year fixed effects. ε_{iwgt} is the error term. We cluster our standard errors at the school by cohort level. The primary parameter of interest is g_1, which measures how each non-AIM student's test scores change in fourth through sixth grade if all of their third grade cohort peers were to score above the qualification threshold in that year. Since, on average, 20% of a third grade cohort at AIM schools scores above the threshold, the estimated coefficients should be scaled by one-fifth.

Results for this analysis are shown in Table 5. The point estimates for both Math and ELA test score outcomes are relatively small, negative, and statistically insignificant. For math scores, the coefficient of -18.83 suggests that non-AIM student's test scores fall, on average, approximately 3.9 points in 4th-6th grade after their peers leave for self-contained AIM classrooms. Likewise, the coefficient of -8.29 for ELA scores suggests that non-AIM student's test scores fall, on average, approximately 1.7 points in 4th-6th grade. We emphasize, though, that these effects are not statistically different from zero.

In summary, results from this analysis show there are no statistically significant effects (positive or negative) of the AIM program on students not in the DJUSD self-contained AIM classrooms.

[11] For this portion of the analysis we limit our sample to those elementary schools in DJUSD that have self-contained AIM classrooms. These schools include: Korematsu, North Davis, Pioneer, Valley Oak (prior to closure), and Willett.
[12] See Hoxby 2000; Hoxby and Weingarth 2006; Vigdor and Nechyba 2007; Burke and Sass 2004; Hanushek et al. 2003; Lars Lefgren 2004; and Carrell and Hoekstra 2010.

Table 1: Size of the AIM (GATE) Program

AY	Students in Fourth Grade				
	Total	AIM Qualified		AIM Self-Contained	
	#	#	%	#	%
2007	690	211	30.6%	119	17.2%
2008	616	183	29.7%	118	19.2%
2009	604	197	32.6%	102	16.9%
2010	634	191	30.1%	122	19.2%
2011	631	203	32.2%	122	19.3%
2012	605	211	34.9%	118	19.5%
2013	667	196	29.4%	131	19.6%
Total	4447	1392	31.3%	832	18.7%

Table 2: AIM (GATE) Qualification by Qualification Type

AY	GATE Qualified Fourth Grade Students by Qualification Type								
	Total	Universal Testing		Retesting		Private Testing		Transfers / Unknown	
	#	#	%	#	%	#	%	#	%
2007	211	75	35.5%	68	32.2%	55	26.1%	13	6.2%
2008	183	82	44.8%	42	23.0%	37	20.2%	22	12.0%
2009	197	83	42.1%	65	33.0%	44	22.3%	5	2.5%
2010	191	70	36.6%	67	35.1%	49	25.7%	5	2.6%
2011	203	50	24.6%	86	42.4%	65	32.0%	2	1.0%
2012	211	59	28.0%	80	37.9%	72	34.1%	0	0.0%
2013	196	56	28.6%	84	42.9%	53	27.0%	3	1.5%
Total	1392	475	34.1%	492	35.3%	375	26.9%	50	3.6%

Table 3:
OLSAT Scores by Qualification Type

Qualification Type:	Mean OLSAT Age Percentile Rank Total Score
Universal Testing	96
Retesting	75
Private Testing	77
All	83

Table 4: OLSAT Scores for 4th Grade AIM Qualified Students by Academic Year

AY	Age Percentile Rank Total Score		
	Mean	Min	Max
2007	87	27	99
2008	90	25	99
2009	87	19	99
2010	82	21	99
2011	78	3	99
2012	78	3	99
2013	78	4	99

Table 5: Effect of AIM on Students Who Do Not Qualify for AIM

VARIABLES	(36) CST: Math	(48) CST: ELA
Percent of other students scoring above the threshold	-18.83	-8.288
	(33.20)	(14.09)
Observations	2,329	2,331
R-squared	0.564	0.659

Estimates produced using data from schools with GATE self contained classes: Korematsu, North Davis, Pioneer, Valley Oak, Willett

Percent of other students scoring above the threshold = Percent of other students scoring above the threshold in a student's school in 3rd grade

Standard errors clustered at the school by cohort level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Figure 1: Size of the AIM (GATE) Program

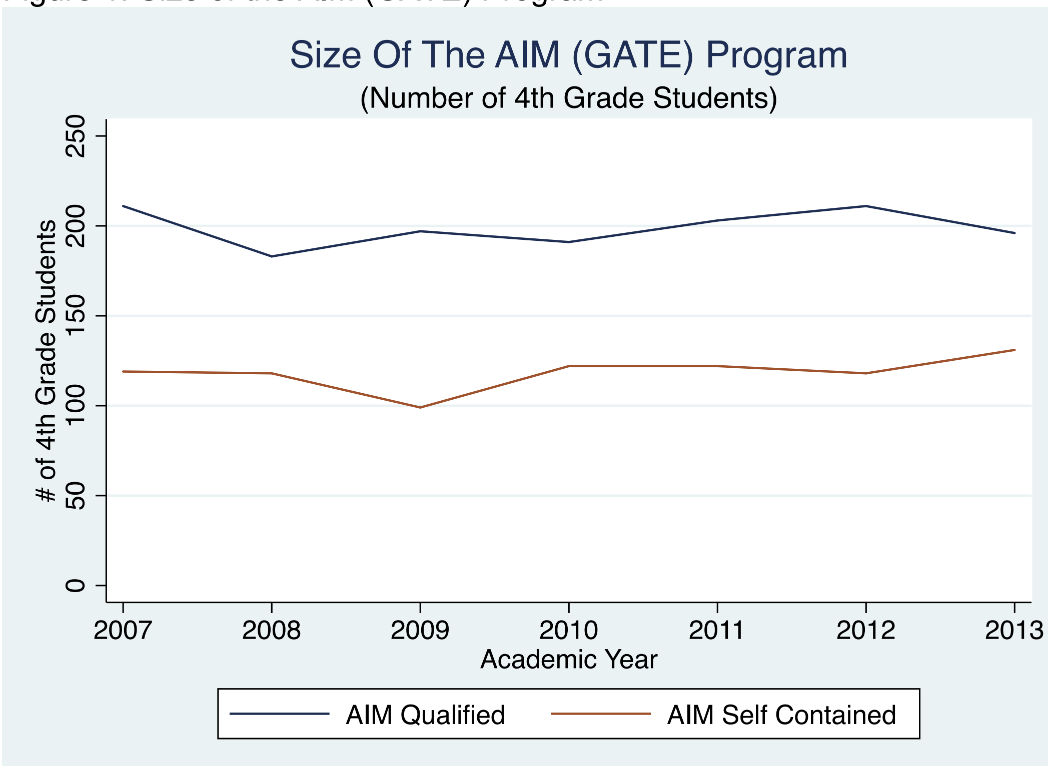


Figure 2: Size of the AIM (GATE) Program

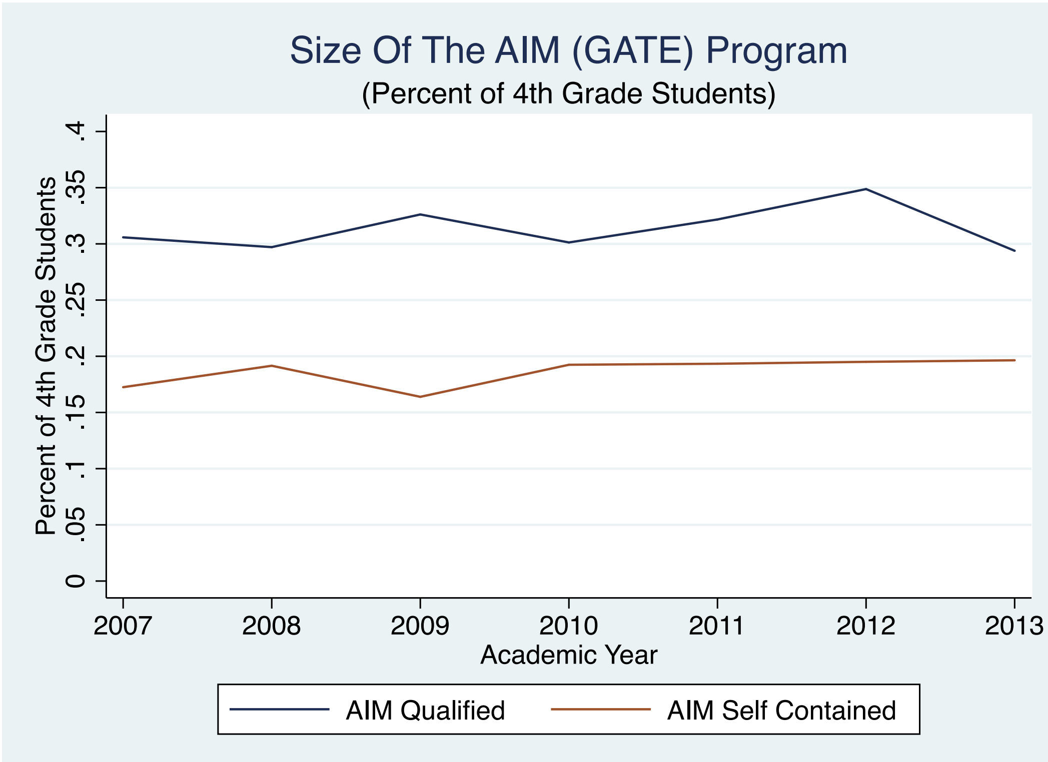
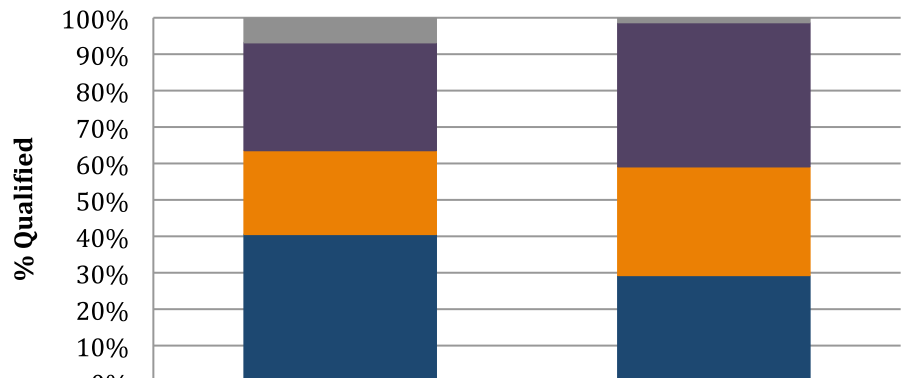
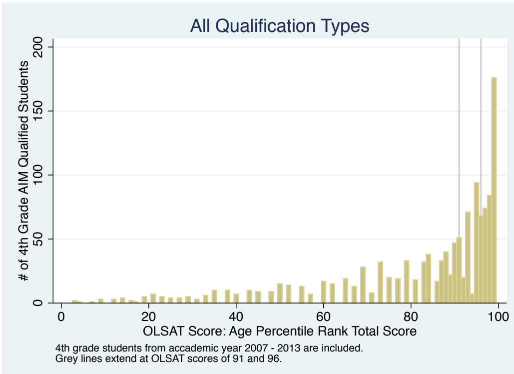


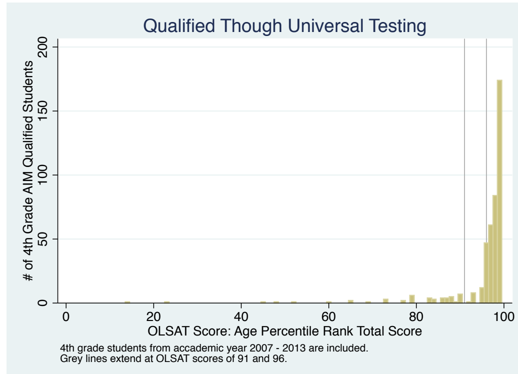
Figure 3:
AIM (GATE) Qualification Type by Master Plan

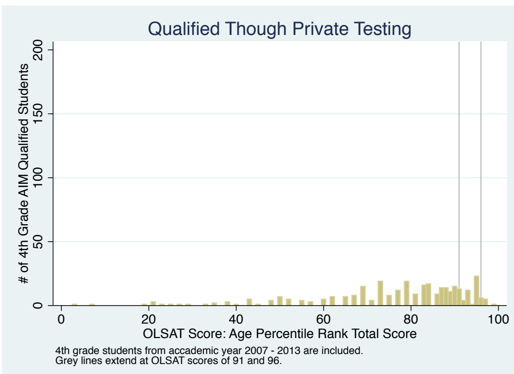


	Master Plan 2005	Master Plan 2008
Transfers / Other	6.8%	1.2%
Retesting	29.6%	39.6%
Private Testing	23.0%	29.8%
Universal Testing	40.6%	29.3%

Figure 4:
OLSAT Age Percentile Rank Scores of 4th Grade AIM Qualified Students By Qualification Type


All Qualification Types
# of 4th Grade AIM Qualified Students
0
50
100
150
200
0
20
40
60
80
100
OLSAT Score: Age Percentile Rank Total Score
4th grade students from accademic year 2007 - 2013 are included.
Grey lines extend at OLSAT scores of 91 and 96.


Qualified Though Universal Testing
# of 4th Grade AIM Qualified Students
0
50
100
150
200
0
20
40
60
80
100
OLSAT Score: Age Percentile Rank Total Score
4th grade students from accademic year 2007 - 2013 are included.
Grey lines extend at OLSAT scores of 91 and 96.


Qualified Though Private Testing
# of 4th Grade AIM Qualified Students
0
50
100
150
200
0
20
40
60
80
100
OLSAT Score: Age Percentile Rank Total Score
4th grade students from accademic year 2007 - 2013 are included.
Grey lines extend at OLSAT scores of 91 and 96.

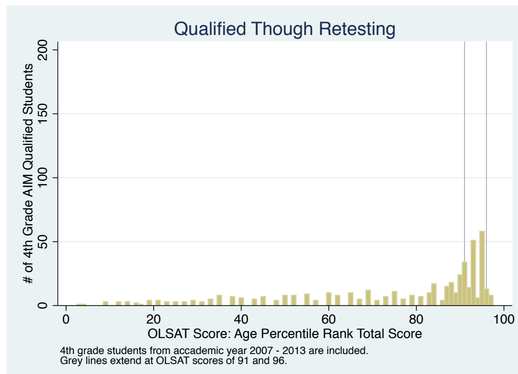

Qualified Though Retesting
# of 4th Grade AIM Qualified Students
0
50
100
150
200
0
20
40
60
80
100
OLSAT Score: Age Percentile Rank Total Score
4th grade students from accademic year 2007 - 2013 are included.
Grey lines extend at OLSAT scores of 91 and 96.

Figure 5: OLSAT Age Percentile Rank Scores of 4th Grade AIM Qualified Students By Year

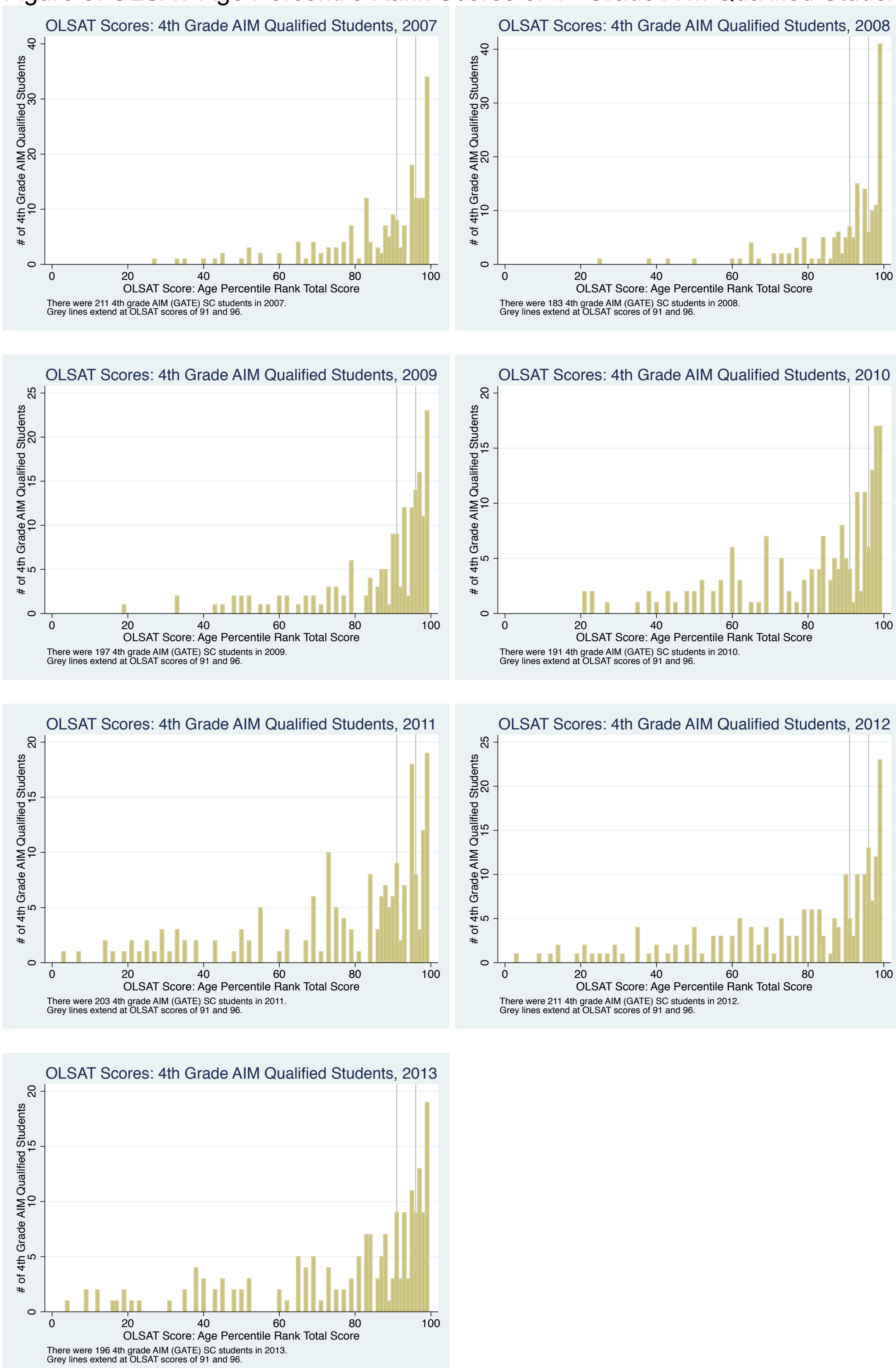
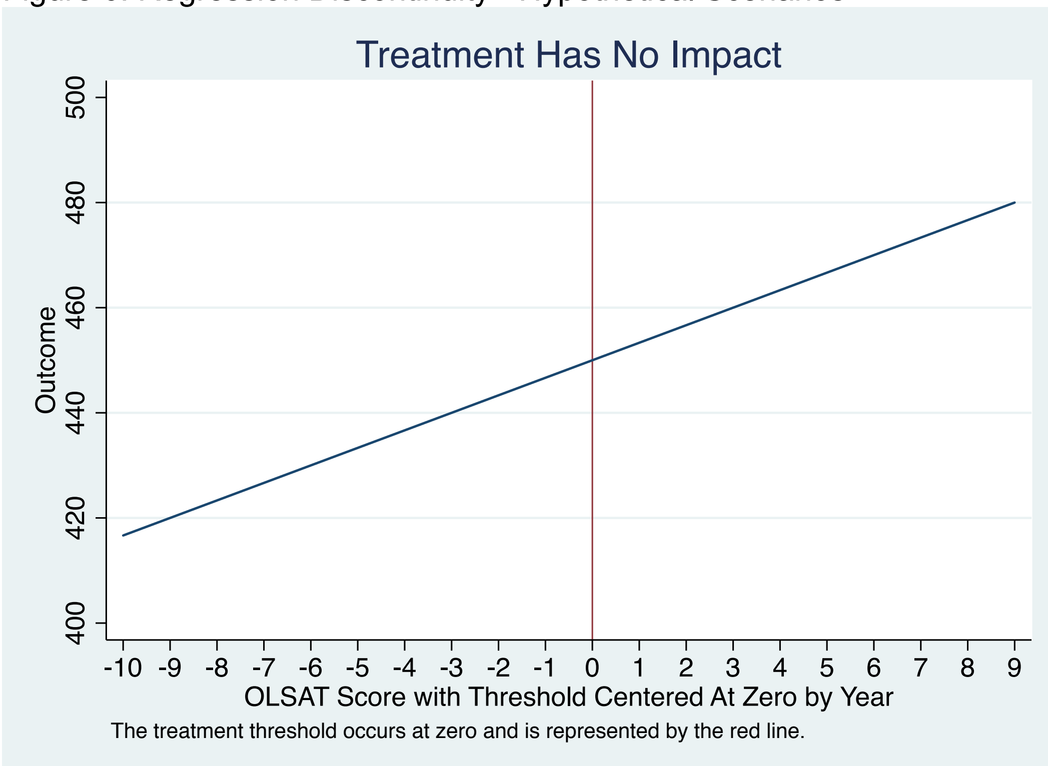


Figure 6: Regression Discontinuity - Hypothetical Scenarios



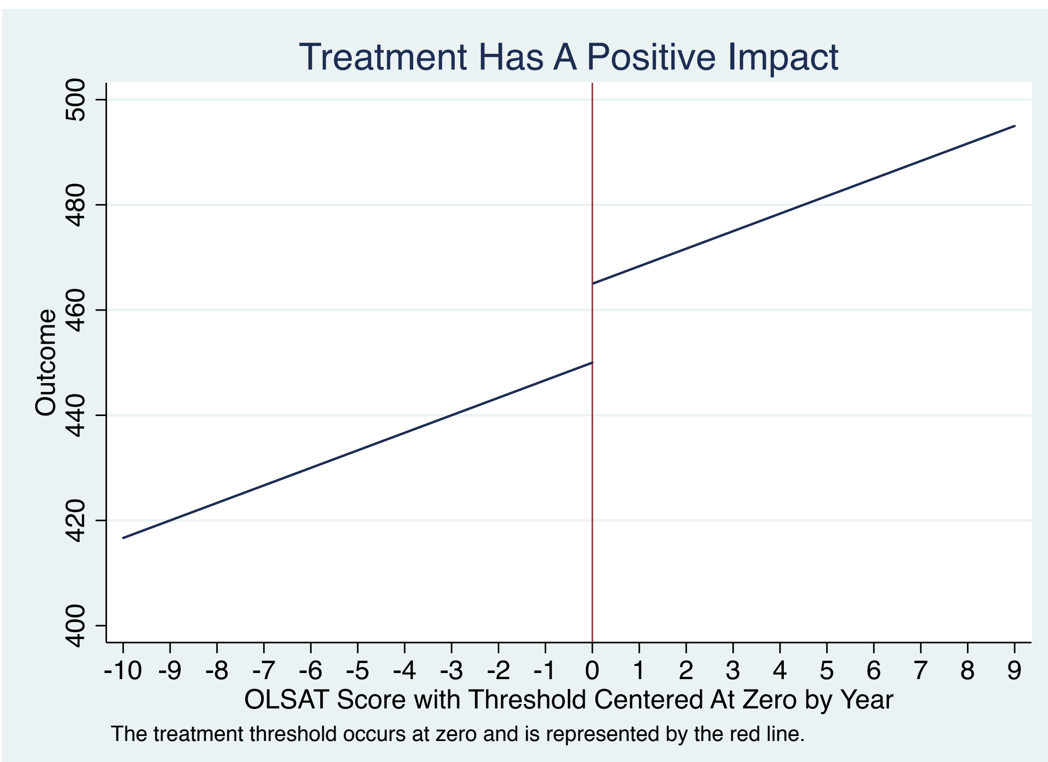


Figure 6: Regression Discontinuity - Hypothetical Scenarios (cont.)

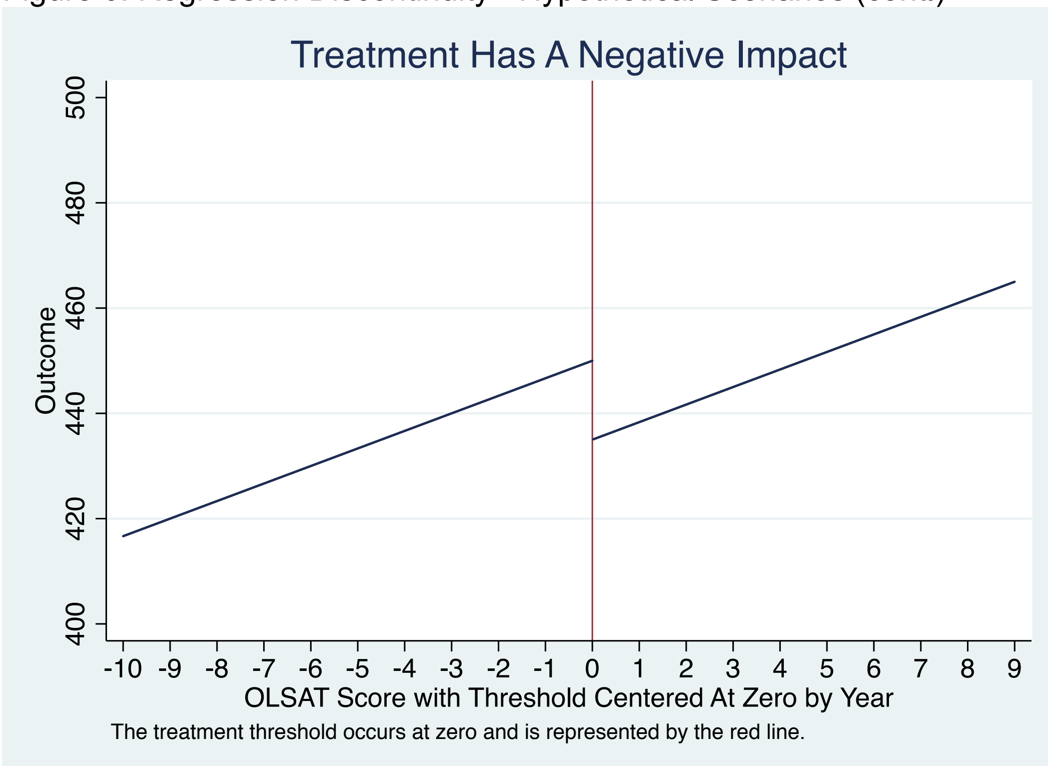
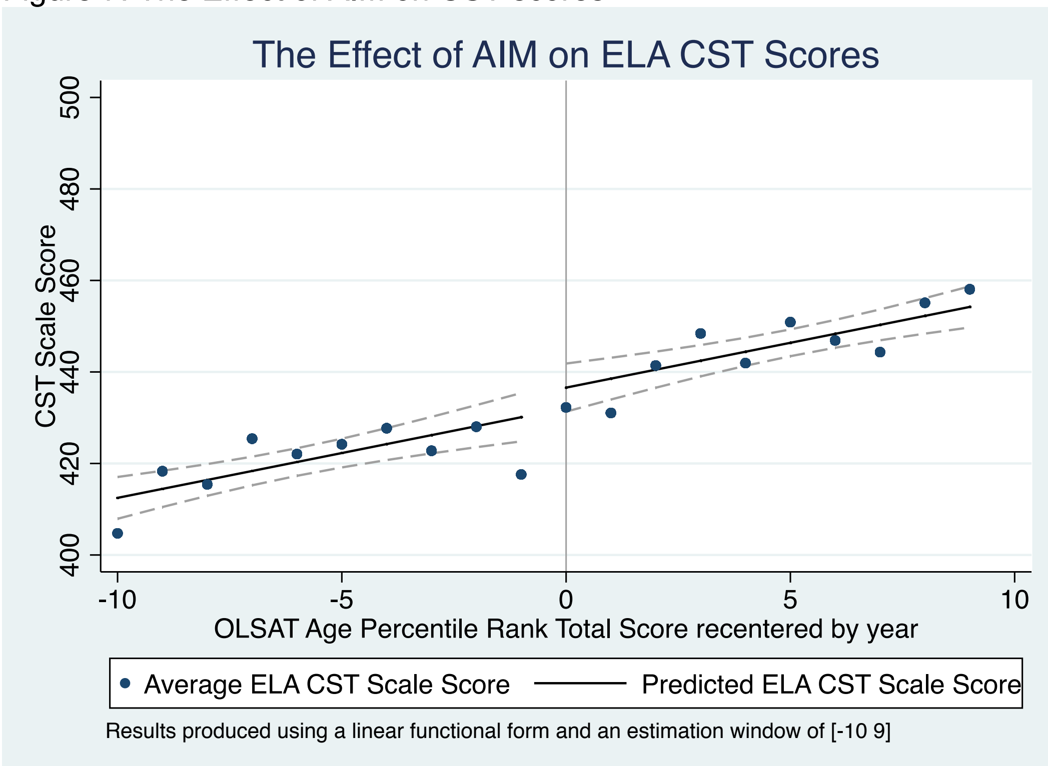


Figure 7: The Effect of AIM on CST scores



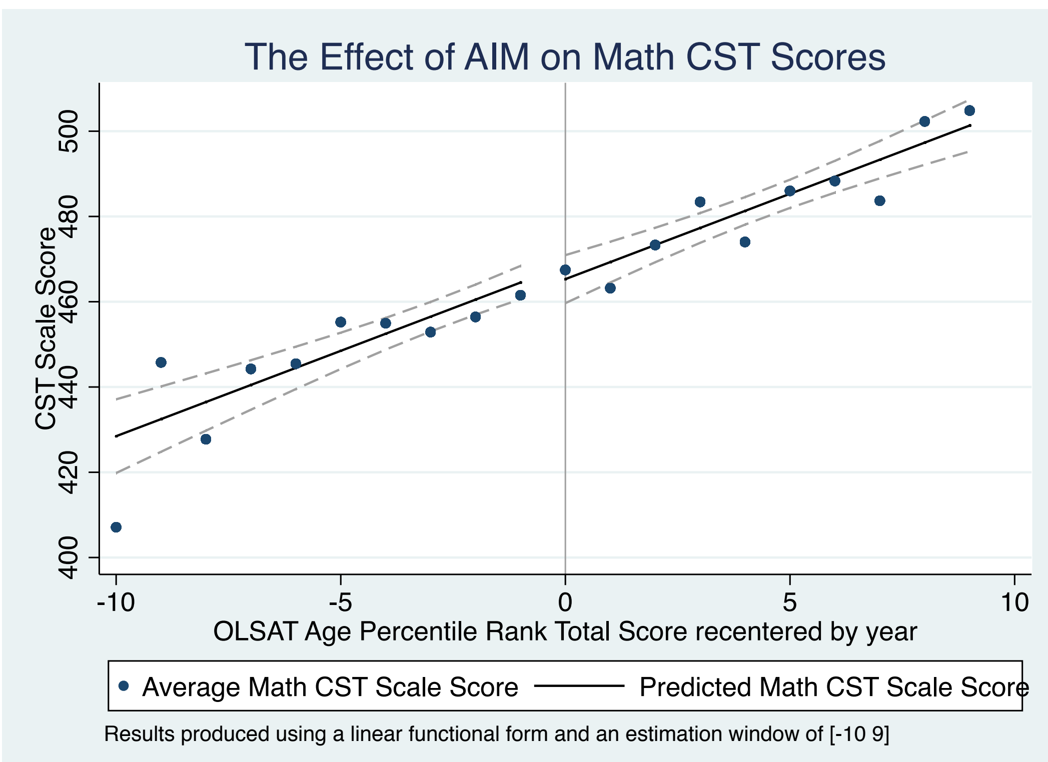


Table A1: Affect of AIM on Students in AIM

VARIABLES	Average CST	Math CST	ELA CST
AIM Effect	-1.557	-6.355	3.230
	(4.407)	(4.269)	(5.514)
Observations	5,830	2,914	2,916
R-squared	0.232	0.182	0.152

Estimates produced using a [-10, 9] window and a linear functional form
Standard errors clustered at the xi level in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure A1: