NBER WORKING PAPER SERIES

DOES GIFTED EDUCATION WORK? FOR WHICH STUDENTS?

David Card
Laura Giuliano

Working Paper 20453
http://www.nber.org/papers/w20453

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2014

We are extremely grateful to Cynthia Park and Jacalyn Schulman for their assistance in accessing and interpreting the data used in this study, and to Sydnee Caldwell, Alessandra Fenizia, Yosub Jung, Hedvig Horvath, Attila Lindner, Carl Nadler and Kevin Todd for outstanding research assistance. We also thank Kelly Bedard, Jesse Rothstein, Enrico Moretti, Chris Walters, and seminar participants at UC Berkeley, UC Riverside, UC Santa Barbara, UMass Amherst, U.Virginia and the NBER Summer Institute for helpful discussions and suggestions. This research was supported by a grant from the Institute of Education Sciences (R305E110019). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Does Gifted Education Work? For Which Students?
David Card and Laura Giuliano
NBER Working Paper No. 20453
September 2014
JEL No. I21,I24

ABSTRACT

Education policy makers have struggled for decades with the question of how to best serve high ability K12 students. As in the debate over selective college admissions, a key issue is targeting. Should gifted and talented programs be allocated on the basis of cognitive ability, or a broader combination of ability and achievement? Should there be a single admission threshold, or a lower bar for disadvantaged students? We use data from a large urban school district to study the impacts of assignment to separate gifted classrooms on three distinct groups of fourth grade students: non-disadvantaged students with IQ scores ≥130; subsidized lunch participants and English language learners with IQ scores ≥116; and students who miss the IQ thresholds but scored highest among their school/grade cohort in state-wide achievement tests in the previous year. Regression discontinuity estimates based on the IQ thresholds for the first two groups show no effects on reading or math achievement at the end of fourth grade. In contrast, estimates based on test score ranks for the third group show significant gains in reading and math, concentrated among lower-income and black and Hispanic students. The math gains persist to fifth grade and are also reflected in fifth grade science scores. Our findings suggest that a separate classroom environment is more effective for students selected on past achievement – particularly disadvantaged students who are often excluded from gifted and talented programs.

David Card
Department of Economics
549 Evans Hall, #3880
University of California, Berkeley
Berkeley, CA 94720-3880
and NBER
card@econ.berkeley.edu

Laura Giuliano
Department of Economics
University of Miami
P.O. Box 248126
Coral Gables, FL 33124-6550
l.giuliano@miami.edu

Over the past century gifted education programs for high-ability students have expanded from a handful of cities to serving nearly 7% of the U.S. student population. The growth of these programs has been accompanied by an ongoing debate over how they should be targeted. Early gifted programs used IQ as the basis for eligibility.[1] By the 1970s this practice was under attack: critics charged that IQ tests were racially biased and argued more broadly that eligibility should be based on a combination of cognitive *and* non-cognitive traits (Renzulli, 1978; U.S. Department of Education 1993). The lack of consensus over the appropriate targeting of gifted education is reflected in the wide variation in state policies. A third of states still mandate the use of IQ scores to identify gifted students -- in some cases allowing lower thresholds for disadvantaged students. Many other states use multiple admission criteria that reward both cognitive ability and achievement.[2]

Despite the longstanding controversy there is little credible evidence on the relative effectiveness of gifted education programs for different target groups. In fact, there is almost no evidence on even the *average* impact of these programs. The best study to date, by Bui et al. (2011), suggests that gifted programs have no overall effect on standardized scores of gifted students. Studies of tracking programs likewise provide no clear guidance on whether higher-ability students should be allocated to special classes or integrated with other students.[3]

In this paper we use detailed administrative data from one of the country's largest school districts ("the District") to study the impacts of an intensive gifted education program on different target groups. The design of the District's program allows us to examine heterogeneity along the two main dimensions that are central to the targeting debate: whether the student is selected on the basis of IQ

[1] Chapman (1988), Jolly (2009) and VanTassel-Baska (2010) discuss the history of gifted education in the U.S.

[2] McClain and Pfeiffer (2012) survey recent changes in state policies. There has been a general trend toward more flexible and inclusive policies for determining gifted eligibility. In 2008, however, New York City switched to a uniform test score cutoff for gifted eligibility (New York Times, June 19, 2008).

[3] See Slavin (1987) and Betts (2011) for reviews. A few recent studies have found benefits of tracking in non-U.S. settings—including Kenya (Duflo, Dupas and Kremer 2011), Trinidad and Tobago (Jackson 2010), Romania (Pop-Eleches and Urquiola 2013), and Iceland (Vardardottir 2013). But two recent studies of exam-based high schools in Boston and New York have found little impact (Abdulkadiroglu et al. 2011; Dobbie and Fryer 2011).

or academic achievement; and whether the student comes from an advantaged or disadvantaged background.

We focus on the District's program in fourth and fifth grades, which places three distinct groups of students in separate gifted classrooms. The first group -- known as "Plan A" gifted students -- consists of non-disadvantaged students who score at least 130 points on a standard IQ test, the state-required threshold for gifted eligibility. The second group -- known as "Plan B" gifted students -- includes English language learners and free- and reduced-price lunch (FRL) participants with IQ's over 116 points, the lower threshold allowed under state law for disadvantaged students. Remaining seats in the gifted classrooms are filled by a third group of *non-gifted* students -- known as high achievers -- who scored highest among their school/grade cohort in statewide achievement tests in the previous year. Comparisons across these groups reveal the impacts of the *same* classroom environment on students selected on the basis of cognitive ability versus achievement. We can also compare the impacts for advantaged and disadvantaged subgroups of gifted and non-gifted participants.

We use the District's eligibility rules to construct regression discontinuity (RD) estimates of program impacts using the IQ scores of gifted students and the test score ranks of high achievers. A key problem is that the IQ scores of marginally eligible students are often boosted to meet the gifted thresholds.[4] Nevertheless, the pre-program test scores and other characteristics of students on either side of the IQ thresholds are similar, suggesting that conventional RD models may still yield consistent program estimates. We confirm the validity of these models by comparing them to "donut-hole" specifications that ignore data close to the threshold (Barreca et al., 2011; Bajari et al., 2011); to first-differenced specifications that compare test score gains for students above and below the threshold; and to donut-hole first-differenced models that measure test score gains for students away from the IQ threshold. Building on Angrist and Rokkanen (2012) we show that test score gains are orthogonal to IQ,

[4] As explained below, we use first recorded IQ scores for students whose first test was conducted by a District psychologist to avoid selective reporting of initial scores and the possibility of selective re-testing.

allowing us to infer program effects for a broader group of students whose status is unlikely to be affected by IQ manipulation.

For non-disadvantaged (Plan A) gifted students our alternative specifications yield uniformly small and insignificant reduced form impacts of gifted program participation on standardized achievement in math, reading, and writing. For disadvantaged (Plan B) gifted students the results are also consistent across specifications and show negligible impacts on math and reading, though a positive effect on writing. Importantly, the data for both groups suggest that even infra-marginal Plan A and Plan B students gain little from the receipt of gifted services.

In contrast, RD models based on past achievement score ranks show that placement in a fourth grade gifted classroom has significant positive effects on reading and math scores of non-gifted high achievers. The estimates are particularly large for low-income (i.e., FRL-eligible) and minority students. Treatment-on-the-treated estimates for these subgroups imply that participating in a fourth grade gifted classroom raises fourth grade math and reading scores by 0.4 - 0.5 standard deviations, with persistent positive effects on math and science scores in fifth grade.

We confirm these conclusions using an alternative school-level design that compares the top 20 non-gifted students at schools where there were no gifted children in the cohort (and no gifted class for fourth grade) to students at schools where there were 1-4 gifted students (and hence a gifted class with about 20 seats for high achievers). This design also allows us to compare the impacts on high achievers closer and further from the eligibility threshold, and to verify that moving the highest achievers to a separate classroom has no effect on the scores of the next-highest group of students (ranked 25-44 in their cohort).

Our finding that the District's gifted program has little effect on the reading or math scores of gifted students is consistent with the results in Bui et al. (2011), and with the findings in two recent studies of elite selective admission high schools (Abdulkadiroglu et al., 2014; Dobbie and Fryer 2011).

More surprising is our finding that placement in a gifted classroom has relatively large positive effects on non-gifted participants. We conclude that the separate gifted classrooms offered in the District are more effective for students targeted on the basis of past achievement than for those targeted on the basis of cognitive ability.

A second important finding is that the estimated impacts of a gifted classroom environment are systematically larger for lower-income and minority high achievers—groups that are the least likely to meet the traditional criteria for gifted program participation. The presence of so many of these students in the District's gifted classrooms is attributable to the policy of providing a separate gifted classroom whenever there is at least one gifted child in the school-grade cohort, and allocating the remaining seats to students at the school using a relative criteria, rather than an absolute threshold. Our results suggest that such a broad-based program -- similar in spirit to the "percent plans" for admission to public universities in Texas and other states -- can be effective in raising the achievement of relatively high-ability students in lower performing schools, potentially creating a pathway for upward mobility.

II. Gifted Education in the District

Most elementary schools in the District offer part-time individualized instruction for students in first through third grades who have been identified as gifted. Starting in 2004 the District required schools to set up separate classrooms for all gifted students in fourth and fifth grades, with any open seats allocated to students with the highest scores in the previous year's standardized tests. Since most schools have only a handful of gifted children per grade, and class sizes are maintained at 20-24 pupils, most gifted classrooms in the District contain a mixture of gifted students and high-achievers.[5] Gifted classroom teachers must complete a 5-course sequence in gifted education offered by the District, though they receive no higher pay, and are on average only slightly more experienced than other

[5] In school year 2007-08 the average number of gifted fourth grade students at the 140 regular elementary schools in the District was 8.5. Only 8% of the schools had over 20 gifted students in fourth grade.

teachers at their school.

The students in a typical gifted classroom are highly selected, with standardized test scores well above the average for their school. Nevertheless, gifted classes follow the same curriculum as other classes, and the students write the same statewide achievement tests each year. Interviews with gifted teachers suggest that they typically divide their classes into ability groups and assign enrichment projects for students who complete the regular curriculum material more quickly.

Until 2004, potential candidates for gifted status were identified through parent and teacher referrals. Recommended students were directed to a District psychologist for an IQ test. Parents could also pay for testing by a private psychologist and submit the results to the school. Students with IQ's above the relevant threshold are eligible for gifted status, with the final determination made in consultation between parents, teachers, and the school's Exceptional Student Education (ESE) specialist. The ESE specialist also draws up an Individualized Education Plan for each gifted student that specifies learning goals and instructional plans tailored to the student's strengths and weaknesses.

The IQ cutoffs used by the District are set by state law. This law also has two special provisions for students who miss the cutoff on their first test attempt. First, those who score within a standard error of the Plan A threshold (i.e., ≥127 points) can be classified as gifted if there is "overwhelming evidence" of superior ability.[6] As we see below, this is most likely to occur for students with IQ's of 129. Second, students may be retested within the same year using a different IQ test; or after one year using the same test. Many non-disadvantaged (Plan A) students who score below 130 points on their first attempt are re-tested by a private psychologist and ultimately achieve gifted status. A relatively small fraction of disadvantaged (Plan B) students are retested, mainly by District psychologists.

In response to concerns about the low numbers of disadvantaged students in gifted education,

[6] A state interpretative memo reads: "Eligibility committees may also judge a student eligible if there is overwhelming evidence in favor of more liberal interpretation, that is, by considering the standard error of measurement." (Florida Department of Education 1996). There is no similar allowance for Plan B students.

the District introduced a universal screening program in 2005. Under this program, all students completed the Naglieri Non-verbal Ability Test (NNAT) in second grade. Disadvantaged (FRL/ELL) students with scores ≥115 points and non-disadvantaged students with scores ≥130 points were referred to a District psychologist for IQ testing.[7] As in earlier years, teachers and parents could still recommend students for IQ testing, and parents could submit scores from private testing agents. Comparisons across cohorts of third-graders, shown in Figure 1, suggest that the screening program raised the gifted fraction of disadvantaged students from around 1.5% to nearly 4%. A financial crisis in 2007 caused the District to cut funding for IQ testing, leading to a drop in the placement rate of Plan B students. The screening program was temporarily suspended in 2010 and re-introduced with a new test in 2011, but as of 2012 the gifted fraction of disadvantaged students had only rebounded slightly.

Figure 2 shows how the composition of gifted fourth grade classrooms varies with the overall fraction of FRL students at different schools.[8] We distinguish four groups of participants: Plan A and Plan B gifted students, advantaged (non-FRL/ELL) high achievers, and disadvantaged (FRL/ELL) high achievers. Plan A gifted children are concentrated at richer schools with lower FRL rates, whereas Plan B children are concentrated at schools with moderate to high levels of FRL participation. Advantaged high achievers are more evenly distributed, and represent about 40% of the students in gifted classrooms at schools with 10-70% FRL participation. Disadvantaged high achievers are the modal group at schools with FRL rates over 65%, and make up the vast majority of participants at very poor schools.

III. Student Achievement Data and Analysis Samples

We use administrative data for students who were in third grade from 2004 to 2011 to study the

[7] For screening purposes the District used a rescaled version of the NNAT with mean 100 and standard deviation 15. The NNAT is designed to measure IQ but is known to be relatively noisy—see Lohman et al. (2008).

[8] The 140 schools included in our analysis have at least 50 students on average in third grade (in years when the school is open) and are not classified as charter schools. The FRL fraction is strongly positively correlated (ρ=0.8) with the fraction of black non-Hispanic students in a school and strongly negatively correlated (ρ=-0.9) with average third grade scores in reading and mathematics.

District's gifted program. We have information on gender, race, ethnicity, and FRL/ELL status for all students, as well as NNAT test scores for second-graders from 2005 to 2009 and IQ test scores for students who were tested by the District or submitted a score. In addition we have state-wide achievement test scores in reading and math for third through tenth grades, fourth grade writing test scores, and fifth grade science scores. We use fourth grade test scores as our main outcome measures, and third grade scores as a measure of baseline (i.e., **pre**-program) achievement for students who participate in gifted classrooms.

A concern with the use of third grade scores as a baseline is that some gifted students have already received individualized instruction in second or third grade. For many students in our sample we also have access to Stanford Achievement Tests (SAT's) in reading and math for first through third grades.[9] As a robustness check, we use these tests to measure baseline achievement *before* gifted students have been exposed to any form of gifted programming. As discussed below, we find very similar results using either set of scores, and in the interests of maximizing our sample sizes we use the third grade scores as the baseline measures for our main specifications.

Table 1 shows the overall characteristics of students who attended one of the District's 140 regular elementary schools in third grade, as well as the characteristics of the samples we use to study the impacts of gifted education on the three groups of participants in fourth grade gifted classrooms.[10] The overall student body in third grade (column 1) is highly diverse, with 30% white non-Hispanics, 37% black non-Hispanics, 26% Hispanics, and 4% Asians. Roughly half of all students are eligible for free or reduced price lunches, 10% are English language learners, and 55% are either FRL or ELL and therefore fall under the Plan B eligibility rules. We use 2000 Census data on median family incomes by zip code to proxy family income: the average level is $57,500. Overall, about 6% of students in the available cohorts

[9] The District required SAT math tests in every grade until 2007 and the SAT reading tests until 2008.

[10] We restrict these analysis samples to students who appear in the District in fourth grade in the year after they are observed in third grade (between 2005 and 2012). As discussed below, we find no evidence that gifted classroom assignment affects attrition from the District.

were classified as gifted by fourth grade and 13% were assigned to a gifted classroom.

The lower rows of the table report mean IQ and NNAT scores and mean scores on third and fourth grade state tests. We also show the mean third grade scores of students' school-wide peers, and the mean scores of students in their schools' fourth grade gifted classrooms. IQ and NNAT test scores are missing for many students, so we report (in italics) the shares of students with valid scores on each test. Both tests are scaled to have mean 100 and standard deviation 15 in a national population. We standardize the statewide achievement test scores to have mean 0 and standard deviation 1 within a grade/year cohort in the District.

Column 2 shows the characteristics of our analysis sample for studying Plan A participants in gifted education. This sample includes non-disadvantaged students who have an IQ test on file by the time they enter fourth grade within 10 points of the cutoff for Plan A eligibility. To avoid problems of selective reporting and retesting we use each student's first recorded IQ test, and limit the sample to students whose first recorded test was administered by a District psychologist.[11] Compared to the overall third grade population, Plan A sample members are more likely to be white/non-Hispanic and are drawn from relatively rich zip codes. Their schools have relatively low FRL rates and relatively high test scores, though not as high as the Plan A sample members themselves, who score 1.2-1.4 standard deviation units (σ's) above the district-wide average in reading and math.

Column 3 presents the characteristics of our analysis sample for studying Plan B participants. These students are either English learners or FRL-eligible with a first-reported IQ score between 105 and 125 points. They are drawn from schools with roughly average test scores but slightly more free lunch participants (61% versus 52%). Students in the Plan B sample have test scores 0.5-0.6 σ's above the district average—a little below the mean scores of the students in the gifted classrooms in their schools.

Columns 4 and 5 show the characteristics of our samples for studying high-achievers. To

[11] Parents have little or no incentive to report private tests below 130 points and nearly all private test scores recorded by the District are above 130.

construct these samples we first identified schools with a fourth grade classroom containing at least one non-gifted student. Under the District's rules, open seats in a gifted classroom are filled by non-gifted students in the same grade with the highest scores on the previous year’s statewide tests. Starting in 2009 the District imposed a uniform ranking formula based on math and reading scores. We therefore identified schools that appear to have followed this formula for fourth graders in 2009-2012. We then calculated school-specific cutoff scores for admission to the fourth grade gifted class and selected the first 10 students with scores above this cutoff and the first 10 with scores below it.[12] Finally, we classify these students as either disadvantaged (ELL or FRL-eligible) or advantaged.

Comparisons with the other groups in Table 1 suggest that advantaged high achievers are from similar neighborhoods and schools as the Plan A analysis sample, while disadvantaged high achievers are from similar neighborhoods and schools as the Plan B analysis sample. Advantaged high achievers have lower test scores than students in the Plan A sample, but still score well above average. Disadvantaged high achievers have somewhat lower scores—closer to those of the students in the Plan B sample.

IV. Validity of RD Design and First-Stage Relationship

a. Evidence of Manipulation of IQ Scores

In an ideal RD design the running variable is exogenously determined in the vicinity of the threshold (see Lee, 2008 and McCrary, 2008). In the case of IQ tests, however, psychologists have some discretion in assigning scores and can easily boost marginal students above the gifted threshold.[13] Figures 3a and 3b show the frequency distributions of scores for our Plan A and Plan B evaluation

[12] In principle the cutoff is the test score of the lowest-scoring non-gifted child in the gifted classroom. As explained in the Data Appendix, we reduce errors caused by missing scores and class size fluctuations by choosing a cutoff score to minimize the misclassification rate of students whose scores are outside an interval around the potential threshold. The number of students above the cutoff is less than 10 if the gifted class has <10 extra seats.

[13] For example, in one section of the Wechsler Intelligence Scale for Children (4th edition), children are given two words – like “red and blue” – and are scored 0, 1, or 2 points based on their explanation for their similarity.

samples.[14] Despite the fact that we use only first-reported IQ scores from District-administered tests, both histograms show evidence of manipulation, with spikes at the minimum threshold scores for each group and deficits below the thresholds.

To aid in interpreting the observed distribution of Plan A scores we fit a simple model to the histogram in Figure 3a, assuming a quadratic frequency distribution of true scores and allowing arbitrary fractions of students with true scores of 127-129 to be “bumped up” to 130 or 131 points. The fitted model, shown by the lighter bars in the figure, provides a reasonable approximation to the true distribution, apart from the fraction of scores in the 125-126 range.[15] We also superimpose the implied frequency distribution from the fitted model assuming no manipulation. Relative to this counterfactual, the model implies that two-thirds of the observations at 130 points and one-third at 131 points are attributable to students with true scores of 127-129 points.

We fit a similar model to the Plan B histogram in Figure 3b, assuming that arbitrary fractions of test takers who score 114 or 115 are bumped up to 116 or 117 points. As shown by the fitted histogram (plotted with lighter bars) the model yields a plausible fit, though a formal test rejects the model at conventional significance levels (chi-square=36 with 15 degrees of freedom). Again, we also superimpose the implied frequency distribution in the absence of manipulation. This benchmark suggests that around 40% of students with a reported score of 116 have true scores of 114 or 115.

While there is clearly some manipulation of IQ scores, the fundamental concern for an RD analysis is whether this leads to systematic differences in the latent abilities of students on either side of the threshold. Some direct evidence on this issue is provided in Figure 4, where we show the relationship between IQ and four key variables: median household income, the NNAT screening test, and third grade reading and math scores. Looking first at the Plan A students, none of the variables

[14] The distribution of ranks in our high-achiever sample is very smooth so we do not show it.

[15] The goodness of fit statistic is 82.2 with 14 degrees of freedom, which is highly significant. If we alter the model by allowing scores as low as 125 points to be bumped up we find too many scores missing from below the threshold relative to the size of the spike above.

exhibits a large discontinuity at the 130 point threshold, although there is a small jump in math scores. (Formal test statistics for the baseline reading and math scores are discussed in the next section). Further evidence is presented in the first two columns of Appendix Table 1, where we show the difference in mean characteristics for students with IQ's of 130-131 points versus 127-129 points, as well as coefficients from a logit model for the probability of being above the threshold, conditional on being in the 127-131 point range. None of the differences in the individual characteristics is significant at the 5% level, though the gap in third grade math scores is marginally significantly (t=1.8). The logit model confirms that as a group the characteristics are not jointly significant predictors of being above the gifted threshold (p-value=0.38). We conclude that the assignment of students to above or below the gifted threshold may be almost "as good as random," with any bias likely to lead to an *overstatement* of the effect of gifted participation. As a result we present conventional RD results for Plan A students below, including models that control for previous test scores. We also present donut-hole specifications and donut-hole first-differenced models that are robust to manipulation of IQ scores.

The plots in Figure 4 for the Plan B analysis sample also show no large discontinuities in the four background characteristics. The similarity of students on either side of the 116 point threshold is confirmed in Appendix Table 1 (columns 3-4) where we compare mean characteristics for students with IQ's of 116 and 117 points versus those with IQ's of 114 or 115, and present logit coefficients from a model predicting who is above the threshold. As in the Plan A sample, none of the mean differences is significant, nor are the characteristics jointly significant in a logit model for scoring above the threshold.

Although we have no reason to suspect manipulation in the scores used to rank non-gifted students for placement in a gifted class, we also checked that the characteristics of the high achievers evolve smoothly as their rank passes the eligibility threshold. As shown in Figure 4, this appears to be the case. These visual impressions are confirmed by the test statistics presented in Table 4, below.

b. Differential Attrition

To be included in our Plan A and Plan B analysis samples a student must receive a District-administered IQ test by the start of fourth grade, remain in a District school through the end of fourth grade, and be promoted one grade per year. Similarly, high achievers must stay in the District between third grade and fourth grades. Any differential attrition of students with IQ's (or test score ranks) just above or just below the eligibility thresholds poses a potential threat to the validity of our RD models.

To check these concerns we conducted an analysis of sample retention rates, summarized in Appendix Figure 1. For our Plan A and Plan B samples we identified students with a first-time District-administered IQ test in first through third grades from 2002 to 2011, and examined the relationship between IQ scores and the probability of being included in our fourth grade sample. For high-achievers (who learn their placement status after enrolling in fourth grade) we examined the relationship between the rank at the start of fourth grade and the probability of remaining in the District through the end of the school year. We find no evidence of discontinuities in retention rates for any of the three groups.[16] Based these comparisons, and results from a series of RD models, we conclude that differential attrition is not a concern for our analysis.

c. First Stage Relationships for Gifted Status and Placement in a Gifted Classroom

We turn now to the first-stage relationships between IQ scores (or test score ranks) and participation in gifted programs. For Plan A and Plan B students we consider two measures of participation: achieving gifted status by fourth grade and placement in a gifted classroom in fourth grade. For our high achiever analysis sample we consider only placement in a gifted classroom.

Figure 5a plots the fraction of students classified as gifted, and the fraction placed in a gifted

[16] This finding differs from that of Davis et al. (2010) who find that in a medium-sized Midwestern district, students eligible for the gifted program are more likely to stay in the district. One possible explanation for the lack of an attrition effect in our sample may be that families whose enrollment decisions hinge on gifted program eligibility are likely to opt for private IQ testing, and are therefore excluded from our estimation sample.

classroom, against the IQ scores of students in our Plan A sample. It also shows the fitted relationships from local linear models (with a bandwidth of 10 points). To the left of the threshold the fraction placed in a gifted classroom is higher than the fraction gifted, reflecting the fact that many high-IQ students enter gifted classes as high achievers. Both series also rise with IQ, reflecting a rising likelihood of being retested and scoring above the threshold, and the possibility of achieving gifted status if IQ is within 3 points of the 130 threshold. The latter phenomenon appears to be most likely for students with an IQ score of 129. As noted in Figure 3a, however, there are very few students with IQ's of 129, and their presence has little impact on the estimated models. To the right of the threshold nearly 100% of all students are both classified as gifted and placed in a gifted classroom. The local linear models imply a first stage discontinuity in the probability of being classified as gifted of about 50 percentage points (pp), and a discontinuity in the probability of participating in a gifted class of about 25 pp.

Figure 5b shows similar first stage models for Plan B students. To the left of the 116 point threshold the fraction classified as gifted is low and exhibits little upward trend, consistent with the fact that very few FRL/ELL students who miss the gifted threshold are retested. The fraction of students who are placed in gifted classes, however, is increasing, reflecting the rising likelihood of being placed in a gifted classroom as a high achiever. There is also an upward jump in the placement rate in a gifted class for students with IQ's of 115, though the fraction classified as gifted is relatively smooth.

Unlike Plan A students, only 55% of Plan B students just to the right of the IQ threshold are classified as gifted, possibly because of concerns that marginally eligible Plan B students will be mismatched to the level of other students in the gifted classroom. Despite this non-compliance, the local linear models show large discontinuities at the eligibility threshold in the probability of being classified as a gifted student (≈50 pp) and in entering a gifted classroom in fourth grade (≈35 pp).

Figure 5c shows the relationship between test score ranks and the probability of placement in a gifted classroom for our high achiever sample. Despite our attempt to measure the cutoff scores accurately, the placement rate just to the left of the threshold is 25%, while the rate just to the right is only 55%, implying a jump of about 30 pp. This fuzziness is attributable to a combination of measurement errors in the cutoff scores and some non-compliance with the District formula (see Data Appendix for details). First stage models fit separately for advantaged and disadvantaged high achievers are broadly similar to Figure 5c, with a somewhat larger jump at the threshold for the advantaged subgroup (≈38 pp) than the disadvantaged subgroup (≈26 pp).

d. Discontinuities in Peer Quality

An important feature of gifted classrooms is the highly selective nature of the *other* students in the class. Figure 6 shows the relationships between IQ or test score rank of individual students and four measures of the quality of their classroom peers: their mean third grade reading and math scores (panels a and b); the fraction who are gifted (panel c); and the fraction with a learning disability (panel d). For all three samples we find positive discontinuities in the lagged test scores of classroom peers (on the order of 0.2 σ's) combined with upward jumps in the fraction of gifted peers and downward jumps in the fraction of learning-disabled peers.

V. Impacts on Student Achievement

a. Framework

We hypothesize that participation in the District's fourth grade gifted program affects test scores through three main channels: (1) a direct effect of the gifted classroom environment and teacher; (2) increased quality of classroom peers; (3) receipt of individualized gifted services. A simple structural model that incorporates all three channels is:

$$(1) \qquad y = \beta_1 D_{class} + \beta_2 Q_{peer} + \beta_3 D_{gifted} + \beta_x X + \lambda\, IQ^* + \varepsilon$$

where y represents a student's test score at the end of fourth grade, D_{class} is an indicator for being placed in a gifted class, Q_{peer} is a measure of peer quality, D_{gifted} is an indicator for being classified as gifted (and thus receiving individualized instruction), X is a set of covariates (e.g., gender, race/ethnicity, school dummies), IQ^* is the student's true cognitive ability (as would be measured by a "perfect" IQ test), and ε reflects all remaining determinants of test scores. Taking expectations conditional on observed IQ:

(2) $E[y|IQ] = \beta_1 P[D_{class}=1|IQ] + \beta_2 P[D_{gifted}=1|IQ] + \beta_3 E[Q_{peer}|IQ] + E[\beta_x X + \lambda IQ^* + \varepsilon|IQ]$.

Assuming that the X's vary smoothly at the gifted threshold (as in Figure 4), and that $E[IQ^*|IQ]$ and $E[\varepsilon|IQ]$ are both continuous at the threshold (as is true if any manipulation of observed IQ scores is ignorable), the discontinuity in the conditional expectation function E[y|IQ] at the gifted threshold (T) is:

(3) $\mathbf{Dis}(y) = \lim_{IQ\downarrow T} E[y|IQ] - \lim_{IQ\uparrow T} E[y|IQ] = \beta_1 \mathbf{Dis}(P_{class}) + \beta_2 \mathbf{Dis}(Q_{peer}) + \beta_3 \mathbf{Dis}(P_{gifted})$,

where $\mathbf{Dis}(P_{class})$ and $\mathbf{Dis}(P_{gifted})$ are the discontinuities in the probabilities of being placed in a gifted class and classified as gifted, respectively, and $\mathbf{Dis}(Q_{peer})$ is the discontinuity in peer quality. A similar model is relevant for high achievers using test score rank as the running variable and setting $D_{gifted} = P_{gifted} = 0$.

In our empirical analysis we estimate reduced form discontinuities in test scores at the eligibility threshold, which we interpret as estimates of **Dis**(y). For purposes of comparing across samples and with the broader education literature we also present 2SLS models which re-scale the reduced form effect by the estimated discontinuity in the probability of being placed in a gifted class, providing estimates of the ratio:

$$\mathbf{Dis}(y) / \mathbf{Dis}(P_{class}) = \beta_1 + \beta_2 \mathbf{Dis}(Q_{peer}) / \mathbf{Dis}(P_{class}) + \beta_3 [\mathbf{Dis}(P_{gifted}) / \mathbf{Dis}(P_{class})].$$

The left hand side is the "treatment on the treated" effect assuming that treatment operates through assignment to a gifted class. When $\beta_2 = \beta_3 = 0$, this gives the causal effect of being placed in a gifted class, β_1. When $\beta_3 = 0$ but $\beta_2 \neq 0$, the rescaled reduced form effect also includes a peer effect component, consisting of the treatment on the treated effect on peer quality for those who move to a gifted class ($\mathbf{Dis}(Q_{peer})/\mathbf{Dis}(P_{class})$), multiplied by the causal effect of peer quality β_2. To judge the likely importance of peer effects in the combined treatment effect we will examine differences in

Dis(Q_{peer})/**Dis**(P_{class}) when we compare treatment effects across different subsamples.

When $\beta_3 \neq 0$ the rescaled reduced form effect includes a third component attributable to the effect of individualized gifted services. All else equal, a positive value of β_3 implies larger (more positive) total treatment effects for gifted students than for non-gifted high achievers. Because our results suggest the opposite is true—and that the treatment effect estimates are mostly close to zero for gifted students—we mainly ignore this channel.[17]

Equation (2) also shows the potential threat to an RD analysis if there is non-random manipulation of measured IQ scores. If, for example, the true ability of students whose IQ scores are boosted to just above the gifted threshold is greater than that of students whose scores are not boosted, $E[IQ^*|IQ]$ will exhibit a positive discontinuity at IQ=T, implying an upward bias in the reduced form discontinuity. We discuss below three extensions that potentially limit such biases. One is to use the change in test scores from third to fourth grade as the dependent variable. This eliminates the effect of any variable that affects third and fourth grade scores equally. The second is to estimate the model after removing observations just above and just below the threshold—a so-called donut-hole specification. Estimates from this specification are robust to manipulation but rely on the assumed linearity of the relationship between IQ and test scores. The third is to combine these two and estimate a differenced donut-hole model.

b. Plan A Students

The reduced-form relationships between IQ and fourth grade test scores in reading, math, and writing for our Plan A analysis sample are shown in Figures 7a-c. As in Figures 4-6, we show the mean outcomes for each IQ point and the fitted relationship from local linear regression models. The graphs suggest an upward-sloping relationship between IQ and average test scores, with small negative

[17] Individualized gifted services may account for the positive effect of gifted status on the writing achievement of Plan B students we see below, and for the effect of gifted status on self-reported satisfaction of Plan A students.

discontinuities at the gifted threshold for reading and math a small positive effect for writing.

Table 2 presents the estimated discontinuities from a variety of alternative RD specifications. (We defer for the moment the issue of bandwidth selection, and focus on results from models that use 10 points to the left of the IQ threshold and 11 points to the right). Each column corresponds to a different dependent variable, whereas each row corresponds to a different specification of the RD model, including models with no added controls (row 1), models with a broad set of controls including school dummies (row 2), donut-hole specifications that exclude data within 2 points of the threshold (row 3), and first differenced models (rows 4 and 5). Columns 1 and 2 take as dependent variables the baseline third grade test scores in reading and math. The next three columns present first stage models for the event of being placed in a gifted classroom in fourth grade (col. 3), being classified as gifted by fourth grade (col. 4), and the average baseline test scores of the classroom peers in fourth grade (col. 5). Finally, columns 6, 7, and 8 present reduced-form models for standardized scores in fourth-grade reading, math, and writing tests, respectively.

The estimated discontinuities shown in columns 1-2 of Table 2 suggest that the relationship between the baseline test scores and IQ is relatively smooth at the 130 point threshold, with a very small jump in reading scores (0.02 to 0.03 standard deviation units) but a slightly larger jump in math scores (around 0.09 standard deviation units). The higher math scores for those above the threshold are consistent with the simple comparisons in Appendix Table 1 and suggest that there may be some positive bias in simple RD models for math achievement.

The estimated first stage discontinuities in the probabilities of placement in a gifted class, being classified as gifted, and in peer quality (columns 3-5) are all insensitive to the addition of controls for school dummies and student characteristics (compare rows 1 and 2), but become slightly larger when data close to the threshold are eliminated (row 3), as expected given the patterns in Figure 5.

The reduced form estimates for reading and math in row 1 are negative, and become more negative when student characteristics and school fixed effects are added (row 2), reflecting the positive discontinuities in the baseline achievement measures. Excluding observations near the 130 point threshold (row 3) leads to even more negative but relatively imprecise estimated impacts.

The first differenced (or "gain-score") models in row 4 of Table 2 use as a dependent variable the change in standardized scores in reading and math from third to fourth grade. (Writing is not tested in third grade we cannot estimate differenced models for this outcome). Assuming that equation (1) is correctly specified for both third and fourth grades and that $D_{class} = D_{gifted} = 0$ for all students in third grade, we can difference the model, obtaining:

(4) $Dy = y - y_0 = \beta_1 D_{class} + \beta_2 (Q_{peer} - Q_{peer3}) + \beta_3 D_{gifted} + \varepsilon - \varepsilon_0$,

where y_0 represents third grade achievement, Q_{peer3} is peer quality in third grade, and ε_0 is the value of the unobserved error component in the model for third grade achievement.[18] Provided that third grade peer quality (Q_{peer3}) varies smoothly with IQ and that there is no discontinuity in the expected **difference** of the transitory determinants of test scores, the reduced-form discontinuity in gain-scores is:

(5) $\textbf{Dis}(Dy) = \lim_{IQ\downarrow T} E[Dy|\ IQ] - \lim_{IQ\uparrow T} E[Dy|\ IQ] = \beta_1 \textbf{Dis}(P_{class}) + \beta_2 \textbf{Dis}(P_{gifted}) + \beta_3 \textbf{Dis}(Q_{peer})$.

Notice that the right hand side of (5) is the same as the right hand side of equation (3). Thus, if both the levels and differences models are correctly specified, they should yield similar reduced form estimates.

An important feature of test score gains is that they are approximately **orthogonal** to measured IQ. Figures 7d and 7e show the relationship for gains in reading and math scores for our Plan A sample, along with the fitted values from linear models fit separately to each side of the 130 point threshold. To provide additional leverage we have expanded the samples in these figures to include students with IQ scores from 115 to 145. Notice that, despite the strong relationship between the receipt of gifted services and IQ (documented in Figure 5), the relationship of test score gains to IQ is flat. This is useful

[18] For simplicity we ignore any time-varying effect of the student covariates in equation (4), but these are all included in our estimating models, as are school fixed effects. We also assume that individualized instruction has no effect on scores in grade 3, which will be true if most gifted students are identified during or after third grade.

because, as noted by Angrist and Rokkanen (2012), it allows us to compare test score gains for groups farther away from the 130 point threshold. Specifically, since the running variable in our RD model is orthogonal to test score gains, the local linear RD specification amounts to a simple difference-in-differences estimator which compares the average test score gains for students above the threshold (virtually all of whom are in gifted classes) to all those below the threshold (less than one-half of whom are in gifted classes).[19] To sidestep concerns about IQ manipulation we also implement a differenced donut-hole model, which is equivalent to a difference in differences of test scores for students with IQ's >131 relative to students with IQ's< 127.

Overall, the various estimators in Table 2 suggest that participation in gifted education has little or no effect on standardized test scores of Plan A students. In fact the point estimates from most of the specifications are negative, though we typically cannot reject small positive effects. We have conducted a wide range of robustness checks on these results, including RD models with alternative bandwidths, and models that use SAT scores written prior to receiving any gifted services to measure baseline achievement. The results from models with alternative bandwidth choices are summarized in Appendix Figure 2, where we show point estimates and standard error bands for the estimated impacts on reading, math and writing using the simple specification in row 1 of Table 2 with bandwidths from 5 to 15 points. For reading, the point estimates become more negative as the bandwidth is increased, with a marginally significant -0.1 estimate using bandwidths of 13 or larger. For math and writing the point estimates tend toward a value closer to 0 as the bandwidth is increased, but remain uniformly negative.

Results from models that use SAT scores measured before a student is classified as gifted as a baseline measure of achievement are presented in Appendix Table 2. These results are quite similar to the results in Table 2, though typically less precise, reflecting the smaller sample sizes. We have also examined effects on fifth grade student outcomes, capturing the effects of two years of exposure to a

[19] As a check we re-estimated these models excluding the running variable terms. The estimated effects are very similar to those reported in the table, but a little more precise.

gifted classroom. These effects (reported in Appendix Table 3) are also small in magnitude and mostly negative. Finally, we have examined RD models for various subsamples of students, based on school characteristics such as the fraction of FRL participants or average test scores in the school as a whole. The results are somewhat imprecise, but we find no evidence of positive effects for any subsample.

One potential explanation for the small (or even negative) treatment effects in our RD models is that marginally eligible Plan A students are mismatched to the gifted classroom environment, and fail to benefit from their placement despite the better peer group.[20] Examination of Figures 7d and 7e, however, shows that marginally eligible Plan A students with IQ's of 130-131 experience about the same achievement gains as infra-marginal students with higher IQ's. In fact, across the entire range of IQ's test score gains are very similar, leading the donut-hole first differenced models to yield estimated effects that are close to 0. In addition, data from student satisfaction surveys conducted by the District (see Appendix A) suggest that that marginally eligible Plan A students (or their parents) are happy with their placement—a result that does not seem to indicate a mismatch problem.

Another concern is that Plan A students are "topped out" on the state-wide achievement tests. Since only 2% of Plan A students with IQ sores in the range of 125 to 135 points have the maximum possible reading score, this is not a major concern for reading achievement. For math the issue is potentially more important because about 15% of students with IQ's in the 125-135 range achieve the maximum score. Nevertheless, topping out would be expected to lead to attenuate the RD estimates toward zero by a percentage proportional to the censoring rate (e.g., Goldberger, 1972; Greene, 1981). Since all but one of the reduced form effects in Table 2 are **negative**, this suggests the true effects are, if anything, even more negative. As a check we fit Tobit models for reading and math using the same specification as in row 1 of Table 2 and obtained estimates that are very close to the OLS estimates. Overall we believe that topping out cannot account for the absence of impacts on Plan A students.

[20] Bui et al. (2011) hypothesis that marginally gifted students in their RD designs suffer an "invidious comparison" effect and end up under-performing in gifted classrooms.

c. Plan B Students

Figure 8 shows the reduced form relationships between IQ and fourth grade test scores for our Plan B analysis sample. Mean reading and math scores evolve relatively smoothly through the gifted threshold at 116 points and provide little evidence of an effect of gifted services on these outcomes. Writing scores, however, exhibit a jump at the gifted threshold, suggesting a potentially positive impact.

Table 3 presents estimated discontinuities for various outcomes and model specifications, following the same format as Table 2. Notice that the estimated discontinuities in baseline reading and math scores are small and insignificant, confirming the visual impression from panels c and d of Figure 4. The first stage discontinuities in the probabilities of being placed in a gifted class or being classified as gifted (columns 3-4) and in classroom peers' previous scores (column 5) are relatively precisely estimated and very similar whether or not we control for student characteristics and school dummies (compare rows 1 and 2). The donut-hole specification (row 3) yields first stage estimates that are a little bigger in magnitude, consistent with the evidence in Figure 5b and 6 that the first stage outcomes for Plan B students with IQ's just below the gifted threshold are higher than expected. Removing these observations therefore tends to widen the estimated discontinuities in the first stage outcomes.

As suggested by Figures 8a and 8b, the reduced-form RD estimates for fourth grade reading and math in columns 6 and 7 of Table 3 are uniformly small in magnitude, regardless of whether we control for school effects and student characteristics or exclude observations within 2 points of the gifted threshold. In Appendix Figure 3 we explore the robustness of these basic RD results to alternative bandwidth choices. Across a range of bandwidths from 5 to 15 points, the estimated impacts on reading and math are uniformly close to zero.

Rows 4 and 5 of Table 3 show estimates from differenced models that use the change in test scores from third to fourth grade as the dependent variable. For additional leverage we expand the sample for these models to include students with IQ scores from 100 to 130 points. Figures 8d and 8e confirm that, as we saw for Plan A students, there is virtually no relationship between test score growth and IQ. Since the running variable in the gain score models is orthogonal to the outcome, the differenced and differenced donut-hole models are equivalent to difference-in-differences specifications. The latter model, for example, compares test score gains for students with IQ's >117 (three quarters of whom are assigned to gifted classes) to the gains for students with IQ's <114 (less than 20% of whom are in gifted classes), and confirms that there is no average treatment effect for a relatively large group of Plan B students.[21]

In contrast, the results in column 8 of Table 3 show reduced form impacts on writing scores that are relatively large in magnitude (0.11 to 0.12 standard deviation units), stable across specifications, and at least marginally significant. As shown in Appendix Figure 3, the point estimates for writing impacts from the simple model in row 1 are also quite stable across alternative bandwidth choices. Our most precise specification in row 2 yields a t-statistic of 2.5 for the test of no effect of gifted services. Scaling the reduced form effect on writing scores by the first stage effect on being assigned to a gifted class (from column 3) implies a treatment effect per Plan B student who is assigned into a gifted classroom of about 0.3 standard deviation units—a relatively large effect.[22]

As with our Plan A sample, we have explored a number of robustness checks to verify the results in Table 3. Results from alternative specifications that use SAT scores measured before a student is classified as gifted as a baseline measure of achievement are presented in Appendix Table 2. These results are quite similar to the results in Table 3. Estimated impacts on fifth grade student outcomes, reported in Appendix Table 3, are also small in magnitude and uniformly insignificant.

We have also examined RD models for various subsamples of students. We find small impacts

[21] We also estimated the differenced models excluding the terms in the running variable. The resulting impact estimates are very similar to the estimates shown in the table, but more precise. E.g., for the donut differenced model the estimates (and standard errors) are -0.004 (0.022) and 0.022 (0.022) for reading and math, respectively.

[22] If we assume that the writing score gains are due mainly to the extra individualized services that gifted students receive (i.e. that $\beta_1 = \beta_2 = 0$) and we scale by the first stage effect of being classified as gifted (column 4) then the implied treatment effect is about 0.2 σ's.

on reading and math scores for all subgroups (including boys and girls, whites versus minorities, and students at schools with higher and lower fractions of FRL participants). We find that the impact on writing scores is largely concentrated among boys and students at schools with a high-FRL fraction.

d. High Achievers

Figure 9 plots the outcomes of students in our high achiever analysis sample by their relative rank on the previous year's tests within their cohort and school. The plots for fourth grade reading and math scores show clear jumps at the threshold for admission to the gifted classroom. In contrast, the plot for fourth grade writing shows little evidence of any discontinuity. Panels d and e show the average test score **gains** in reading and math. In contrast to the differenced models for the two gifted groups, test score gains of high achievers from third to fourth grade are strongly negatively correlated with their rank on the third grade tests, reflecting the fact that the top scorers in any one year have on average benefited from positive measurement errors (or "good luck") that is not expected to persist to the next year. Nevertheless, the differenced models also show clear evidence of discontinuities at the threshold for admission to the gifted class.

Table 4a shows the estimated discontinuities in our main outcome variables from alternative model specifications, using the same format as Tables 2 and 3. The results for the baseline test scores in columns 1 and 2 show only small changes in reading and math scores at the threshold rank. The first stage models for the high achievers in columns 3 and 4 of Table 4a are precisely estimated and robust to the addition of controls. We find a relatively large (32 pp) discontinuity in the fraction of high-achievers placed in the gifted class at the cutoff, and a slightly smaller discontinuity (0.28 σ's) in mean lagged achievement scores of their classroom peers.

The estimated reduced-form discontinuities in reading and math are also relatively precisely estimated and robust across alternative specifications, with a magnitude of 0.07 to 0.10 σ's, while the effects on writing are close to zero. Appendix Figure 4 shows the robustness of these results to alternative bandwidth choices, using the simple specification in row 1 of Table 4a. Across a range of bandwidths from 5 to 15, the estimated impacts on reading and math are quite stable and statistically significant at conventional levels, while the estimated impacts on writing are uniformly small. Taking account of the first stage effect, the reduced form impacts on reading and math achievement imply treatment-on-the-treated effects of being placed in a gifted classroom of roughly 0.3 σ's.

To what extent do the achievement gains of high achievers in fourth grade persist to later years? Table 4b presents a series of RD models focusing on fifth grade outcomes: in all cases the running variable is the student's rank in **third grade** achievement tests. Graphs of the underlying relationships are presented in Figure 10. We begin in column 1 with a model for the probability of remaining in the District until fifth grade. The estimated effects of being ranked above the threshold for placement in a fourth grade gifted class are very small, showing no evidence of differential attrition. Column 2 presents a model for the probability of placement in a gifted class in fifth grade. The estimated RD impact is 7 pp, about one-fifth as big as the jump in the probability of placement in a fourth grade class. Since placement in fifth grade depends on *fourth grade* scores, this discontinuity is not purely mechanical and is presumably driven in part by the achievement gains experienced by students who attended a gifted class in fourth grade. Columns 3 and 4 present models for baseline (third grade) achievement of the subsample of our high-achiever analysis sample who can be followed to fifth grade. These models show a small negative discontinuity in third grade math scores, similar to the results for the overall sample (column 2 of Table 4a) but a positive discontinuity in third grade reading.

While not reported in the Table, the effects of participating in a fourth grade gifted class on *fourth grade* reading and math are about the same in the subsample that can be followed to fifth grade and our overall high achiever sample. As shown in panels c and d of Figure 10, both outcomes exhibit a clear jump at the cutoff rank for admission to the class. The reduced form effects on *fifth grade* reading scores are relatively small (see panel e of Figure 10 for a graphical representation), and suggest that the achievement gains in reading registered at the end of fourth grade fade relatively quickly.[23] In contrast, the estimated effects on fifth grade math scores are about the same size as the effects on fourth grade scores, and are either marginally significant or significant (see panel f of Figure 10).

Perhaps most interesting is the effect on fifth grade science achievement, which is about the same size as the effect on fourth grade math, and (in the richer specification in row 2) significantly different from zero at conventional levels (see panel g of Figure 10). Taken together with the results for fifth grade math we conclude that an important share of the extra learning that high achievers experienced in fourth grade persists to later years.

Table 5 explores the heterogeneity in impacts for different subgroups of high achievers, using the specification with controls for student characteristics (including lagged test scores) and school dummies in row 2 of Table 4a. We present results for subgroups based on a student's FRL status and race, as well as contrasts across schools based on the school-wide fraction of FRL students and the number of gifted children in the gifted classroom.

Column 1 of the table shows the estimated first stage effects, which vary only slightly across subsamples. Columns 2 and 3 report two stage least squares estimates for the effect of placement in a gifted classroom on average peer quality (as measured by mean third grade scores in math and reading). There is some variation in the size of the peer effects across subgroups, with larger gains in peer quality for non-FRL eligible students and white non-Hispanics, and smaller gains for FRL eligible students and minorities.

[23] The reduced form effects on fifth grade outcomes of high achievers represent a combination of persistent effects from any gains made in fourth grade, and the treatment effect of being assigned to a gifted class in fifth grade. A plausible benchmark model for the discontinuity in fifth grade scores is $\mathbf{Dis}(y_5)=\delta\beta_1\mathbf{Dis}(P_{class4})+\beta'_1\mathbf{Dis}(P_{class5})$, where δ is the fraction of the effect in fourth grade scores (β_1) that persists to the next year's scores, β'_1 is the direct effect of placement in a gifted class in fifth grade on fifth grade scores, and $\mathbf{Dis}(P_{class5})$ is the discontinuity in the probability in placement in a fifth grade gifted class at the cutoff for admission to the fourth grade class. This suggests that the expected reduced form effect on fifth grade scores, as a fraction of the effect on fourth grade scores, is $\delta+\beta'_1\mathbf{Dis}(P_{class5})/\ \beta_1\mathbf{Dis}(P_{class4}) \approx \delta+0.2\ \beta'_1/\beta_1$.

Columns 4-8 report 2SLS estimates for fourth grade achievement outcomes, while columns 9-11 report corresponding estimates for fifth grade outcomes (in both cases using ranks on third grade test scores as the running variable). The estimated treatment effects on fourth grade reading and math are particularly large for FRL-eligible students and minorities -- in the range of 0.4-0.6 σ's. These effects are large enough to close the corresponding achievement gaps between FRL-eligible and ineligible high achievers, or between minority and white high achievers. Interestingly, both groups also experience relatively large impacts on fifth grade science. Comparisons based on school-wide FRL rates are less systematic, with larger gains in math for students at low-FRL schools and larger gains in reading at high-FRL schools.[24]

The contrast in rows 2 and 3 between the estimated effects for FRL and non-FRL high achievers is worth noting because (as shown in column 2) FRL-eligible high achievers experience a smaller average gain in peer quality when they enter a gifted classroom (0.75 σ's) than FRL-ineligible students (0.98 σ's). Nevertheless, FRL-eligibles gain more, casting some doubt on the importance of peer effects as a mediator for the impact of the gifted classroom experience.

e. An Alternative Design for High Achievers

A potential concern with the policy of placing gifted and high-achieving students in a separate classroom is that moving these children can affect the scores of *other* children, leading to possible biases in our RD procedure. If, for example, removing the most able students has a negative effect on the highest-ability children who remain in the class, then our RD estimates could overstate the causal effect of gifted placement on high ability children. The District's policy of offering a gifted classroom for fourth graders if and only if there is at least one gifted child in the fourth grade cohort provides a design for

[24] We also examined differences by gender using both our RD approach and the alternative between school design (explained below) that allows us to examine infra-marginal impacts. These results, available on request, show larger gains for marginally eligible boys than girls, but the opposite pattern for infra-marginal girls and boys.

checking these impacts.[25] It also provides an opportunity to compare the treatment effects measured in our RD design to the average treatment effects for wider groups of students.

Specifically, consider the top 20 fourth grade students (ranked by third grade test scores) at schools with either 0 or a small number (e.g., 1-4) of gifted students in the fourth grade cohort. At schools with no gifted children these students will be assigned to regular classrooms, whereas at the schools with a small number of gifted children the top 20 high achievers have a high probability of assignment to a separate gifted classroom. We can therefore compare how the average test score gains of the top 20 groups vary as the number of gifted children varies, looking for a discrete jump when the number is 0. We can also examine the test score gains of students ranked 25-44 in the same schools. These students should remain in regular classes regardless of whether there is a gifted classroom or not, so any effect of the presence of a gifted child on these students is due to a spillover effect.

To implement this design we identified a set of fourth grade classes at elementary schools in the District with from 0 to 4 gifted children in the entire cohort.[26] We then identified the students ranked from 1 to 20 and from 25 to 44 on the previous year's achievement scores in each class, and constructed the mean fractions of each group who were placed in a gifted classroom in fourth grade, and the average changes in math and reading test scores between third and fourth grades.

The first stage relationship for this alternative design and the corresponding reduced form effects are illustrated in Figure 11. Panel a shows the fraction of the top 20 students in a school/grade cohort who were placed in a gifted class for school/grade cohorts with 0, 1, 2, 3, or 4 gifted children. The relationship is very close to linear for schools with 1 or more gifted students, with an intercept of around 30%. Since high achievers at schools with 0 gifted children have no chance of placement in a gifted classroom, the implied effect of having at least 1 gifted child in the cohort is about 30 pp. Panel b shows

[25] We are grateful to Kelley Bedard for a suggestion that motivated this section.

[26] This sample comprises 255 fourth grade classes at 94 elementary schools during the years 2009-12 (for which we know the District's ranking formula) and includes roughly half of all the fourth grade classes during these years.

the corresponding relationship for students ranked 25-44 in each school/grade cohort. This relationship is also nearly linear for school/cohorts with 1-4 gifted children, but at a much lower level, reflecting the low probability that students ranked in this range will be placed in a gifted classroom, even if one is present. The implied effect of having at least 1 gifted child is only about 5%. The <100% impact on the 1-20 group and the positive impact on the 25-44 group are due to our difficulties (noted earlier) in replicating the rankings used by the schools to fill their gifted classes. Despite this slippage, the presence of a gifted child has relatively large impact on the probability of placement in a gifted classroom for the top-ranked group and only a small effect for their lower-ranked classmates.

The corresponding reduced form impacts on the test score gains of the two groups are illustrated in panels c and d. (We combine math and reading to maximize our power, and show regression-adjusted impacts that control for class-wide average reading and math scores in third grade, the class-wide fraction of FRL students, and class size). For the 1-20 ranked group, the pattern in panel c suggests a reduced form impact of around 0.1 σ's. The reduced form impact on the 25-44 ranked group, by comparison, is very close to 0.

Table 6 presents a series of models for the mean outcomes of the two groups of students. In each case we show the coefficient of a dummy for having at least one gifted child in the (school-wide) class. Columns 1 and 4 present models for third grade (i.e., pre-intervention) average test scores in reading and math. Columns 2 and 5 present models for the fraction of the group that is placed in a gifted classroom. Finally columns 3 and 6 show models for the group-wide average change in the average of reading and math test scores between third and fourth grades. Row 1 presents a baseline specification with only a linear control for the number of gifted students in the cohort. Row 2 presents specifications that include class-level controls for third grade test scores, the fraction of FRL students and class size, as were used in Figure 11. Finally, we present results based on a subsample of schools and classes with either zero or one gifted student in the class in row 3.

Looking at the baseline (third grade) test score results in columns 1 and 5, notice that while average student achievement in either the 1-20 rank group or the 25-44 rank group is positively correlated with the presence of at least one gifted child in the cohort (row 1), this correlation is eliminated by the cohort and school-level controls (row 2). The first stage models for the probability of assignment to a gifted classroom (columns 2 and 5) show that for the 1-20 rank group, the presence of a gifted child raises the probability of being assigned by 33-35 pp, whereas for the 24-44 rank group the corresponding effect is 6-7 pp.

The reduced form models for test score gains of the 1-20 ranked group in column 3 show a positive effect of having at least one gifted student in the class of 0.06 to 0.09 σ's—not too different from the reduced form effects in Table 4a for our overall high achiever sample. As shown in Table 5 (row 9), however, the implied effects from our RD models are actually a little larger for high achievers in school/grade cohorts with 1-4 gifted children. Relative to this benchmark, the reduced form effects in column 3 imply somewhat smaller treatment on the treated effects (e.g., a 0.27 treatment on the treated effect from the estimate in row 2 of Table 6 versus 0.415 for reading and math from row 9 of Table 5). The estimated effects for the 24-44 ranked group, by comparison, are very close to zero, suggesting that the placement of the higher ranked students in gifted classes has no effect on the lower ranked students (i.e., no spillover effects).

We can also use the between school design to examine potential differences in the in the effect of placement in a gifted class for students who are closer or further from the threshold in our RD models. In Table 7 we classify students in the 1-20 rank group into four subgroups: 1-5, 6-10, 11-15, and 16-20, and estimate separate models for the effect of the presence of at least one gifted child on each subgroup. Since there are generally 19-24 students in a class, the "marginally eligible" students whose treatment effects are identified by our RD models are those in the 11-15 and 16-20 rank groups, while the higher-ranked groups are infra-marginal.

Column 1 shows the average third grade achievement levels of the various student groups. On average, students ranked in the top 20 have combined reading and math scores of about 0.8 σ's -- not too different from the average across all high achiever groups in our RD-based analysis. Students ranked in the top 5 for their school-cohort have much higher achievement (1.3σ's) -- about the same as the Plan A analysis group -- whereas those ranked 16-20 have lower scores (around 0.4 σ's).[27]

The first stage models in column 2 show that students in the highest ranked group (1-5) are the most likely to move into a gifted class when one is available (42 pp effect), whereas those in the lowest ranked group have a smaller effect (21 pp), reflecting the higher misclassification probability for students nearest the cutoff. In contrast, the reduced form impacts in column 3 are larger for the lower ranked groups. As a result, the 2SLS (treatment on the treated) estimates in column 4 show a pattern of increasing effects for lower-ranked groups. Indeed, the 2SLS estimates of 0.39 and 0.50 for the two marginally eligible rank groups bracket the RD-based 2SLS estimates for schools with 1-4 gifted students (Table 5, row 9).

The variation in the estimated treatment effects across the rank groups in Table 7 is too large to be explained by the (minor) variation in the fractions of low-income and minority children across the groups.[28] It is also not explained by heterogeneity across schools or teachers, since the subgroups are balanced across classes. Instead, we suspect that the variation reflects systematically larger treatment effects for students with lower initial levels of achievement, who have the most to gain.

VI. Interpreting the Differences across Groups

We find that participation in separate classrooms has no little or no effect on the standardized achievement scores of the gifted students these classes were originally designed to serve -- even Plan A

[27] The fractions of FRL-eligible and minority students are relatively similar across rank groups, at about 70% each.

[28] The fraction of black and Hispanic students in the lowest rank group (16-20) is about 8 pp higher than the fraction in the highest group. Assuming the treatment on the treated effect on white non-Hispanic students is 0 and the effect on black and Hispanic children is 0.5 σ's, this explains only a 0.04 σ rise in treatment effects.

students who meet the traditional IQ threshold for gifted education. Paradoxically, however, these classes have relatively large effects on non-gifted high achievers. One explanation for the small impact on Plan A students is that statewide test are the wrong metric for this group. Since most Plan A students already score in the top percentiles of the state tests, parents and teachers may be less concerned with mastering the basic curriculum and more concerned with higher-level learning objectives (and in simply keeping Plan A students engaged in school). This explanation is consistent with the presence of a positive jump in student satisfaction at the 130 IQ bar for Plan A students, discussed in Appendix A, and with the fact that many Plan A parents are willing to pay for private IQ testing to get their children into the gifted program. A related explanation is that it is very hard to raise the scores of students who are already performing very well - a "flat of the curve" effect that may also explain the variation in treatment effects across the different rank-groups in Table 7 and the differential treatment effects for FRL-eligible and ineligible high achievers in Table 6. Similar arguments could explain the lack of impacts in several other recent studies of tracking programs for high-ability students, including Bui et al.'s (2011) analysis of gifted programs, and the findings of Abdulkadiroglu et al. (2011) and Dobbie and Fryer (2011) on elite exam-based public high schools in NYC and Boston.[29]

These arguments are less compelling for Plan B gifted students, whose test scores prior to entering gifted classrooms are closer to the disadvantaged high achievers. A fundamental difference between Plan B students and high achievers is the way they are selected -- in one case by IQ, in the other by relative achievement on standardized tests. Students selected on the basis of previous test scores have a combination of cognitive abilities **and** non-cognitive traits like attention-to-task and willingness to meet social expectations that are necessary to perform well on low-stakes tests of routine knowledge. Students selected entirely on the basis of IQ may lack these traits. This "non-cognitive deficit" explanation is consistent with the view expressed in personal interviews with a number of gifted

[29] In a later version of their paper, Dobbie and Fryer (2013) show that attending an exam school also has little or no effect on college enrollment, even at relatively elite colleges and universities.

teachers in the District who reported a preference for teaching high achievers over Plan B gifted students. To the extent that non-cognitive traits are even more important in a gifted classroom environment, it may also help to explain the fact that Plan B students appear to experience a fall in satisfaction from entering gifted classes, contrary to the rise for Plan A students and the zero effect for high achievers (see Appendix A).

Examination of the characteristics of the compliers in our RD design for the Plan B group confirms that students whose placement in a gifted classroom depends on whether their IQ is above or below the 116 point threshold are "underachievers" whose standardized test scores are low relative to their cognitive ability.[30] Specifically, as shown in Appendix B, Plan B compliers score roughly 1.6 σ's above the mean on the NNAT test (a measure of cognitive ability) but only 0.7 σ's above the mean in third-grade reading and math achievement. By comparison, the compliers in our RD analysis of disadvantaged high achievers have NNAT scores and standardized achievement scores that are both around 0.7 σ's above the mean.

VII. Conclusions

As in many other areas of education policy, a fundamental issue for gifted education is the question of who is allowed access to the program. Traditionally, gifted education was targeted to students with very high cognitive ability, often using IQ as a screen. Since the 1970s, many researchers have argued for a more holistic admission standard that recognizes both cognitive and non-cognitive skills. While some states and school districts have clung to the traditional IQ-based criteria, others have adopted multi-dimensional formulas. Most often, however, even these systems impose the same thresholds across richer and poorer schools, leading to the systematic under-representation of minority and economically disadvantaged students in gifted and talented programs.

[30] We use the method described by Walters (2014) to estimate the mean characteristics of compliers. See Appendix B for details.

The gifted program for fourth and fifth grade students in the District provides a unique opportunity to address these issues, since the program serves some students who are selected on IQ scores and others who are selected on past achievement *in the same classrooms*. The program creates a separate gifted classroom whenever there is at least one gifted child (meeting an absolute IQ standard) in the school. As a result, most schools -- even those with very high fractions of free-lunch and minority students -- have separate gifted classrooms. The majority of seats in these classes are filled by *non-gifted* students who scored highest on state-wide achievement tests in the past year. For poorer schools with only a handful of gifted students in any grade the program is similar to a tracking program that sets aside a separate class for the 20 or so top performing students in the grade/cohort.

Our findings suggest that the program has no effects on the reading or math achievement of non-disadvantaged gifted students, who must meet a 130 point IQ standard to achieve traditional gifted status. Nor does it affect the reading or math scores of disadvantaged gifted students (free/reduced price lunch participants and English language learners) who face a lower 116 point IQ threshold for "Plan B" gifted status. However, it has positive and relatively large effects on the achievement of the non-gifted high achievers who fill the remaining seats in the class, concentrated among free/reduced price lunch participants and black and Hispanic students.

We reach similar conclusions about the impact of the program using a regression discontinuity approach (based on the students' ranks on previous achievement tests) and using an alternative between-class design that compares achievement gains for top-ranked students at schools where there are no gifted children in fourth grade (and no gifted class) to similarly-ranked students with 1, 2, 3, or 4 gifted children in the grade/cohort. The latter design allows us to verify that there are no spillover effects on the "second tier" of students who narrowly miss the cutoffs for placement in a gifted classroom (when one is present). It also allows us to compare impacts on groups of students who are close to the threshold for admission (and whose outcomes drive our RD estimates) to impacts for infra-marginal students. The gains appear to be largest for students who are closest to the threshold and furthest away from the traditional thresholds that determine entry to gifted programs in other settings.

Overall, we conclude that a separate gifted classroom environment can be highly effective in raising the standardized test scores of students selected on the basis of past achievement, particularly disadvantaged and minority students who would normally not qualify for gifted education programs that use an absolute admission standard. Our findings suggest that a comprehensive tracking program that establishes a separate classroom in every school for the top-performing students could significantly boost the performance of the most talented students in even the poorest neighborhoods, at little or no cost to other students or the District's budget.

References

Abdulkadiroglu, Atila, Joshua D. Angrist and Parag A. Pathak. “The Elite Illusion: Achievement Effects at Boston and New York Exam Schools.” *Econometrica* 82(1): 137-196.

Angrist, Joshua D. and Miikka Rokkanen. "Wanna Get Away? RD Identification Away from the Cutoff." NBER Working Paper No. 18662. December 2012.

Bajari, Patrick, Han Hong, Minjung Park and Robert Town. “Regression Discontinuity Designs with an Endogenous Forcing Variable and an Application to Contracting in Health. NBER Working Paper 17643, December 2011.

Barreca, Alan, Jason M. Lindo and Glen R. Waddell. “Heaping Induced Biases in Regression Discontinuity Designs. NBER Working Paper 17408, September 2011.

Betts, Julian R. “The Economics of Tracking in Education.” Education.” In Hanushek, Eric A., Stephen Machin and Ludger Woessmann (eds.) *Handbook of the Economics of Education*, Volume 3, Amsterdam: North Holland, pp. 341-381.

Bui, Sa A., Steven G. Craig, and Scott A. Imberman. “Is Gifted Education a Bright Idea? Assessing the Impact of Gifted and Talented Progams on Achievement.” NBER Working Paper 17089, May 2011.

Chapman, P. D. *Schools as Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890– 1930.* New York: New York University Press, 1988.

Davis, Bilie, John Engberg, Dennis N. Epple and Ron Zimmer. “Evaluating the Gifted Program of an Urban District Using a Modified Regression Discontinuity Design.” NBER Working Paper 16414, September 2010.

Dobbie, Will and Roland G. Fryer Jr. “Exam High Schools and Academic Achievement: Evidence from New York City.” NBER Working Paper 17286, August 2011.

Duflo Esther, Pascaline Dupas, and Michael Kremer. “Peer Effects, Teacher Incentives and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya.” *American Economic Review* 101 (August 2011), pp. 1739-1174.

Florida Department of Education. “Standard Error of Measurement”. Technical Assistance Paper Number FY 1996-7. Available at http://www.fldoe.org/bii/Gifted_Ed/.

Goldberger, Arthur S. “Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations,” Institute for Research on Poverty Discussion Paper 123-72, 1972.

Greene, William. “On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model.” *Econometrica* 49 (1981), pp. 505–513

Jackson, C. Kirabo. “Do Students Benefit from Attending Better Schools? Evidence from Rule-based Student Assignments in Trinidad and Tobago.” *Economic Journal*, 120(549) (2010), pp. 1399-1429.

Jolly, Jennifer L. “A Resuscitation of Gifted Education.” *American Educational History Journal* 36(1) (2009), pp. 37-52.

Lee, David S. “Randomized Experiments from Non-random Selection in U.S. House Elections,” *Journal of Econometrics* 142(2) (February 2008), pp 675–697.

Lohman, David F, Katrina A. Korb and Joni Lakin. “Identifying Academically Gifted English-Language Learners Using Nonverbal Tests : A Comparison of the Raven, NNAT, and CogAT.” *Gifted Child Quarterly* 52(4) (Fall 2008), pp. 275-296.

McClain, Mary-Caherine and Seven Pfeiffer. “Identification of Gifted Students in the United States Today: A Look at State Definitions, Policies and Practices.” *Journal of Applied School Psychology* 28(1) (2012), pp. 59-88.

McCrary, Justin. “Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test.” *Journal of Econometrics* 142(2) (2008), pp. 698–714..

Pop-Eleches , Cristian and Miguel Urquiola. “Going to a Better School: Effects and Behavioral Responses.” *American Economic Review* 103(4) (June 2013), pp. 1289-1324.

Renzulli, Joseph S. (1978). “What Makes Giftedness? Re-examining a Definition.” *Phi Delta Kappan* 60(3), pp. 180-184.

Slavin, Robert E. “Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis.” *Review of Educational Research* 57(3) (Fall 1987), pp. 293-336.

U.S. Department of Education, Office of Educational Research and Improvement. *National Excellence: A Case for Developing America's Talent*. Washington, DC: U.S. Government Printing Office, 1993.

VanTassel-Baska, Joyce. “The History of Urban Gifted Education." *Gifted Child Today* 33(4) (Fall 2010), pp.18-27.

Vardardottir, Arna. “Peer Effects and Academic Achievement.” *Economics of Education Review* (2013) 36: 108–121.

Walters, Christopher. "The Demand for Effective Charter Schools." Unpublished Working Paper, UC Berkeley Department of Economics, 2014.

Figure 1. Fraction of third graders who are classified as gifted by the end of the school year

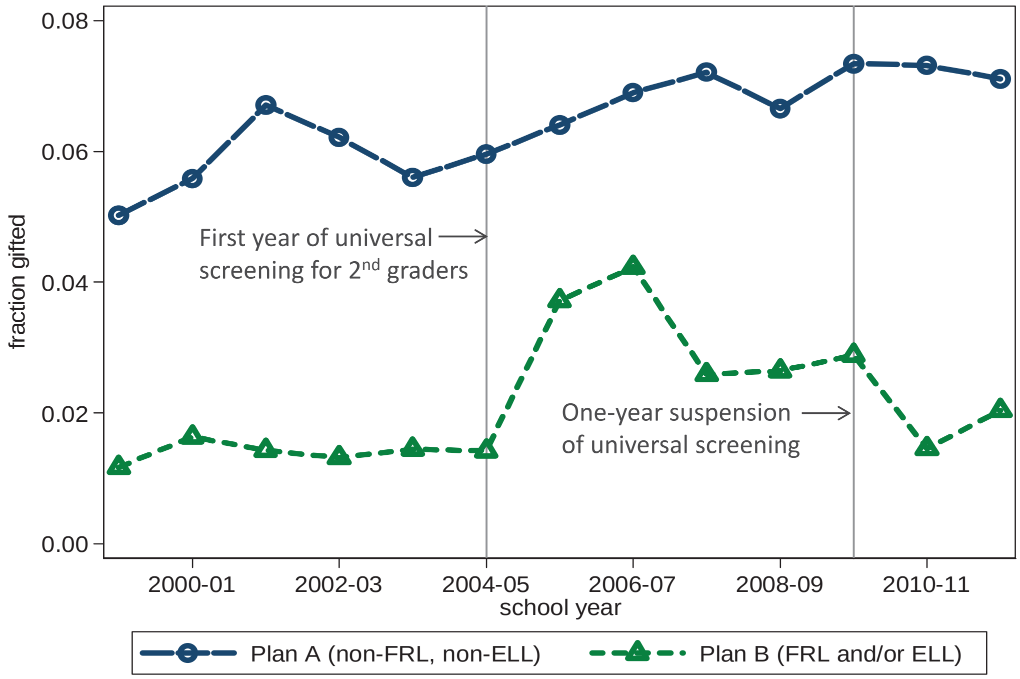


Figure 2. Composition of fourth grade gifted classrooms by %FRL in school

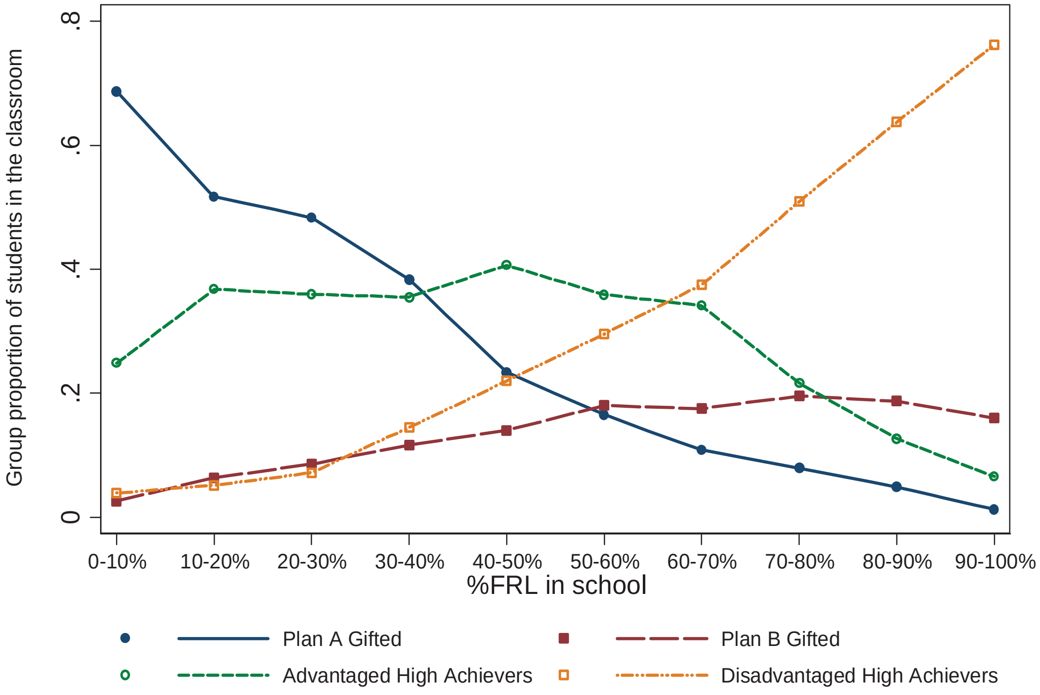
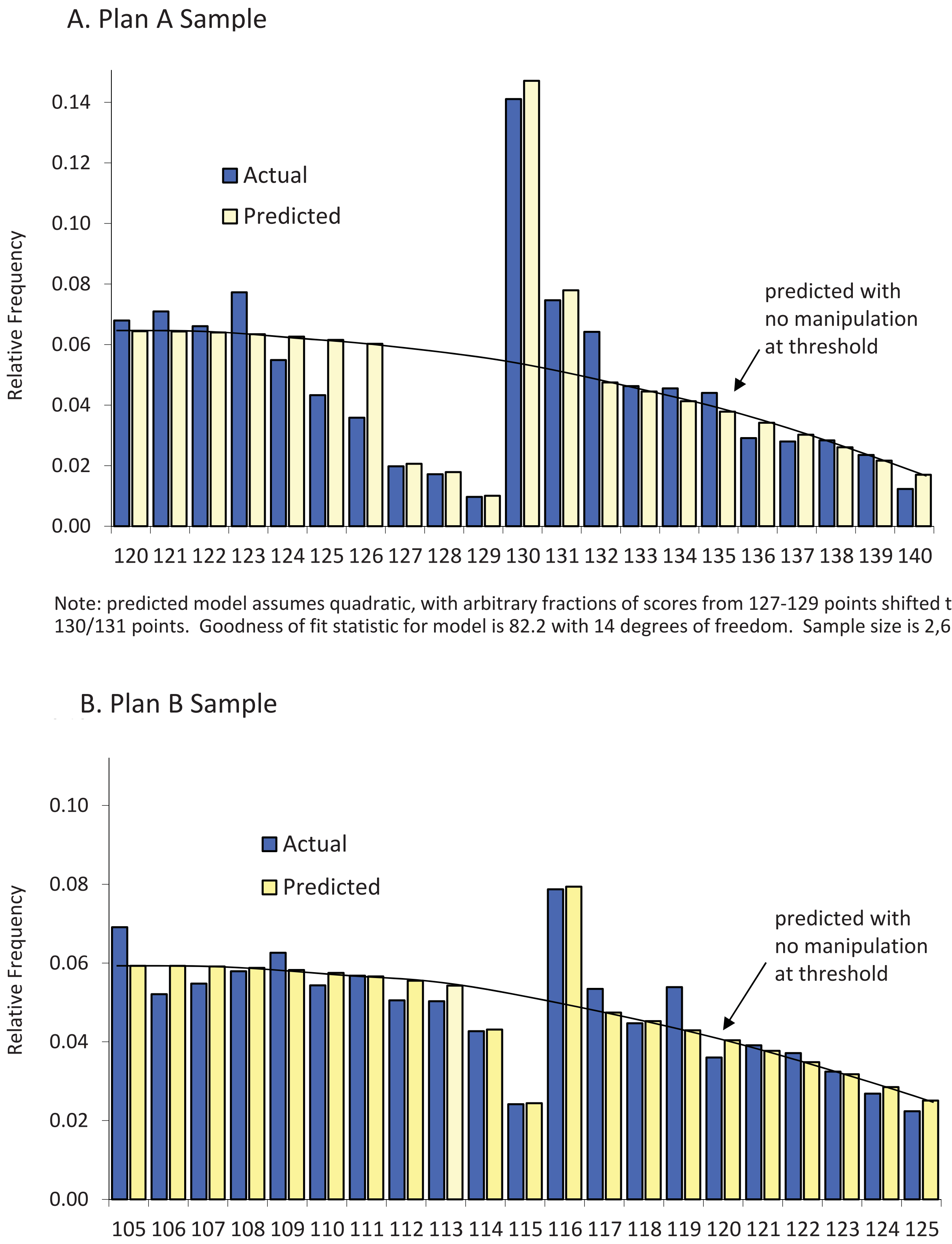


Figure 3. Histograms of First IQ Scores



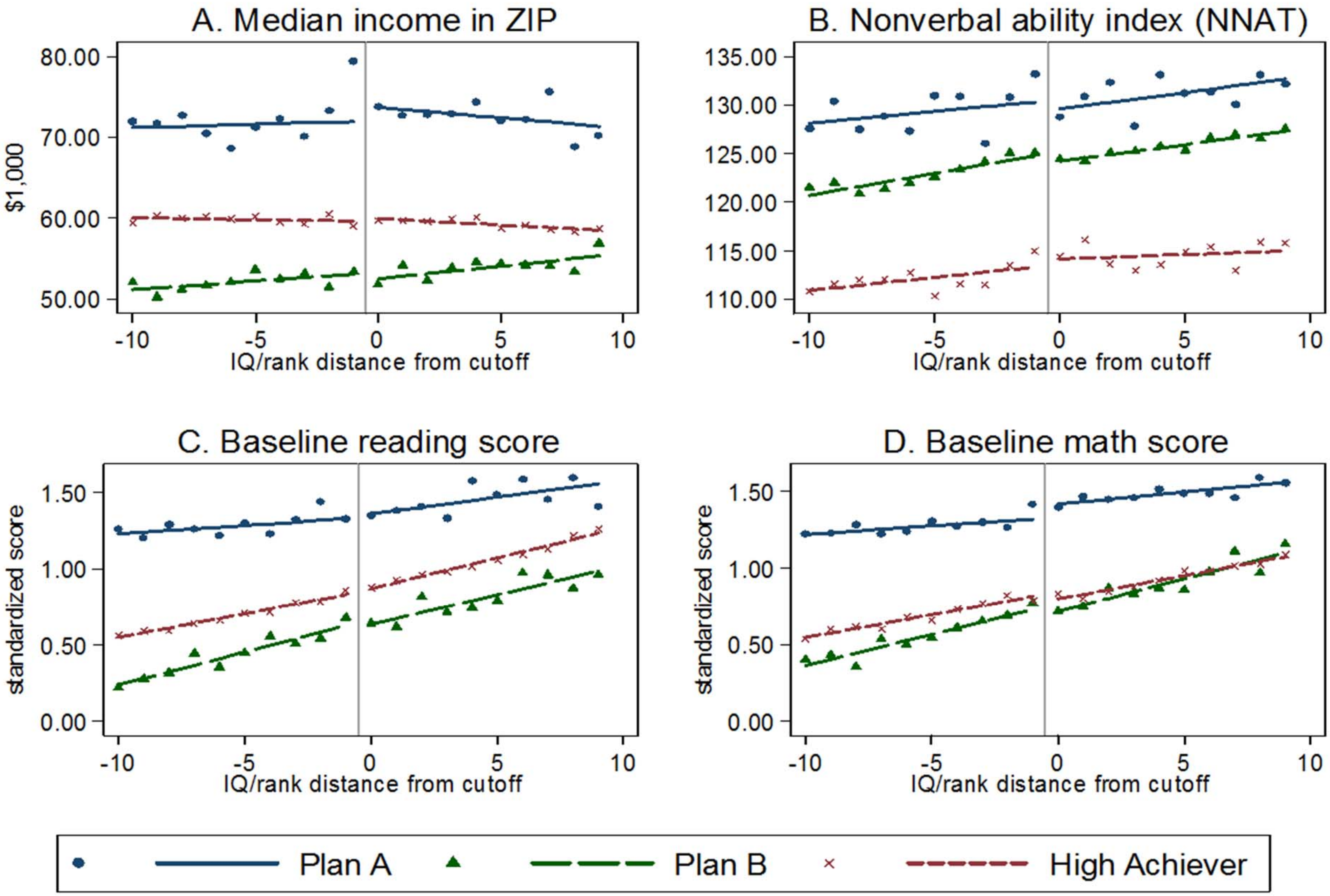
A. Plan A Sample

Note: predicted model assumes quadratic, with arbitrary fractions of scores from 127-129 points shifted to 130/131 points. Goodness of fit statistic for model is 82.2 with 14 degrees of freedom. Sample size is 2,679.

B. Plan B Sample

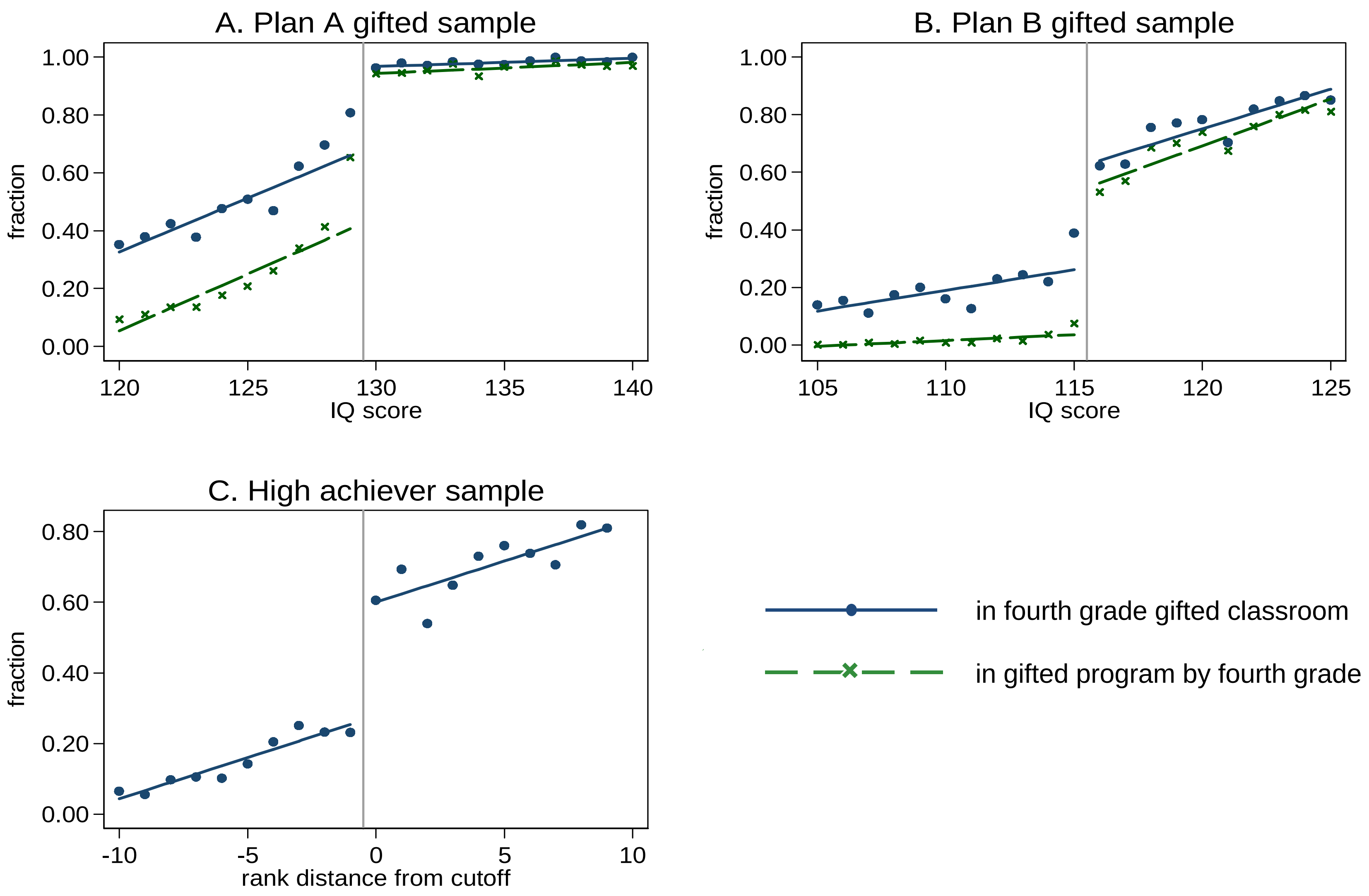
Note: predicted model assumes quadratic, with arbitrary fractions of scores from 114-115 points shifted to 116/117 points. Goodness of fit statistic for model is 35.7 with 15 degrees of freedom. Sample size is 4,472.

Figure 4. Student characteristics and baseline scores, by group



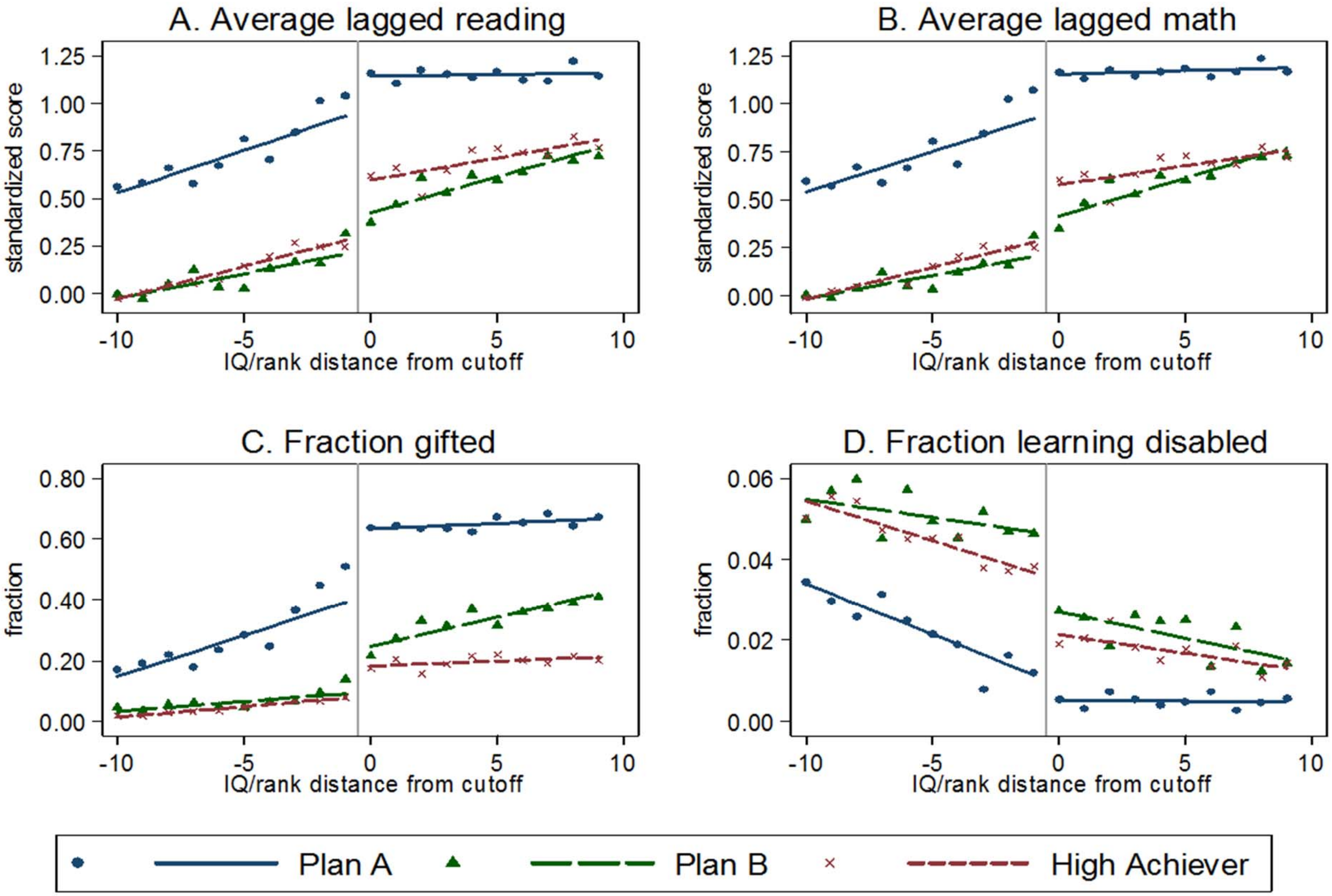
Note: Means and fitted values from local linear regressions. Reading and math scores are standardized within district and year.

Figure 5. First stage relationships



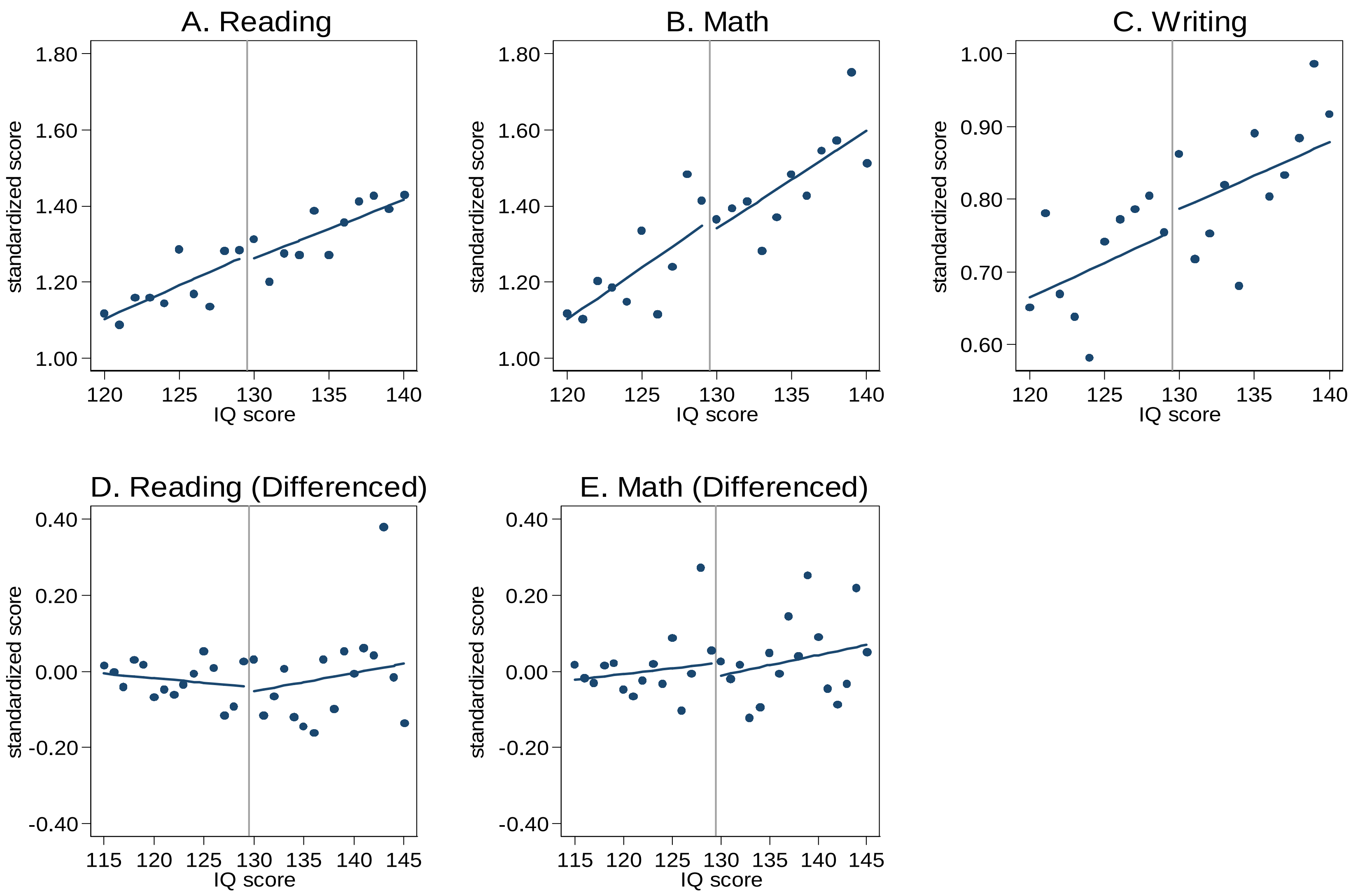
Note: Means and fitted values from local linear regressions. See text for more details.

Figure 6. Characteristics of fourth grade classmates, by group



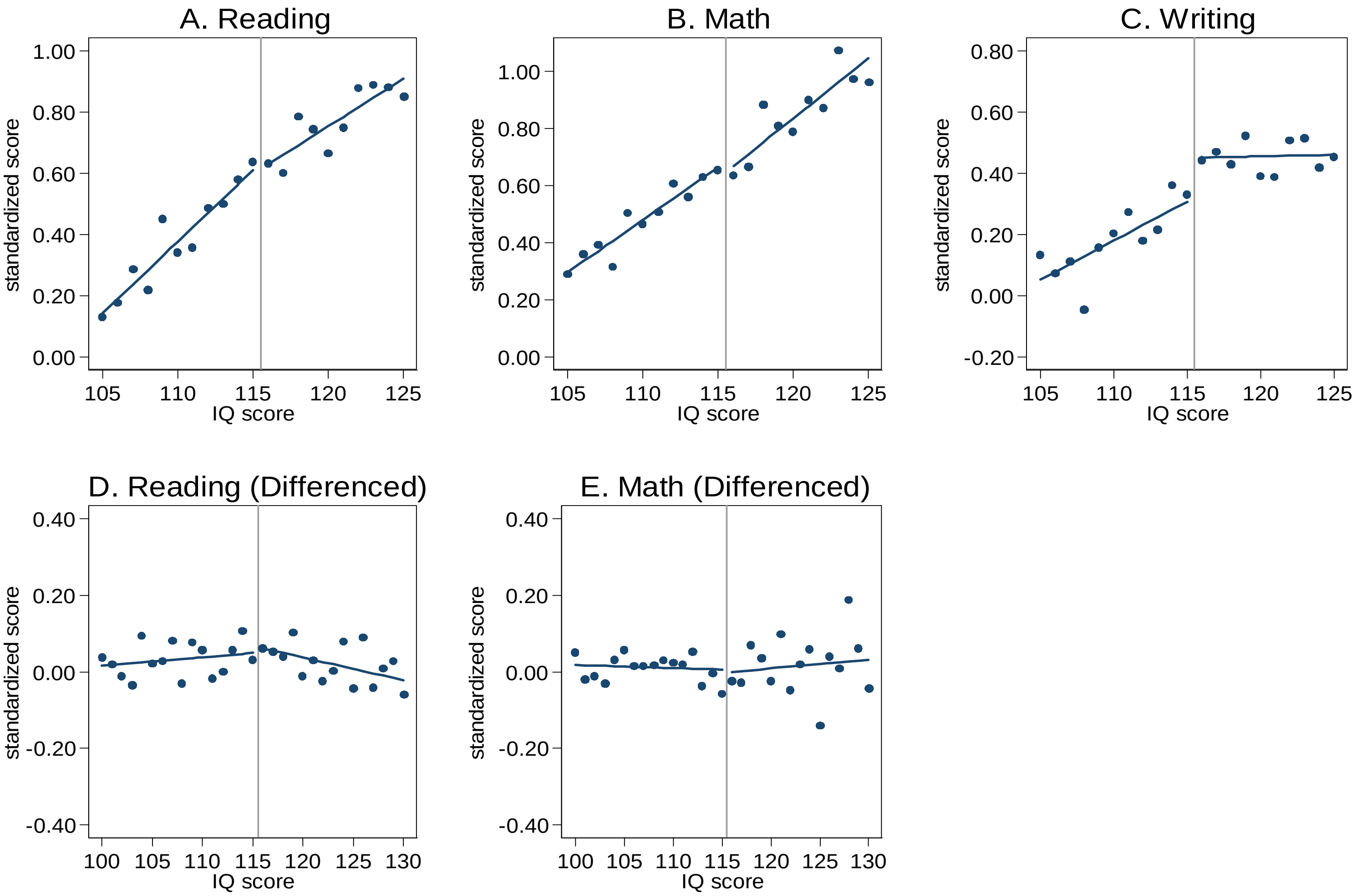
Note: Means and fitted values from local linear regressions. Reading and math scores are standardized within district and year. See text for more details.

Figure 7. Fourth grade standardized test scores, Plan A sample



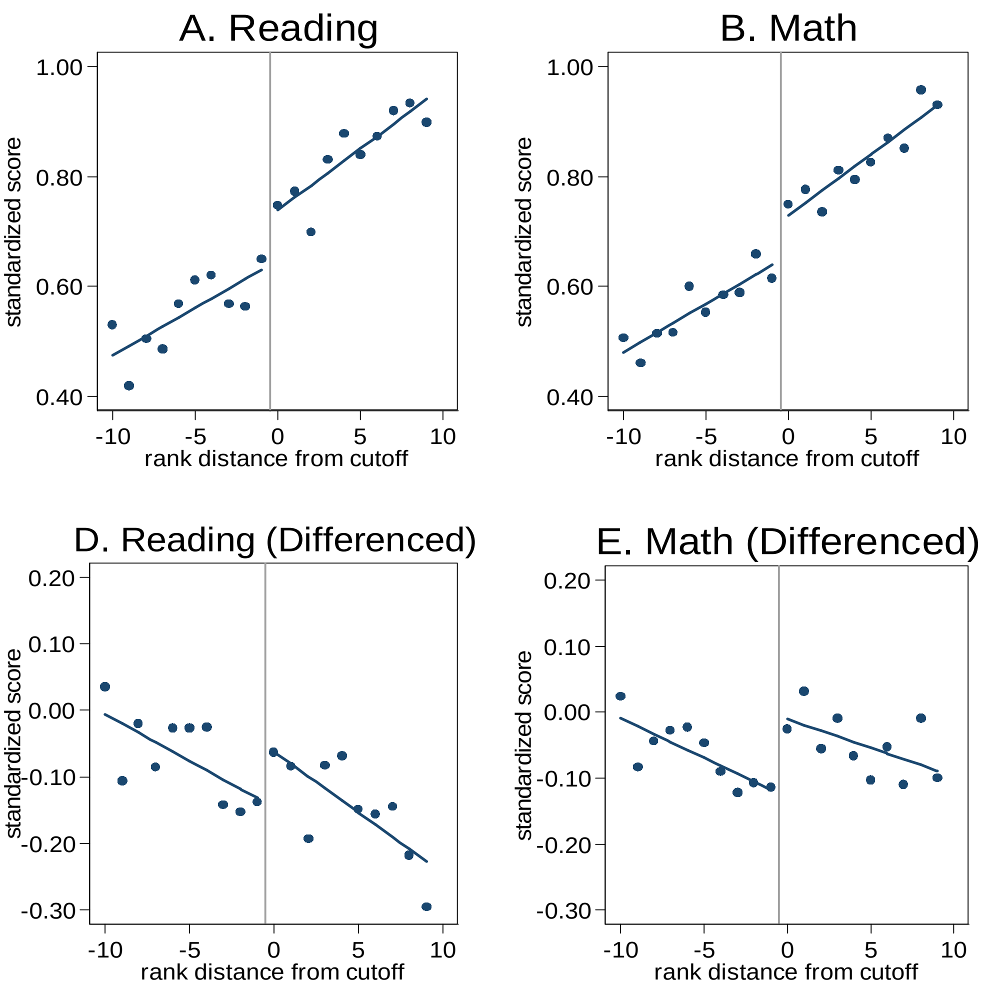
Note: Means and fitted values from local linear regressions. All test scores are standardized within district and year. See text for more details.

Figure 8. Fourth grade standardized test scores, Plan B sample



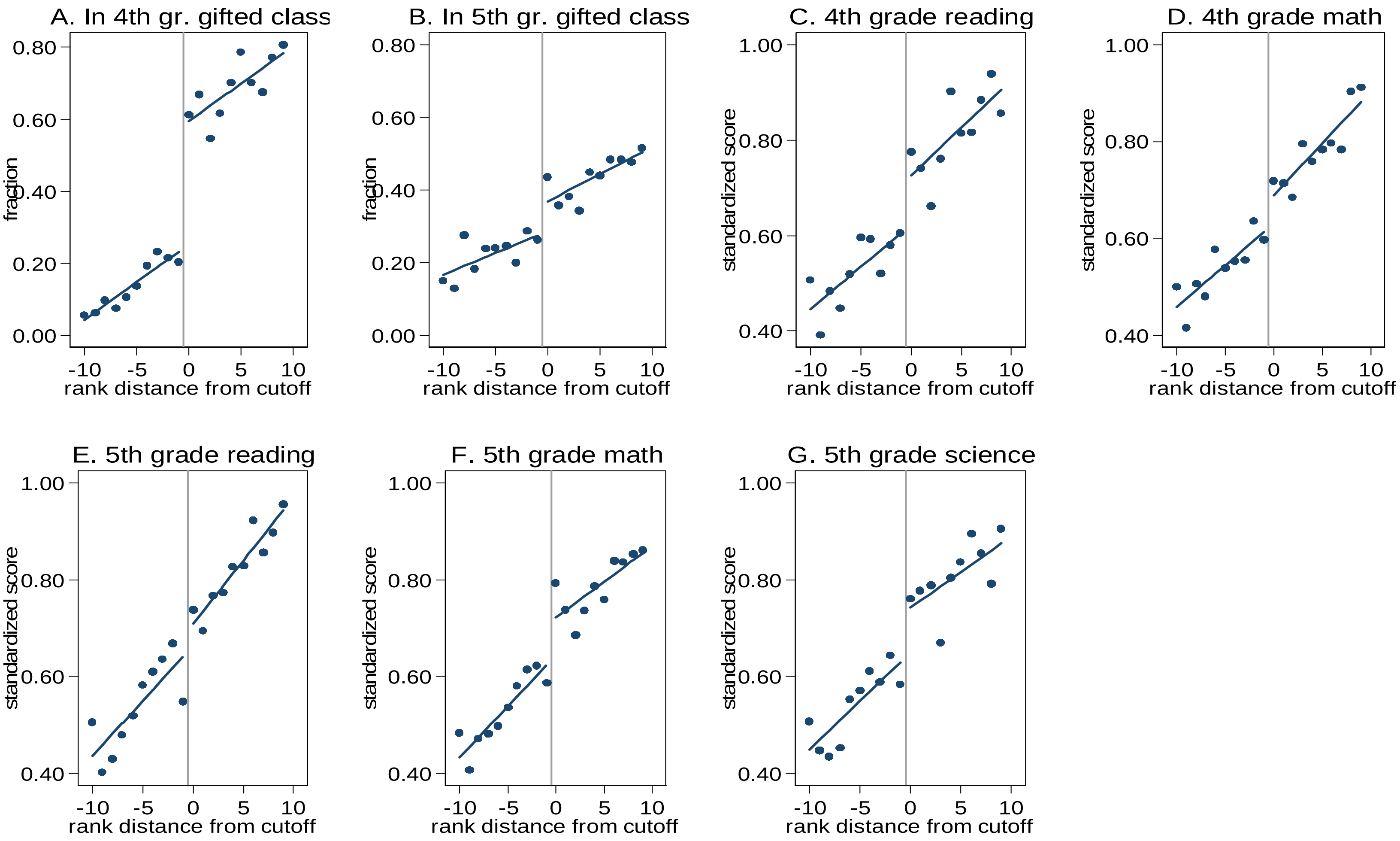
Note: Means and fitted values from local linear regressions. All test scores are standardized within district and year. See text for more details.

Figure 9. Fourth grade standardized test scores, High achiever sample



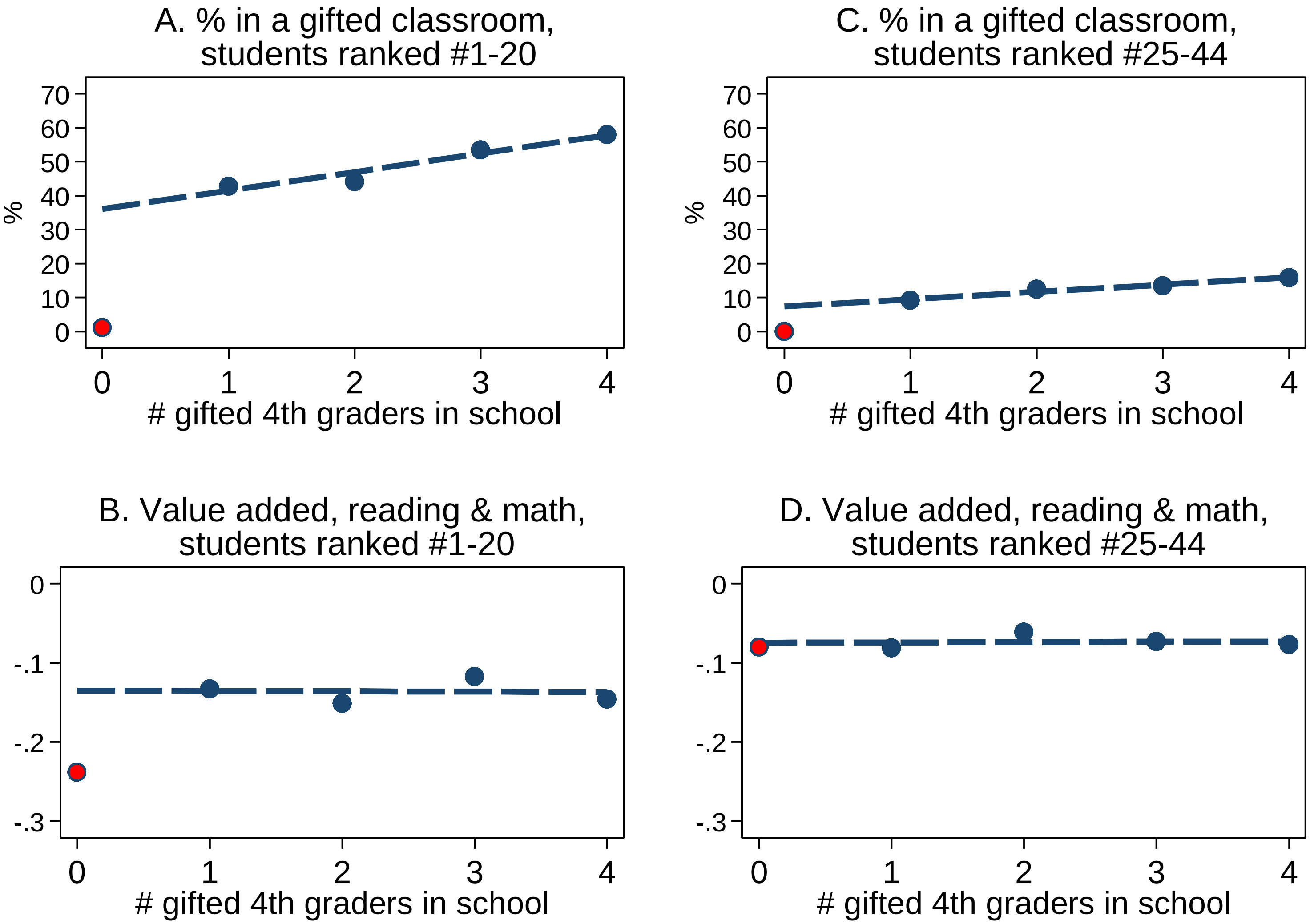
Note: Means and fitted values from local linear regressions. All test scores are standardized within district and year. See text for more details.

Figure 10. Discontinuities at Rank Threshold for Admission to 4^{th} Grade Gifted Class for Subsample of High Achievers Observed in 5th Grade

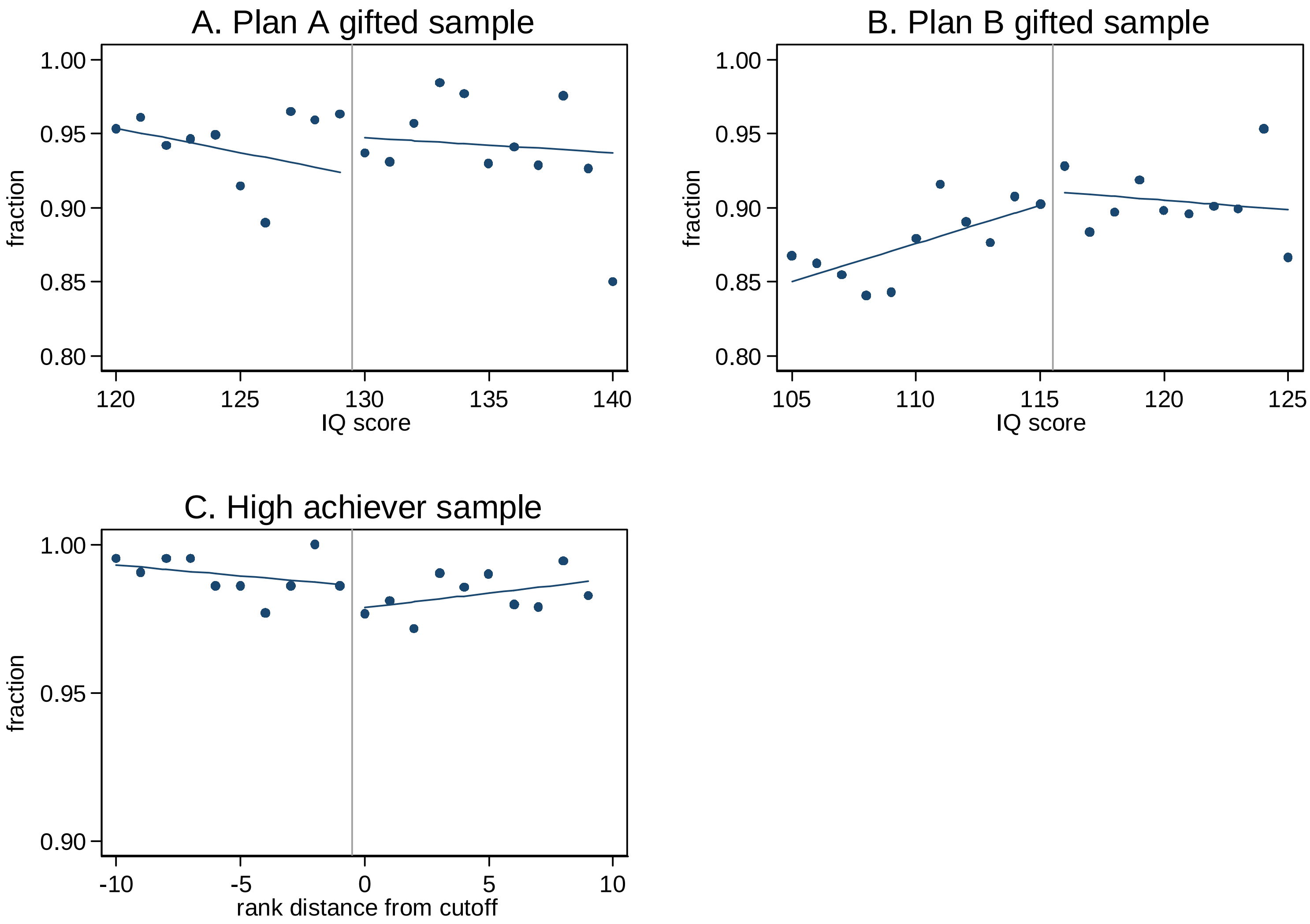


Note: Means and fitted values from local linear regressions. All test scores are standardized within district and year. See text for more details.

Figure 11. Between-school analysis of high achievers

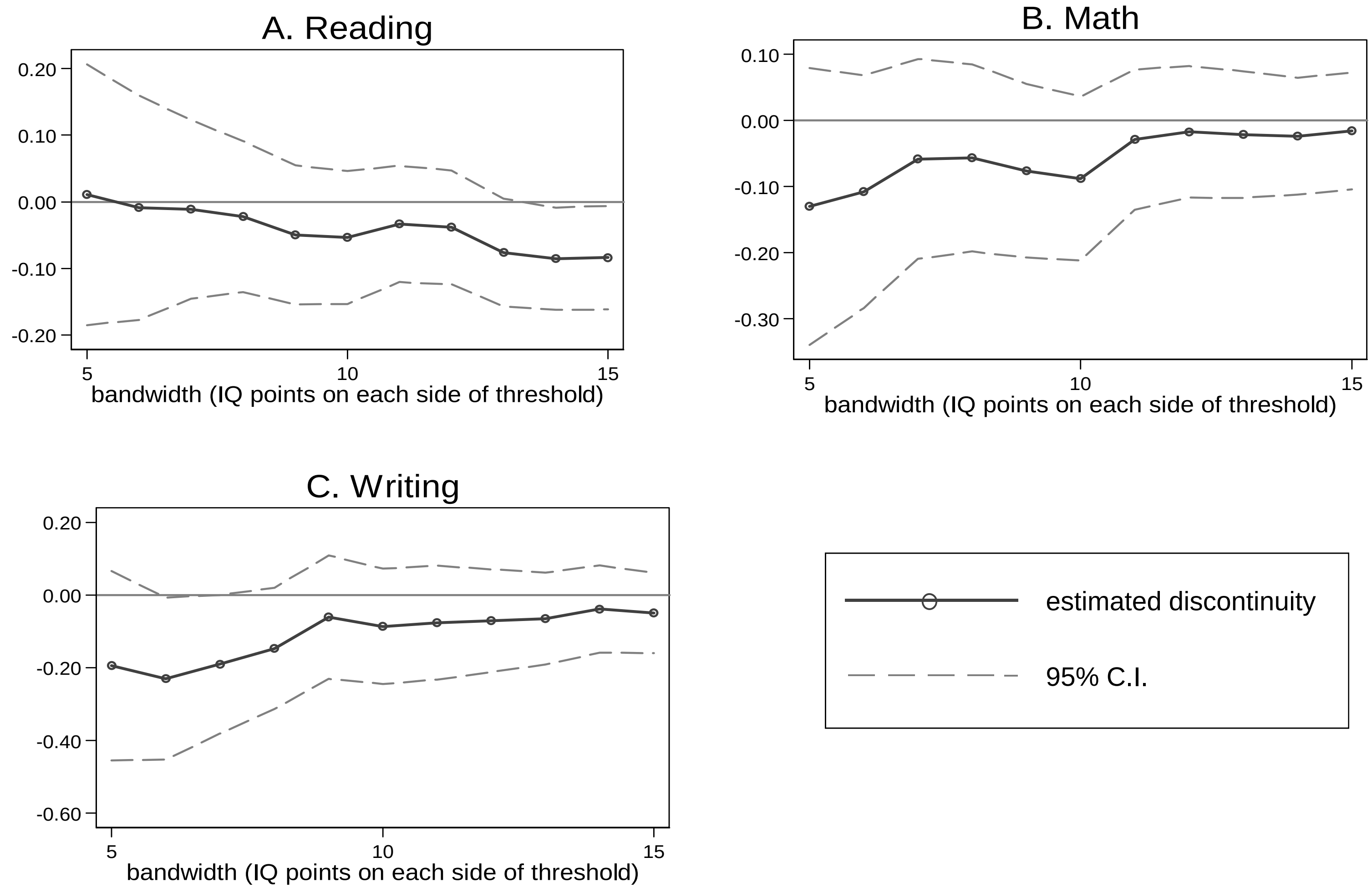


Appendix Figure 1. Fractions of IQ-tested students and potential high achievers for whom 4th grade outcomes are observed



Note: Means and fitted values from local linear regressions.

Appendix Figure 2. Estimated discontinuities in fourth grade scores from local linear regressions with varying bandwidths, Plan A sample

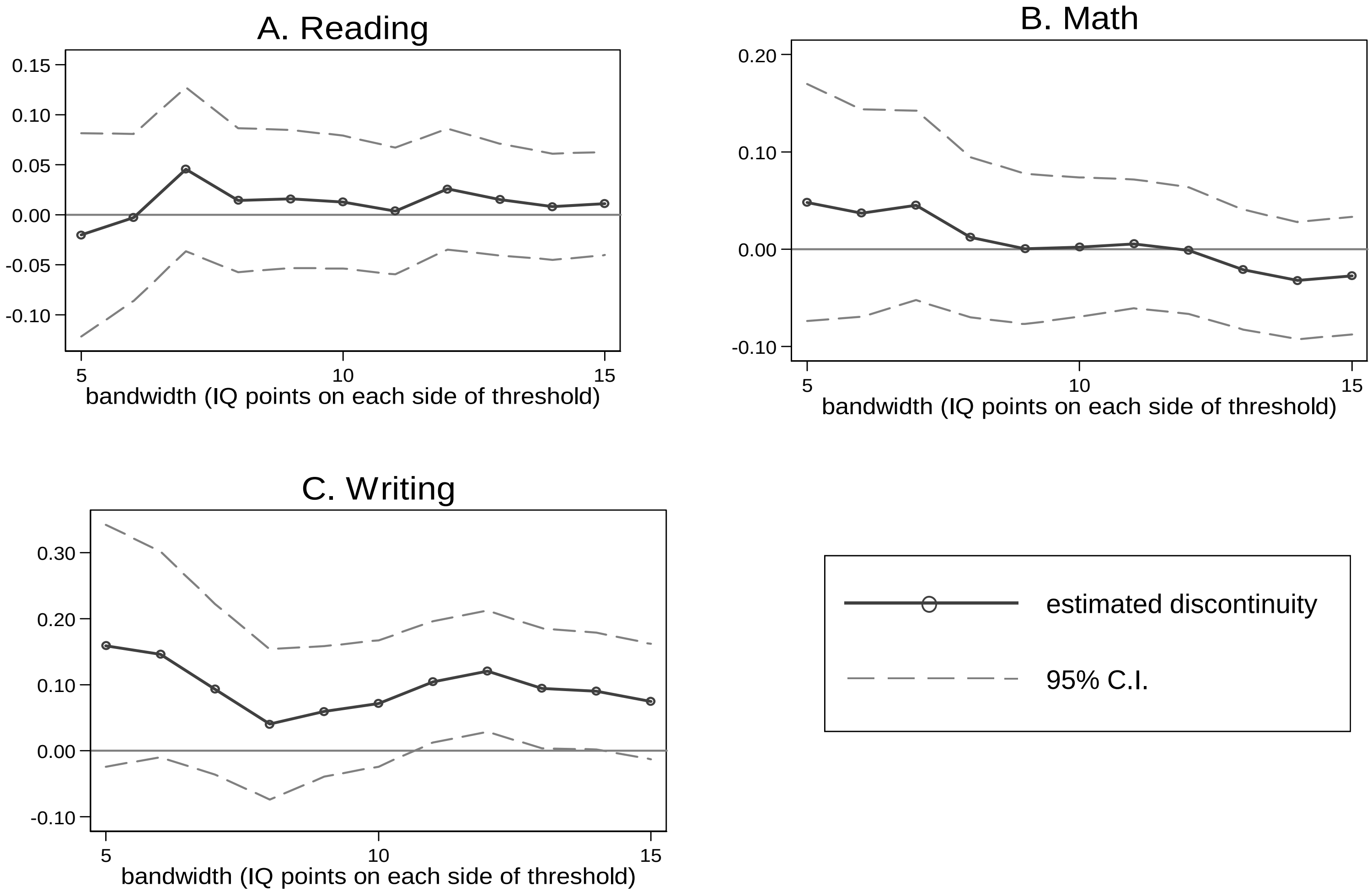


A. Reading

B. Math

C. Writing

Appendix Figure 3. Estimated discontinuities in fourth grade scores from local linear regressions with varying bandwidths, Plan B sample



A. Reading

B. Math

C. Writing

Appendix Figure 4. Estimated discontinuities in fourth grade scores
from local linear regressions with varying bandwidths, High achiever sample

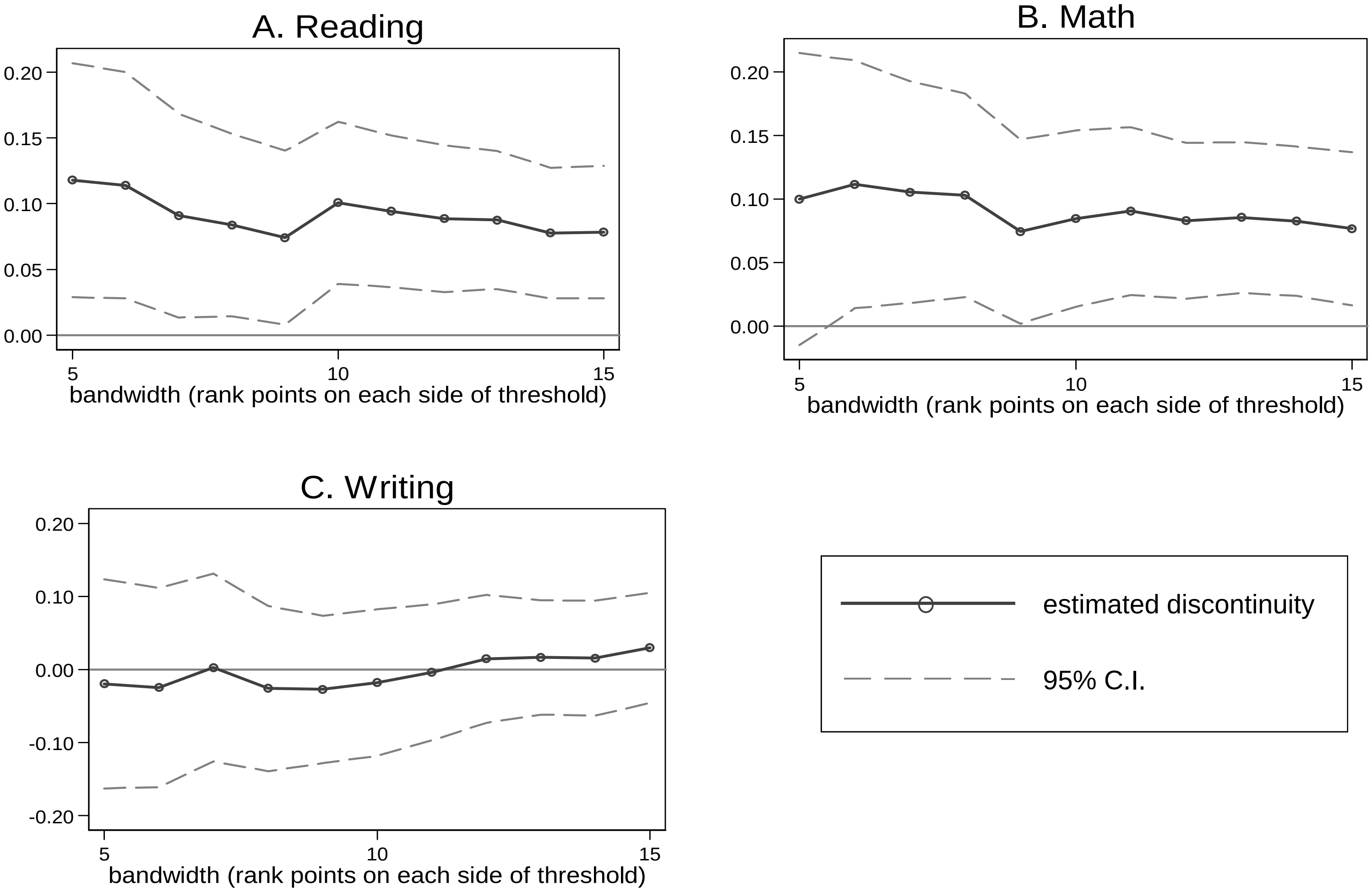


Table 1. Sample characteristics

	All students in district for 3rd grade (2004-11)	Plan A Gifted Sample, IQ Є [120,140]	Plan B Gifted Sample, IQ Є [105,125]	Advantaged High Achievers, +/- 10 from cutoff	Disdvantaged High Achievers, +/- 10 from cutoff
	(1)	(2)	(3)	(4)	(5)
Student demographics					
Female (%)	48	48	49	51	52
White (%)	30	58	19	51	16
Black (%)	37	7	34	13	45
Hispanic (%)	26	22	38	25	31
Asian (%)	4	8	5	7	4
Free lunch eligible (%)	52	--	86	--	100
English language learner (%)	10	--	17	--	2
Plan B eligible (%)	55	--	100	--	100
Median income in ZIP ($1,000s)	57.5	72.3	52.6	68.7	50.2
Mean IQ score	102.7	128.5	113.8	114.5	106.4
Percent taking test	*20*	*100*	*100*	*11*	*11*
Mean NNAT (screening test)	104.9	129.8	123.6	116.7	110.0
Percent taking test	*53*	*77*	*86*	*67*	*72*
Third grade state test scores (standardized)					
Mean reading test	-0.01	1.35	0.55	1.07	0.65
Mean math test	0.00	1.37	0.66	1.02	0.57
Fourth grade state test scores (standardized)					
Mean reading test	0.00	1.24	0.52	0.89	0.48
Mean math test	0.02	1.31	0.61	0.90	0.46
Mean writing test	0.04	0.76	0.29	0.55	0.25
Characteristics of school-wide peers					
Mean 3rd grade reading test		0.33	0.04	0.23	-0.06
Mean 3rd grade math test		0.34	0.02	0.21	-0.08
Percent free lunch eligible		31	61	45	73
Characteristics of school's fourth grade gifted classroom					
Mean 3rd grade reading test		1.14	0.74	1.14	0.84
Mean 3rd grade math test		1.16	0.73	1.09	0.79
Percent gifted by end of 3rd grade		61	37	32	21
Number of observations	*159,895*	*2,679*	*4,472*	*2,098*	*2,046*

Note: Sample in column 1 includes one observation per student for students observed in a non-charter district elementary school in 3rd grade between 2004 and 2011. Fourth grade test scores are reported only for those who stayed in the district and advanced to fourth grade in the following year. Free lunch status (FRL) and English language learner (ELL) status are measured at the end of third grade. Sub-samples in columns 2&3 include students who received a first IQ test in 2004-2010 while completing grades 1-3 (if tested in spring) or entering grades 2-4 (if tested in summer/fall), with IQ score within 10 points of the gifted eligibility threshold, and who wrote state standardized tests in both 3rd and 4th grade. Non-verbal ability index (NNAT) is measured at the end of 2nd grade for students who took the test between 2005-2009. Plan B eligibility status is measured in the year that the first IQ test was administered. Sub-samples in columns 4&5 include students in 4th grade in 2009-2012 in schools that complied with District's ranking formula and had at least one self-contained gifted/high-achiever classroom for 4th graders. State 3rd and 4th grade test scores are in standard deviation units (standardized across district within year and grade.)

Table 2. Regression Discontinuity Estimates for Plan A Gifted Sample

	Baseline Achievement		First Stage Models			Reduced Form Outcomes		
	Baseline Reading (1)	Baseline Math (2)	Probability in Gifted Classroom (3)	Probability Classified as Gifted (4)	Mean Peer Lagged Scores (5)	4th Grade Reading (6)	4th Grade Math (7)	4th Grade Writing (8)
1. No controls	0.023	0.088	0.270**	0.498**	0.173**	-0.016	-0.033	0.026
	(0.073)	(0.061)	(0.044)	(0.044)	(0.049)	(0.059)	(0.070)	(0.088)
2. Cohort and school fixed effects	0.033	0.094	0.267**	0.494**	0.143**	-0.052	-0.084	-0.083
and student controls	(0.075)	(0.059)	(0.046)	(0.044)	(0.047)	(0.050)	(0.064)	(0.081)
Sample size (rows 1-2)	*2,679*	*2,679*	*2,679*	*2,679*	*2,579*	*2,679*	*2,679*	*2,679*
3. Donut hole specification	0.074	0.100	0.359**	0.609**	0.217**	-0.138	-0.129	-0.112
(exclude IQ scores 127-131)	(0.117)	(0.109)	(0.077)	(0.061)	(0.075)	(0.089)	(0.101)	(0.154)
Sample size	*1,976*	*1,976*	*1,976*	*1,976*	*1,898*	*1,976*	*1,976*	*1,976*
4. Differenced specification	--	--	--	--	--	-0.022	-0.047	--
(based on change in test scores)						(0.054)	(0.052)	
Sample size	--	--	--	--	--	*3,954*	*3,954*	--
5. Differenced model with donut hole	--	--	--	--	--	-0.093	-0.034	--
(exclude IQ scores 127-131)						(0.063)	(0.063)	
Sample size	--	--	--	--	--	*3,251*	*3,251*	--

Notes: Based on Plan A analysis sample described in column 2 of Table 1. Entries are estimated coefficients on a dummy for IQ≥130 from models that include a linear term in IQ interacted with the dummy. Models in rows 2-5 control for student demographics (age, gender, race/ethnicity, and median household income in ZIP code), dummies indicating year of the first IQ test and year in fourth grade, and a complete set of school dummies. Models in rows 2-3, columns 3-8 also control for baseline test scores. Donut-hole samples (rows 3 & 5) exclude students with IQ∈[127,131]. Differenced models (rows 4-5) use expanded bandwidth sample that includes IQ∈[115,145]. Standard errors, clustered by school, in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 3. Regression Discontinuity Estimates for Plan B Gifted Sample

	Baseline Achievement		First Stage Models			Reduced Form Outcomes		
	Baseline Reading (1)	Baseline Math (2)	Probability in Gifted Classroom (3)	Probability Classified as Gifted (4)	Mean Peer Lagged Scores (5)	4th Grade Reading (6)	4th Grade Math (7)	4th Grade Writing (8)
1. No controls	-0.040	-0.062	0.365**	0.522**	0.200**	-0.028	-0.033	0.118*
	(0.044)	(0.041)	(0.029)	(0.028)	(0.038)	(0.046)	(0.038)	(0.052)
2. Cohort and school fixed effects	-0.039	-0.045	0.359**	0.519**	0.240**	0.013	0.004	0.116*
and student controls	(0.045)	(0.041)	(0.027)	(0.027)	(0.030)	(0.032)	(0.035)	(0.047)
Sample size (row 1-2)	*4,472*	*4,472*	*4,472*	*4,472*	*4,251*	*4,472*	*4,472*	*4,472*
3. Donut hole specification	-0.040	-0.030	0.453**	0.612**	0.350**	0.052	0.063	0.179*
(exclude IQ scores 114-117)	(0.070)	(0.068)	(0.037)	(0.035)	(0.041)	(0.052)	(0.058)	(0.071)
Sample size	*3,582*	*3,582*	*3,582*	*3,582*	*3,405*	*3,582*	*3,582*	*3,582*
4. Differenced specification	--	--	--	--	--	0.015	-0.005	--
(based on change in test scores)						(0.030)	(0.033)	
Sample size						*6,154*	*6,154*	
5. Differenced model with donut hole	--	--	--	--	--	0.038	0.018	--
(exclude IQ scores 114-117)						(0.048)	(0.043)	
Sample size						*5,264*	*5,264*	

Notes: Based on Plan B analysis sample described in column 3 of Table 1. Entries are estimated coefficients on a dummy for IQ≥116 from models that include a linear term in IQ interacted with the dummy. Models in rows 2-5 control for student demographics (age, gender, race/ethnicity, and median household income in ZIP code), dummies indicating year of the first IQ test and year in fourth grade, and a complete set of school dummies. Models in rows 2-3, columns 3-8 also control for baseline test scores. Donut hole sample (rows 3 & 5) excludes students with IQ∈[114,117]. Differenced models (rows 4-5) use expanded bandwidth sample that includes IQ∈[100,130]. Standard errors, clustered by school, in parentheses. + p < 0.10, * p < 0.05, ** p < 0.01.

Table 4a. Regression Discontinuity Estimates for High Achiever Sample

	Baseline Achievement		First Stage Models		Reduced Form Outcomes (Fourth Grade Scores)		
	Baseline Reading	Baseline Math	Prob. in 4th Grade Gifted Classroom	Mean Peer Lagged Scores	4th Grade Reading	4th Grade Math	4th Grade Science
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. No controls	0.008	-0.046	0.323**	0.273**	0.092**	0.073+	-0.011
	(0.029)	(0.043)	(0.025)	(0.028)	(0.034)	(0.039)	(0.054)
2. Cohort and school fixed effects and student controls	0.015	-0.044	0.319**	0.277**	0.093**	0.087*	-0.012
	(0.027)	(0.040)	(0.026)	(0.029)	(0.031)	(0.035)	(0.051)
3. Differenced specification (based on change in test scores)	--	--	--	--	0.092**	0.105*	--
					(0.033)	(0.041)	
Sample size	*4,144*	*4,144*	*4,144*	*4,041*	*4,144*	*4,144*	*4,144*

Notes: Based on high achiever analysis sample, described in columns 6-7 of Table 1. Standard errors, clustered by school, in parentheses. See text for more details of model specifications. + p < 0.10, * p < 0.05, ** p < 0.01

Table 4b. Regression Discontinuity Estimates for 5th Grade Outcomes of Fourth Grade High Achiever Sample

			Baseline Achievement (3rd Grade)		Reduced Form Outcomes (Fifth Grade Scores)		
	Prob. Stayed in District for 5th Grade	Prob. in 5th Grade Gifted Classroom	Baseline Reading	Baseline Math	5th Grade Reading	5th Grade Math	5th Grade Science
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. No controls	-0.001	0.070**	0.054	-0.049	0.046	0.078+	0.096+
	(0.019)	(0.025)	(0.033)	(0.057)	(0.046)	(0.046)	(0.054)
2. Cohort and school fixed effects and student controls	0.001	0.068*	0.050+	-0.064	0.052	0.077+	.113*
	(0.019)	(0.027)	(0.029)	(0.053)	(0.044)	(0.045)	(0.045)
3. Differenced specification (based on change in test scores)	--	--	--	--	0.013	0.117**	--
					(0.044)	(0.043)	
Sample size	*3089*	*2770*	*2770*	*2770*	*2770*	*2770*	*2,037*

Notes: Based on students in high achiever analysis sample who are observed in fourth grade up to 2011. Column 1 includes all students. Columns 2-7 includes only students who are observed though the end of fifth grade. Standard errors, clustered by school, in parentheses. See text for more details of model specifications. + p < 0.10, * p < 0.05, ** p < 0.01

Table 5. High Achiever Heterogeneity Analysis

	First Stage	2SLS RD Estimates - Peer Lagged Scores and Outcomes in 4th and 5th Grades									
	Prob. in 4th Grade Gifted Class	4th Gr. Peer Lagged Scores	4th Gr. Peer Fraction Gifted	4th Grade Reading	Diff. 4th-3rd Gr. Reading	4th Grade Math	Diff. 4th-3rd Gr. Math	4th Grade Writing	Diff. 5th-3rd Gr. Reading	Diff. 5th-3rd Gr. Math	5th Grade Science
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1. Full sample	0.32**	0.88**	0.30**	0.29**	0.29**	0.27*	0.33**	-0.04	0.04	0.35**	0.30*
	(0.03)	(0.05)	(0.02)	(0.10)	(0.10)	(0.11)	(0.13)	(0.16)	(0.12)	(0.13)	(0.12)
Free/Reduced Lunch Status											
2. Not FRL eligible	0.38**	0.98**	0.36**	0.19	0.12	0.20	0.22	0.01	0.06	0.30+	0.12
	(0.03)	(0.05)	(0.02)	(0.13)	(0.12)	(0.13)	(0.16)	(0.19)	(0.14)	(0.16)	(0.15)
3. FRL eligible	0.26**	0.75**	0.19**	0.45**	0.56**	0.37*	0.46*	-0.18	0.02	0.31	0.59*
	(0.05)	(0.10)	(0.03)	(0.17)	(0.20)	(0.19)	(0.21)	(0.30)	(0.26)	(0.25)	(0.24)
Race											
4. White	0.33**	0.92**	0.38**	-0.08	-0.08	-0.14	-0.05	-0.14	-0.16	0.02	-0.07
	(0.04)	(0.09)	(0.04)	(0.17)	(0.19)	(0.17)	(0.21)	(0.28)	(0.22)	(0.22)	(0.19)
5. Black & Hispanic	0.30**	0.88**	0.25**	0.60**	0.64**	0.51**	0.55**	-0.03	0.21	0.52*	0.62**
	(0.04)	(0.07)	(0.03)	(0.15)	(0.16)	(0.16)	(0.17)	(0.19)	(0.21)	(0.22)	(0.22)
School Share of FRL Students											
6. Low (<60%)	0.33**	0.86**	0.41**	0.14	0.16	0.28+	0.45*	-0.04	0.05	0.39*	0.20
	(0.03)	(0.06)	(0.03)	(0.14)	(0.13)	(0.16)	(0.19)	(0.20)	(0.17)	(0.17)	(0.17)
7. High (>=60%)	0.31**	0.90**	0.19**	0.46**	0.46**	0.25+	0.20	-0.02	0.05	0.27	0.46*
	(0.04)	(0.07)	(0.02)	(0.14)	(0.15)	(0.15)	(0.16)	(0.24)	(0.19)	(0.18)	(0.18)
School/class number of gifted children											
8. High (5 or more)	0.33**	0.87**	0.39**	0.25*	0.23+	0.19	0.29+	-0.11	-0.08	0.34*	0.19
	(0.03)	(0.06)	(0.02)	(0.12)	(0.12)	(0.14)	(0.16)	(0.18)	(0.16)	(0.14)	(0.15)
9. Low (1-4)	0.30**	0.90**	0.13**	0.38*	0.42**	0.42*	0.41*	0.06	0.25	0.37	0.56*
	(0.04)	(0.10)	(0.02)	(0.15)	(0.16)	(0.18)	(0.21)	(0.28)	(0.21)	(0.25)	(0.22)

Notes: see notes to table 4a and 4b. In all 2SLS models (columns 2-8) the first stage model is for the probability of being in the 4th gifted classroom. + p < 0.10, * p < 0.05, ** p < 0.01

Table 6. Estimated Group-level models for effect of having at least one student designated as gifted in 4th grade school cohort

	Students ranked 1-20			Students ranked 25-44		
	Baseline Scores: 3rd grade (reading and math combined)	First Stage: Fraction in 4th grade gifted classroom	Reduced Form: Difference in scores (4th-3rd grade, reading and math combined)	Baseline Scores: 3rd grade (reading and math combined)	First Stage: Fraction in 4th grade gifted classroom	Reduced Form: Difference in scores (4th-3rd grade, reading and math combined)
	(1)	(2)	(3)	(4)	(5)	(6)
Estimation sample: schools/classes with ≤4 gifted students						
1. Controls=linear control for number of gifted students in class	0.148+	0.352**	0.064+	0.249*	0.071**	-0.002
	(0.078)	(0.062)	(0.032)	(0.095)	(0.025)	(0.042)
2. add class-level controls & year dummies	-0.012	0.330**	0.089**	0.047	0.063*	0.006
	(0.032)	(0.063)	(0.033)	(0.028)	(0.024)	(0.040)
number of fourth grade school/cohorts	*255*	*255*	*255*	*255*	*255*	*255*
Estimation sample: schools/classes with ≤1 gifted student						
3. Controls as in line 2	-0.024	0.408**	0.067*	0.017	0.083**	0.012
	(0.029)	(0.057)	(0.031)	(0.028)	(0.016)	(0.037)
number of fourth grade school/cohorts	*116*	*116*	*116*	*116*	*116*	*116*

Note: Entries are coefficients on a dummy variable equal to one if the class (i.e., school-wide cohort in a given grade in a given year) had at least one gifted student in fourth grade. Full estimation sample includes 255 fourth grade cohorts at 94 elementary schools during the years 2009-2012. Models in columns 1-3 use average outcomes for top 20 ranked students in class (school-year cohort); models in columns 4-6 use average outcomes for students ranked 25-44 in class. Dependent variables are: group average of reading and math z-scores in 3rd grade (columns 1,4); fraction of students in the group who are in a gifted/high-achiever class in 4th grade (columns 2,5); and the group average change in reading and math z-scores from 3rd to 4th grade (columns 3,6). Student rank is based on the formula used by the district in 2009-2012 to determine placement of non-gifted students into 4th grade gifted classrooms. Group averages exclude gifted students whose rank falls within the indicated range and students who are missing test scores from either 3rd or 4th grade. Class-level controls include class average 3rd grade math and reading scores; fraction of students who are free lunch eligible; number of students in the class, and the number of students with non-missing test scores in both 3rd and 4th grade. Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 7: Estimated Group-level Models for Effect of Having at Least One Gifted Student in 4th Grade School Cohort allowing for heterogeneity by relative achievement within classroom

	Characteristics of Student in Group: Average standardized reading & math scores in 3rd grade	First stage: Fraction in gifted classroom in 4th grade	Reduced Form: Difference in scores (4th-3rd grade, reading and math combined)	2SLS Effect: Difference in scores (4th-3rd grade, reading and math combined)
	(1)	(2)	(3)	(4)
1. All ranked 1-20	0.77	0.33**	0.09**	0.27**
		(0.06)	(0.03)	(0.07)
FRL Status				
2. Not FRL eligible	1.11	0.44**	0.07	0.16
		(0.08)	(0.08)	(0.18)
3. FRL eligible	0.64	0.33**	0.12**	0.36**
		(0.06)	(0.04)	(0.12)
Race				
4. White	1.07	0.31**	0.13	0.43
		(0.09)	(0.09)	(0.30)
5. Black & Hispanic	0.66	0.34**	0.08*	0.23*
		(0.07)	(0.04)	(0.11)
Rank within top 20				
6. Ranked 1-5	1.30	0.42**	0.04	0.11
		(0.09)	(0.07)	(0.15)
7. Ranked 6-10	0.83	0.41**	0.07	0.16
		(0.07)	(0.04)	(0.11)
8. Ranked 11-15	0.59	0.32**	0.13**	0.39*
		(0.07)	(0.05)	(0.15)
9. Ranked 16-20	0.42	0.21**	0.10*	0.50*
		(0.06)	(0.04)	(0.22)

Notes: See notes to Table 6. As in row 2 of Table 6, all models control for class-wide average test scores in reading and math, class-wide fraction of FRL participants, size of class, and number of students in class with valid test scores. Sample includes 255 4th grade school/year cohorts. Standard errors in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix Table 1. Differences in Characteristics of Students Just Above and Just Below IQ Threshold and Estimated Coefficients from Conditional Logit Model of Probability of Being Above Threshold

	Plan A Sample (127-131 IQ)		**Plan B Sample (114-117 IQ)**	
	Difference in Means Above-Below	Coefficients from Logit for Pr(Above)	Difference in Means Above-Below	Coefficients from Logit for Pr(Above)
	(1)	(2)	(3)	(4)
White	0.05	0.33	0.02	0.34
	(0.05)	(0.29)	(0.03)	(0.28)
Black	0.03	0.75	0.04	0.46+
	(0.03)	(0.48)	(0.03)	(0.27)
Hispanic	-0.05	0.03	-0.03	0.24
	(0.04)	(0.32)	(0.03)	(0.26)
Female	-0.01	-0.02	-0.01	-0.07
	(0.05)	(0.20)	(0.04)	(0.15)
Median Income (1000's)	0.25	-0.01	0.56	0.00
	(2.16)	(0.01)	(1.27)	(0.01)
On Free/Reduced Lunch	--	--	0.03	0.28
			(0.03)	(0.20)
School Fraction FRL	-0.02	-0.99	-0.01	-0.17
	(0.02)	(0.69)	(0.02)	(0.39)
NNAT Score (normalized)	0.27	0.00	-0.74	-0.01
	(1.28)	(0.01)	(0.77)	(0.01)
3rd grade FCAT Reading (Normalized)	-0.01	-0.12	0.04	0.10
	(0.07)	(0.14)	(0.05)	(0.11)
3rd grade FCAT Math (Normalized)	0.11+	0.33+	0.01	-0.02
	(0.06)	(0.17)	(0.05)	(0.12)
SAT Reading (Normalized)*	0.03	--	0.09	--
	(0.07)		(0.06)	
SAT Math (Normalized)*	0.13+	--	0.02	--
	(0.08)		(0.07)	
P-value for Chi-square test: all coefficients=0	--	0.38	--	0.73
Number of observations	703		890	

Notes: Columns 1 and 3 show differences in mean characteristics between students with IQ's just above versus just below the Plan A or Plan B eligibility thresholds. Columns 2 and 4 show estimated coefficients for logistic regression model of probability of being above the threshold, conditional on being just above or just below the threshold. Standard errors in parentheses. * SAT Reading Reading and SAT Math scores are from the year the student first took the IQ test (i.e. the year before potential entry to the gifted program). SAT Reading scores are available only from 2004-2008; SAT Math scores are available only from 2004-2007. Sample sizes for these variables are 610 (Plan A, SAT Reading); 560 (Plan A, SAT Math); 742 (Plan B, SAT Reading); 626 (Plan B, SAT Math).

Appendix Table 2. Regression Discontinuity Estimates for Plan A & Plan B Gifted Samples Using Alternative Measure of Baseline Achievement

	Plan A Sample					Plan B Sample				
	Baseline Achievement (Standardized Test Score in Year of First IQ Test)		Reduced Form Outcomes			Baseline Achievement (Standardized Test Score in Year of First IQ Test)		Reduced Form Outcomes		
	Baseline Reading	Baseline Math	4th Grade Reading	4th Grade Math	4th Grade Writing	Baseline Reading	Baseline Math	4th Grade Reading	4th Grade Math	4th Grade Writing
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. No controls	0.04	0.00	-0.05	-0.04	-0.05	-0.02	0.01	-0.03	-0.02	0.13*
	(0.07)	(0.08)	(0.06)	(0.07)	(0.10)	(0.05)	(0.06)	(0.05)	(0.04)	(0.06)
2. Cohort and school fixed effects	0.03	-0.04	-0.06	-0.05	-0.16+	0.00	0.01	-0.02	-0.02	0.12*
and student controls	(0.07)	(0.08)	(0.06)	(0.06)	(0.09)	(0.05)	(0.06)	(0.04)	(0.04)	(0.05)
Sample size (row 1-2)	*2362*	*2154*	*2362*	*2362*	*2362*	*3788*	*3201*	*3788*	*3788*	*3788*
3. Donut hole specification	0.09	-0.17	-0.13	-0.04	-0.18	-0.04	0.17*	0.03	-0.01	0.20*
	(0.11)	(0.14)	(0.09)	(0.11)	(0.16)	(0.09)	(0.08)	(0.07)	(0.07)	(0.08)
Sample size	*1752*	*1594*	*1752*	*1752*	*1752*	*3046*	*2575*	*3046*	*3046*	*3046*
4. Differenced specification			-0.09	-0.01				-0.01	-0.07	
(based on change in test scores)			(0.06)	(0.06)				(0.04)	(0.05)	
Sample size			*3472*	*3161*				*5191*	*4374*	
5. Differenced model with			-0.19*	0.00				0.04	-0.14*	
donut hole			(0.08)	(0.09)				(0.06)	(0.06)	
Sample size			*2862*	*2601*				*4449*	*3748*	

Note: Estimates from regression discontinuity models as described in note to Table 2 (Plan A sample) and Table 3 (Plan B sample). Baseline scores for students who received first IQ test in 1st or 2nd grade between 2004-2008 are SAT scores from the year the IQ test was given. Baseline scores are missing for students who received first IQ test in 1st or 2nd grade between 2009-2010, because SAT was not administered in those years, and students with missing scores are excluded from the estimation sample for all models. In addition, baseline math scores are missing for students who received IQ test in 1st or 2nd grade in 2008. All models include dummies for year and grade of first IQ test. Baseline scores are 3rd grade FCAT scores for all remaining students (who received first IQ test in 3rd grade). Standard errors, clustered by school, in parentheses. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix Table 3. Regression Discontinuity Estimates for 5th Grade Outcomes of Plan A and Plan B Gifted Samples

	Plan A Sample						Plan B Sample					
	Probability Stayed in District for 5th Grade	Baseline Achievement (3rd grade scores)		Reduced Form Outcomes (5th Grade Scores)			Probability Stayed in District for 5th Grade	Baseline Achievement (3rd grade scores)		Reduced Form Outcomes (5th Grade Scores)		
		Reading	Math	Reading	Math	Science		Reading	Math	Reading	Math	Science
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1. No controls	0.01 (0.02)	0.04 (0.08)	0.11+ (0.07)	-0.05 (0.08)	-0.08 (0.07)	-0.07 (0.07)	-0.01 (0.02)	-0.04 (0.05)	-0.05 (0.04)	-0.06 (0.05)	-0.05 (0.04)	-0.06 (0.05)
2. Cohort and school fixed effects and student controls	0.01 (0.02)	0.07 (0.08)	0.14* (0.06)	-0.10 (0.07)	-0.14* (0.06)	-0.16** (0.05)	0.00 (0.02)	-0.05 (0.04)	-0.03 (0.04)	0.00 (0.03)	-0.01 (0.03)	-0.01 (0.04)
Sample size (row 1-2)	*2546*	*2398*	*2398*	*2398*	*2398*	*1946*	*4378*	*4066*	*4066*	*4066*	*4066*	*3251*
3. Donut hole specification	0.01 (0.03)	0.12 (0.12)	0.14 (0.11)	-0.13 (0.11)	-0.20+ (0.10)	-0.21* (0.09)	-0.01 (0.02)	-0.05 (0.07)	-0.02 (0.07)	0.01 (0.06)	0.01 (0.05)	-0.02 (0.07)
Sample size	*1882*	*1775*	*1775*	*1775*	*1775*	*1480*	*3509*	*3253*	*3253*	*3180*	*3180*	*2591*
4. Differenced specification (based on change in test scores)				-0.01 (0.06)	-0.13* (0.05)					0.04 (0.03)	0.02 (0.03)	
Sample size				*3505*	*3504*					*5456*	*5455*	
5. Differenced model with donut hole				-0.01 (0.08)	-0.13* (0.06)					0.06 (0.05)	0.04 (0.04)	
Sample size				*2888*	*2887*					*4660*	*4657*	

Note: Column 1 (7) based on Plan A (plan B) analysis sample described in columns 2 (3) of Table 1. Columns 2-6 and 8-12 present estimates from regression discontinuity models as described in note to Table 2 (Plan A sample) and Table 4 (Plan B sample), and are based on the sub-samples of students with observed 5th grade outcomes. Standard errors, clustered by school, in parentheses. + p < 0.10, * p < 0.05, ** p < 0.01.

Data Appendix

Matching Students to Classrooms and Identification of Gifted Classrooms

For each course taken by each student, the data contain a course identifier, a subject identifier, and a teacher identifier, but they do not contain classroom identifiers. We therefore match students to classrooms by constructing all unique combinations of a school, year, course and teacher identifier and matching each student to one of these combinations for each of the three core subjects (Mathematics, Reading and Language Arts).

In a few schools, students rotate teachers in fourth grade so that the same teacher teaches a given subject to multiple classes throughout the day. For students in these schools, which make up about 5% of our sample, it is impossible to identify peers who sit in the same classroom at the same time of day. We therefore exclude these schools from the sample when estimating models of peer characteristics.

In the remaining fourth grade school-year cohorts, students are assigned the same teacher for all three core subjects and each school-year-course-teacher combination is assigned to 23 students on average (standard deviation = 3). In principle, students in these cohorts have the same group of peers in each core subject; but because the matching is imperfect (due to reassignments, coding errors, etc.) we use average characteristics of peers in the three core subjects as our measures of peer characteristics.

Finally, we classify a student as being placed in a gifted classroom if, in each of the three core subjects, at least one peer is classified as gifted *and* at least one of the following conditions is also satisfied:

- the gifted student has an EP on file stating he or she is in a gifted/high achiever classroom;
- the average lagged tests scores of peers in the classroom are significantly higher than the average of all other students in the cohort.

These two conditions rule out a small number of cases in which a student has a gifted peer but is not in a gifted classroom. This may occur when there are very few gifted students in the cohort and either the student(s) were placed in the gifted program after the school year began (too late for a gifted class to be formed) or the school was unable to hire a certified teacher and obtained a waiver from the District requirement of having a separate gifted classroom.

Construction of High Achiever Sample and Estimation of Cutoff Scores

To construct the estimation sample for the analysis of non-gifted high-achievers, we started with all students who were in fourth grade in the 2008-09 through 2011-12 school years—a total of 68,263 students in 527 school-year cohorts. We restrict the sample to to these four years because prior to 2008-09, the District did not prescribe a uniform ranking formula for determining which non-gifted students were placed in the gifted classrooms.

We then eliminated school-year cohorts for which classrooms could not be identified and those that did not have a gifted/high achiever classroom (either because there were no gifted students or because there were enough gifted students to fill an entire classroom and the school opted for a gifted-only classroom)—leaving 385 school-year cohorts.

For each of these school-year cohorts, we estimated the cutoff rank for placement in the gifted classroom as follows:

1. First, we assigned a class rank to each non-gifted fourth grade student with non-missing third grade standardized test scores using the district's prescribed rule. This rule is a lexicographic formula that first groups students based on their "achievement levels" on the reading and math portions of the third grade statewide achievement test. These achievement levels range from 1-5 and are based on the scale scores, with cutoffs set each year by the state. Students who achieve level 5 (the highest) in *both* reading and math are given highest priority, followed by students with a level 5 in reading and a 4 in math; those with a 4 in reading and 5 in math; those with a 4 in both reading and math, and so on. Within each of these groups, students are ranked using the sum of their scale scores in reading and math.

2. Next, we calculated an initial estimate of the cutoff rank, *c*, as the rank of the lowest-ranked non-gifted student in the gifted classroom.

   This estimate is subject to several sources of measurement error, including:
   - non-compliance with the District rule
   - classroom reassignments and/or errors in matching students to classrooms
   - missing third grade test scores

   We therefore replace *c* with $c' \in (c\text{-}10, c\text{+}9)$, where *c*′ is chosen using an iterative procedure to minimize the misclassification rate of students whose scores are outside an interval around the potential cutoff. Specifically, letting *c*′=*c* be the initial estimate of the cutoff rank, we replace *c*′ with *c*′+1 if $\sum_{r=c'-3}^{c'-2} T_r < \sum_{r=c'+1}^{c'+2} T_r$ or with *c*′-1 if $\sum_{r=c'-3}^{c'-2} T_r > \sum_{r=c'+1}^{c'+2} T_r$, where T_r is an indicator variable that is equal to one if the student with rank *r* is in the gifted classroom. We repeat this step until no further reduction in mismatch is possible.

3. Once we have estimated the cutoff for each fourth grade school-year cohort we then eliminate from the sample cohorts where there is still substantial mismatch (either because the school did not comply with the district's rule or due to other sources of measurement error). Specifically, for each school-year cohort we estimate the difference in the probability of treatment above and below the estimated cutoff for students with $r \in (c'\text{-}10, c'\text{+}9)$, and we keep cohorts for which a one-tailed test of H_0: $\left(E(T_r|r \geq c') - E(T_r|r < c')\right) = 0$ has a z-statistic of >1. This results in our estimation sample of 4,144 fourth grade students in 220 school-year cohorts.

   An analysis of the determinants of being excluded from our sample suggests that in most cases, our inability to accurately estimate the cutoff is due to schools' non-compliance with the District formula—especially in the first year that the formula was prescribed (i.e. in 2009). We dropped 60% of the 2009 fourth grade cohorts compared to 47% of the cohorts in 2010 and 32% in 2011 and 2012. Other significant predictors of being dropped from our sample are the number of students we coded as being in the gifted classroom (based on the criteria described above) and the number of students who are missing third grade test scores. Importantly, the likelihood of being excluded from our sample is *not* significantly correlated with school characteristics such as average tests scores or the fraction of students who are FRL eligible.

Appendix A: Student Satisfaction Survey

The District administers a “customer survey” each spring to students in grades 3-12 to assess students’ satisfaction with their learning environment. We use the fourth grade survey responses to assess whether the failure of marginally eligible gifted students to benefit from placement in gifted classrooms may be due to mismatch or "invidious comparison" effects. In particular, we examine the impact of placement a gifted classroom on students’ self-report satisfaction with their learning environment by analyzing responses to the following four survey items:

- My teacher(s) believe I can succeed.
- My teacher(s) answer(s) my questions in a way that I can understand.
- I enjoy learning at my school. (2007-2011 only)
- I am accepted and feel like I belong at this school.

Responses are measured using a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Because the distribution of responses among fourth graders is highly skewed (with very few responses of <4) we examine the relationship between IQ scores and the probability that a student “strongly agrees” with each of the four statements. We also analyze a “satisfaction index,” which is the composite average of all four responses.

Appendix Figure A shows the relationship between IQ score or achievement rank and the satisfaction index for the Plan A gifted, Plan B gifted and High Achiever samples (panels A, B and C respectively), while Appendix Table A shows the results from fitting RD models to the responses of each of the four survey items (rows 1-4) as well as the satisfaction index (row 5).[1]

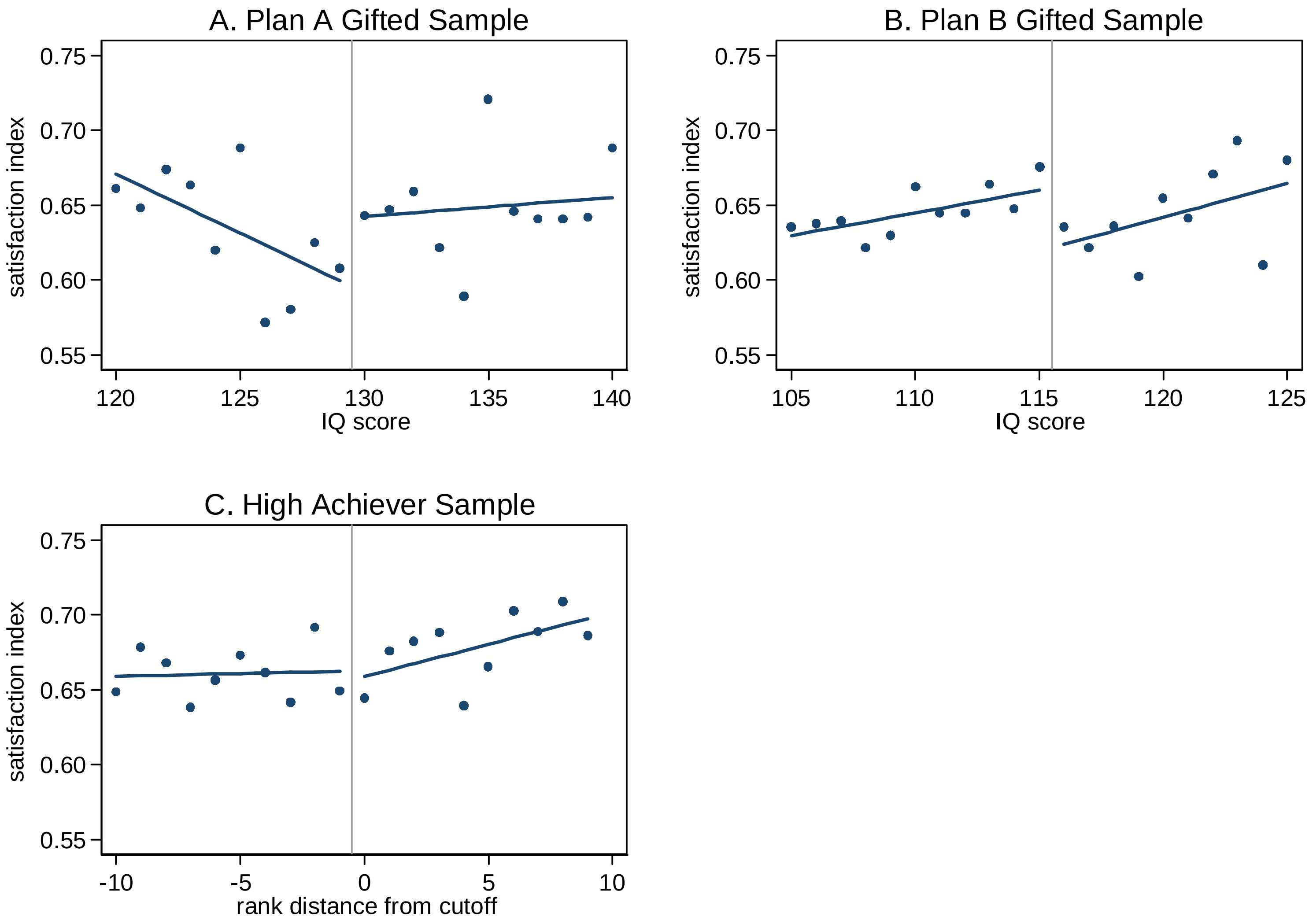
Although the RD estimates are imprecise, the pattern of results is quite different for each of the three analysis samples. In the Plan A gifted sample, all four measures of student satisfaction exhibit *positive* discontinuities at the gifted eligibility threshold (see columns 2-3). These patterns are the opposite of what we would expect if marginally eligible gifted students have difficulty understanding the material or experience invidious comparison effects. Instead, marginally eligible students appear to be *more satisfied* with their learning environment.

The pattern of results from the Plan B analysis sample (columns 5-6) are strikingly different from the Plan A results and point to uniformly *negative* effects of gifted placement on all dimensions of student satisfaction. In particular, the estimates suggest that marginally eligible students are 5 percentage points (or about 10%) less likely to strongly agree that “my teacher answers questions in a way I understand” and 5-6 percentage points less likely to agree that they enjoy learning in their school or feel accepted at their school. These results suggest that the environment of the gifted classroom may not be a good match for marginally eligible Plan B gifted students.

[1] Because not all students completed the survey (the response rate is around 95%) we also test for discontinuities in the response rate (row 6 of Table 3). The estimates are small in magnitude and statistically insignificant.

Finally, the models based on the high achiever sample (Panel C Figure A and columns 8-9 of Table A) suggest that, unlike their Plan A or Plan B counterparts, high achievers who are eligible for placement in the gifted classroom do not differ consistently from those who just miss the cutoff in reported satisfaction with their learning environment. Across the four questions the estimated discontinuities are uniformly close to zero, and the index of all four responses is very smooth at the eligibility threshold. The absence of large effects on satisfaction levels of the high achievers is notable in light of the relatively large gains in peer quality experienced by those who enter a gifted. High achievers' satisfaction levels do not seem to be strongly affected by peer composition, or indeed by any other feature of the gifted classrooms in the District.

Appendix Figure A: Index of self-reported satisfaction based on fourth grade survey responses



Note: Means and fitted values from local linear regressions as in Table A1, row 5, columns 2, 5, and 8.

Appendix Table A. Regression Discontinuity Models for Self-reported Survey Responses

	Plan A Analysis Sample:			Plan B Analysis Sample:			High Achievers Analysis Sample:		
	Mean response (left of threshold)	Local linear RD		Mean response (left of threshold)	Local linear RD		Mean response (left of threshold)	Local linear RD	
		no controls	with controls		no controls	with controls		no controls	with controls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. My teacher(s) believe I can succeed	0.79	0.04	0.03	0.83	-0.03	-0.04	0.82	0.00	0.00
		(0.04)	(0.04)		(0.02)	(0.03)		(0.02)	(0.02)
2. My teacher(s) answer questions in a way I understand	0.56	0.01	0.01	0.64	-0.05+	-0.05	0.62	0.00	0.01
		(0.05)	(0.05)		(0.03)	(0.03)		(0.03)	(0.03)
3. I enjoy learning in my school (2007-11 only)	0.38	0.08	0.08	0.60	-0.04	-0.06	0.56	-0.01	-0.01
		(0.07)	(0.07)		(0.03)	(0.04)		(0.04)	(0.04)
4. I feel accepted/like I belong at my school	0.55	0.08+	0.06	0.61	-0.05	-0.05	0.60	-0.01	-0.02
		-0.05	-0.05		-0.03	-0.03		(0.03)	(0.03)
5. Index (average of nonmissing responses, questions 1-4)	0.59	0.05	0.04	0.67	-0.04+	-0.05*	0.66	0.00	0.00
		-0.03	-0.03		-0.02	-0.02		(0.02)	(0.02)
6. Did not participate in survey	0.05	0.00	0.00	0.05	0.00	0.00	0.05	0.00	0.00
		-0.03	-0.03		-0.01	-0.01		(0.01)	(0.01)

Notes: Table reports estimated mean responses (for students with IQ or rank just below the threshold) and estimated regression discontinuities in responses (=1 if strongly agreed with statement, 0 otherwise) to survey questions administered at the end of fourth grade by the District. Standard errors, clustered by school, in parentheses.

Appendix B: Characteristics of Compliers

Walters (2014) shows that the characteristics of the "compliers" in a fuzzy RD design can be estimated by estimating either of two fuzzy regression specifications, in which the dependent variable is the product of the characteristic of interest and an indicator for either treatment status or non-treatment status. Specifically, using standard treatment effects notation, define $Z_i \in \{0,1\}$ as an indicator for whether the "running variable" X_i exceeds the threshold c (e.g., whether student *i's* IQ exceeds 130 points). Let $D_i \in \{0,1\}$ be an indicator for treatment status (e.g., whether the student is assigned to gifted status), and let $D_i(z)$ denote potential treatment status of individual *i* given $Z_i = z$ -- i.e., the treatment status that would result if individual *i* were assigned the value of $Z_i = z$. Finally, let W_i denote a predetermined covariate (or set of covariates) that is unaffected by D_i or Z_i , and let $g(W_i)$ be some function of the observed covariate(s). Walters defines

$$\gamma_d = \frac{\lim_{x \to c^+} E[g(W_i) \bullet 1[D_i = d] \mid X_i = x] - \lim_{x \to c^-} E[g(W_i) \bullet 1[D_i = d] \mid X_i = x]}{\lim_{x \to c^+} E[1[D_i = d] \mid X_i = x] - \lim_{x \to c^-} E[1[D_i = d] \mid X_i = x]}$$

which is the fuzzy RD estimand for the "outcome" $g(W_i) \bullet 1[D_i = d]$. Under standard conditions (including a monotonicity condition), Walters shows that

$$\gamma_0 = \gamma_1 = E[g(W_i) \mid D_i(1) > D_i(0), X_i = c].$$

Note that individuals with $D_i(1) > D_i(0)$ are the "compliers" whose treatment status switches from 0 to 1 when the instrument switches from 0 to 1. Thus, the fuzzy RD estimands γ_0 and γ_1 provide estimates of the expected value of $g(W_i)$ for compliers with X_i near the cutoff.

To maximize efficiency we implement Walters' method using restricted 3SLS, in which the two "structural" equations of interest are models for the outcomes $W_i \bullet 1[D_i = 1]$ and $W_i \bullet 1[D_i = 0]$, the equations include as controls X_i and $X_i \bullet 1[Z_i = 1]$, and the endogenous variables are the indicators $1[D_i = 1]$ and $1[D_i = 0]$, respectively. The associated first stage equations include the same control variables and use as instruments the indicators $1[Z_i = 1]$ and $1[Z_i = 0]$. We impose the restriction that the estimated RD effects in the two equations (which correspond to estimates of γ_0 and γ_1) are equal.

Appendix Table B presents the estimated mean baseline test scores and mean NNAT scores for compliers in our Plan B sample, and in the sample of disadvantaged high achievers (i.e., high achievers who are either FRL eligible or English language learners) We also show the adjusted means for compliers in the Plan B group, reweighting the distribution of Plan B sample members to have the same representation across schools in the District as disadvantaged high achievers, and the adjusted means for compliers in the disadvantaged high achiever group, reweighting this sample to have the same representation across schools in the District as Plan B group. As can be seen in Figure 2, disadvantaged high achievers are more likely to attend schools with very high FRL rates. Thus, reweighting the Plan B students to have the same distribution as the disadvantaged high achievers tends to slightly lower their average standardized test scores. Conversely, reweighting the disadvantaged high achievers to have the same distribution as the Plan B students tends to slightly raise their average standardized test scores.

The means for the Plan B students suggest that these students score much lower on standardized reading and math tests than would be expected, given their NNAT scores. In particular, their NNAT scores are in (roughly) the top 5% of the overall distribution, whereas their reading and math scores are in the top 25th percentile. By comparison, disadvantaged high achievers have NNAT scores and standardized test scores that are both in the 25-30th percentile ranges.

Appendix Table B: Characteristics of Compliers

	Baseline Reading	Baseline Math	NNAT
Plan B Gifted Students	0.71	0.80	1.63
Reweighted to DHA distribution across schools	0.70	0.76	1.62
Disadvantaged High Achievers (DHA)	0.68	0.57	0.66
Reweighted to Plan B distribution across schools	0.82	0.72	0.77

Note: Table reports estimated means of baseline reading and math scores, and NNAT scores, for compliers in the fuzzy RD design, treating assignment to a gifted classroom as the “treatment” variable for both Plan B students and disadvantaged high achievers.