Can Tracking Raise the Test Scores of High-Ability Minority Students?

David Card and Laura Giuliano*

Abstract: We evaluate a tracking program in a large urban district where schools with at least one gifted fourth grader create a separate "gifted/high achiever" classroom. Most seats are filled by non-gifted high achievers, ranked by previous-year test scores. We study the program's effects on the high achievers using (1) a rank-based regression discontinuity design, and (2) a between-school/cohort analysis. We find significant effects that are concentrated among black and Hispanic participants. Minorities gain 0.5 standard deviation units in fourth-grade reading and math scores, with persistent gains through sixth grade. We find no evidence of negative or positive spillovers on non-participants.

The small fraction of minority students who score in the top percentiles of college entry tests poses a challenge to the U.S. education system (see e.g., Bowen and Bok 1998). Although significant test score gaps have already emerged by age five (Fryer and Levitt 2006), recent studies show that racial disparities in the upper tail continue to widen as students progress through school (Hanushek and Rivkin 2009; Clotfelter et al. 2009). At the same time, blacks and

* Card: Department of Economics; 549 Evans Hall, #3880, University of California, Berkeley, Berkeley, CA 94720-3880, and NBER; (email: card@econ.berkeley.edu); Giuliano: Department of Economics, University of Miami, P.O. Box 248126, Miami, FL 33124-6550 (email: l.giuliano@miami.edu). We are extremely grateful to Cynthia Park, Jacalyn Schulman and Donna Turner 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, Carlos Dobkin, Jesse Rothstein, Enrico Moretti, Chris Walters, five anonymous referees, 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. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110019 to the National Bureau of Economic Research. The opinions expressed are those of the authors and do not represent views of the NBER or the U.S. Department of Education. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

Hispanics are significantly under-represented in advanced academic programs at all levels of K-12 education (US Department of Education 2014). These patterns raise the question: is the low fraction of high-performing minorities at the end of high school due in part to the failure to identify and adequately serve minority students with high learning ability?

One approach for helping high-ability students is tracking – assigning students to different classes based on past achievement (see e.g., Slavin 1987).[1] Despite the widespread use of within-school tracking in the U.S. (e.g., Dieterle et al. 2015), there is no clear consensus in the literature on whether tracking leads to significant achievement gains (Betts 2011).[2] Moreover, in the case of minority students, a particular concern is that any gains for students in upper track classes may be offset by losses for students in lower track classes, where most black and Hispanic children are placed (Oakes 1985). While some older studies suggested that tracking programs harm lower-tracked students, Betts (2011) concludes that the effects in more recent studies are small. Indeed, the most rigorously designed recent study -- a randomized study of first graders in Kenya (Duflo et al. 2011) -- suggests that tracking benefits students in upper *and* lower tracks. Whether this finding generalizes to other settings, however, is still unclear.

In this paper we present new evidence on the efficacy of selective tracking for high achieving students, using data from a unique initiative in one of the nation's largest school Districts ("the District"). In 2004 the District began

[1] Tracking can take various forms: between-school; within-school between classes; and within classes (also known as ability grouping). For simplicity we use the term "tracking" in this paper to refer to within-school, between-class tracking.

[2] As noted by Betts (2011), many studies -- particularly those using U.S. data -- rely on observational designs that are easily criticized. Two recent non-U.S. studies (Duflo, Dupas and Kremer 2011; Vardardottir 2013) use more rigorous designs and find positive effects on upper-tracked students. In a related literature on tracking *between* high schools, a series of recent, carefully designed studies yield mixed results. Several non-U.S. studies find positive effects of gaining admission to schools with higher achievement (e.g., Jackson 2010; Pop-Eleches and Urquiola 2013), while two U.S. studies find negligible impacts (Abdulkadiroglu et al. 2014; Dobbie and Fryer 2014).

requiring schools to establish separate classrooms for any fourth or fifth grade gifted students. Crucially, the extra seats in each class were allocated to *non-gifted* students in the same school who scored highest in statewide achievement tests in the previous year – a group known as high-achievers. Since most schools have only a handful of gifted students per grade, the resulting "gifted/high achieving" (GHA) classes are largely populated by non-gifted high-achievers, and function as upper track classes for students selected on the basis of past achievement. Moreover, because GHA participants are drawn from the same school, and schools in the District are highly segregated by race and socioeconomic class, the program serves many low-income and minority students who typically would be excluded from gifted and advanced academic programs.[3]

We evaluate the effects of the District's tracking program on the non-gifted "high achievers" in two complementary ways.[4] First, we use the eligibility rules for GHA classes to construct regression discontinuity (RD) estimates of the effect of participating in a fourth-grade GHA class relative to a regular class. While the RD estimates are highly credible, their interpretation hinges on whether students who are left behind in regular classes are affected by the presence of a GHA class. Our second approach directly addresses this issue using a between-school/cohort design that compares students in fourth-grade cohorts where there

[3] Minorities tend to be excluded from advanced programs for several reasons: the use of IQ cutoffs, absolute admission criteria, competition across schools rather than within schools, the lack of programs in many low-performing schools, and the reliance on parent and teacher referrals (Donovan and Cross 2002; Card and Giuliano 2014, 2015). In the District, for example, Blacks and Hispanics made up 58 percent of all fourth graders in 2009-2012, but only 40 percent of gifted fourth graders (despite a special program that used universal screening to boost minority referrals). However, because of the GHA program, minority representation in fourth grade GHA classrooms was 50 percent.

[4] We study the program's impact on *gifted* participants in Card and Giuliano (2014). Using an IQ-based RD design, we find no gains to students who are marginally eligible to be classified as gifted. Bui et al. (2014) also find no effects on test scores of gifted and talented programs in another large school district. Dotter (2013), however, finds a significant effect on post-secondary graduation from participation in the gifted and talented program operated in San Diego public schools.

were no gifted children (and no GHA class) to students in other cohorts with between 1 and 4 gifted students (and hence a GHA class with about 20 high achievers). By focusing on students in different rank groups (e.g., 1-20, or 25-44) we can identify both the direct effects of participating in GHA classes *and* potential spillover effects on non-participants.

We reach two main conclusions. First, we find that placement in a fourth-grade GHA class has significant positive effects on the reading and math scores of high achievers, with the gains concentrated among black and Hispanic students. Treatment-on-the-treated estimates for minority students are in the range of 0.5 standard deviation units – comparable to the impacts of "best practice" charter schools (Angrist et al. 2013). The effects for white students, by comparison, are small and insignificant in all our specifications. Importantly, the minority impacts persist to at least sixth grade. Second, we find no evidence of either positive or negative spillover effects on other students in the same school/grade cohort, including those who narrowly miss the cutoff for admission to the GHA class.

The literature suggests a number of channels through which tracking could affect student achievement, including teacher quality, peer composition, and the "match" between student ability and the level of instruction (see e.g., Betts 2011; Duflo, et al. 2011). Since white and minority GHA participants experience very similar changes in teachers and classroom peers, however, the striking absence of any effect on white students suggests that these standard channels are unlikely to explain the large and persistent effects of GHA participation for minorities. This conclusion is confirmed by a direct examination of the effects of changes in teacher quality and average peer characteristics associated with moving from a regular class to a GHA class. Although we find that teacher value added has a large impact on test scores, we show that there is no discontinuity in teacher quality between GHA and non-GHA classes for either whites or minorities. Average peer characteristics like lagged test scores and the fraction of females

change significantly between regular and GHA classes but we find that the impacts of these characteristics are relatively small for both whites and minorities, and explain only a small fraction of the achievement gains experienced by minority students in GHA classes. The absence of large effects from these peer characteristics is also consistent with our finding of no spillover effects on non-GHA participants.

Instead, we hypothesize that higher-ability minority students face obstacles in the regular classroom environment that cause them to underperform relative to their potential, and that some of these obstacles – including low teacher expectations and negative peer pressure – are reduced or eliminated in a GHA class. We show that minority students have lower achievement scores than white students with the same cognitive ability, and that placement in a GHA class effectively closes this minority under-achievement gap. A mediating role for teacher expectations is consistent with evidence that shows teachers systematically under-refer black and Hispanic students for the District's gifted program (Card and Giuliano 2015). Further, an analysis of unexcused absences and suspensions suggests that the impacts of GHA placement are partly mediated through changes in student behavior and reflect a more supportive environment for high achieving minority students.

I. Background, Research Design, and Analysis Samples

A. *Background*

In 2004 the District introduced a new policy requiring schools to offer separate classrooms for fourth or fifth grade if there was at least one gifted student in the school. Because of the strict IQ thresholds for gifted status imposed by state law, a typical elementary school in the District had only 5 or 6 gifted

children per grade, with even fewer at the schools in poor neighborhoods.[5] Two features of the policy, however, meant that it effectively created a broader, within-school tracking program. First, the classes were required to be the *same size* as the regular classes for that grade (20-24 students). And second, any open seats in the classes were to be filled by *non-gifted* students at the school with the highest scores on the previous year's standardized tests – i.e., the high achievers that are the focus of this paper.

Figure 1 shows the distribution of fourth-grade school/cohorts in the District grouped by the number of gifted children in the cohort. We also superimpose the fractions of GHA participants who are gifted or high achievers, cross classified by minority status. Schools with relatively few gifted students per cohort (on the left side of the graph) are mainly located in poor neighborhoods; their GHA classes are mostly filled by minority (i.e. black or Hispanic) high achievers.[6] Schools with more gifted students per cohort (on the right side) are typically located in richer neighborhoods and have more white students in GHA classes. Even at these schools, however, there are many high achievers in GHA classes because the schools often set up two GHA classrooms.

[FIGURE 1 ABOUT HERE]

GHA teachers are drawn from the regular teaching staff: about 70 percent are observed teaching a regular (i.e., non-GHA) fourth-grade class at some point in our sample period. GHA teachers must complete a set of five courses on gifted education, but they receive no extra salary.[7] Students in GHA classes are

[5] State law requires regular students to have an IQ score of 130 or higher to be eligible for gifted status (2 standard deviations above the national mean). A lower 116 point threshold applies to English Language Learners and participants in the federal free and reduced price lunch program.
[6] We henceforth use the term "minority" to refer to students who are black or Hispanic; Asians and other non-white minorities make up less than 6 percent of students in the District.
[7] The courses are also required for gifted endorsement by the state and are offered on line by the District. These courses focus on issues associated with teaching gifted students, though as discussed in Section IV, they may also have some impact on instruction for minority high achievers.

responsible for mastering the same statewide curriculum as those in regular classes, and are evaluated using the same statewide tests. Accordingly, GHA classes also use the same textbooks as regular classes, and are responsible for covering the same amount of the textbook in a school year. Teachers in both GHA and regular classes are free to divide their class into subject-specific ability groups and often do so.

The main difference between GHA and regular classes, interviews and policy documents suggest, is that teachers in GHA classes have greater flexibility regarding the pace of instruction and curricular enhancement.[8] Because state policy requires that gifted students in the classroom receive differentiated instruction, GHA teachers are encouraged to cover the standard curriculum at a faster pace and to use the extra time to explore topics in depth or for other curricular enrichment. While there is no systematic information on the differences in pacing or enrichment, it is important to note that both the rate of pacing and the nature of enrichment activities are left to the discretion of GHA teachers and that they vary widely across teachers.

B. *Causal Mechanisms and Research Design*

Previous research on tracking programs (Betts, 2011) and the determinants of student achievement (e.g., Betts, Zao and Rice 2003; Rivkin, Hanushek and Kain 2005) suggests there are multiple channels through which GHA classes could affect student achievement. One channel is teacher quality: if more effective teachers are assigned to GHA classes, students will make bigger achievement gains in these classes.

[8] Our description of the GHA program is based on state and District policy documents, discussions with the District's program director and two gifted coordinators, and on interviews with 30 teachers at 10 schools across the District. All 30 teachers had significant experience teaching both GHA and regular classes. The least experienced had taught for two years in GHA classes and six years total.

A second channel is through the "match" between students' prior knowledge or ability and the level of instruction (Duflo et al. 2011). If teachers choose a level of instruction appropriate for the middle of their class, the median ability student in a GHA class may learn faster than she would in a regular classroom. [9] For marginally eligible students, however, there is a risk of mismatch if the GHA teacher aims too high.[10] Lower-ability students may also benefit from improved matching when a GHA class is established, though in our setting any such effect is limited by the small share of students moved to GHA classes (20 percent in a typical school with 5 fourth-grade classes).

A third channel is through peer characteristics. A standard assumption is that students are positively influenced by the mean ability of their classmates (e.g., Sacerdote 2011). In this case the establishment of a GHA classroom would be expected to raise the scores of higher-achieving students (who are now grouped together) and lower the scores of those who remain in regular classes. Other changes in peer composition could produce similar effects. Such changes include a reduced presence of disruptive students (Lazear 2001; Carrell and Hoekstra 2010) or a larger fraction of girls (Hoxby 2000; Lavy and Schlosser 2011). Again, the spillover effect on students in regular classes may be relatively small in our setting given the modest fraction of students who are shifted to a GHA class. Moreover, recent findings suggest that the link between mean peer characteristics and average achievement can easily break down in specific settings.[11]

[9] The choice of what students to target in a given class is endogenous, and could be affected by institutional pressures, making precise predictions difficult in the absence of information on the target levels actually used by teachers. Another factor is ability grouping, which is widely used in both regular and GHA classes in the District, and arguably improves the match between teaching levels and student abilities even in heterogeneous classes.

[10] Such concerns are often expressed in studies of programs to expand minority representation in elite universities (e.g., Arcidiacono et al., 2014).

[11] Two recent, carefully designed studies of between school tracking (Abdulkadiroglu et al. 2014 and Dobbie and Fryer 2014) find negligible impacts from large changes in peer characteristics for

Peer composition may also affect achievement through a "rank" effect. Hoxby and Weingarth (2005) suggest that the lowest-ranked students in a class may suffer from invidious interpersonal comparisons. Such comparisons could affect marginally eligible GHA participants and push down the estimated effects from an RD design, similar to the predicted pattern arising from mismatch of student ability and the teacher's targeted level of instruction. Murphy and Weinhardt (2014) propose a related "top of the class" effect which predicts that the highest ranked students who remain in a regular class when a GHA class is established will perform better, again attenuating the impacts from a RD design.

Finally, the GHA classroom environment may be uniquely beneficial for high-ability minorities if it removes obstacles that cause these students to under-perform in the regular classroom. One possible source of underperformance is negative peer pressure. Fryer and Torelli (2010) argue that high-achieving black students – particularly boys – suffer peer sanctions if they perform well in class, while Bursztyn and Jensen (2015) find that peer pressure against "academically oriented" choices is reduced when students are in classes designated as "honors." Under-performance may also be reinforced by lowered teacher expectations. In a companion paper, Card and Giuliano (2015) find that minority students with high IQ but modest achievement were systematically under-referred for the District's gifted program when the referral process relied heavily on teacher nomination—suggesting that teachers often fail to recognize the potential of minority students who under-perform in class.

In view of the wide variety of potential channels that could arise when high achieving students are placed in separate classes, we present two

---

students assigned to elite high schools. In the college context, Booij et al. (2015) find that tracking improves the outcomes of lower-achieving students, and attribute the effect to richer peer interactions among the students in lower track classes. Carrell et al. (2013) find that lower-achieving Air Force Academy students do *worse* when assigned to squads with more high-achieving peers.

complementary research designs for analyzing the effects of fourth-grade GHA classrooms. First, we use a regression discontinuity approach based on eligibility rules for the fourth-grade GHA class. This approach is highly credible (DiNardo and Lee 2011) and leads to relatively precise estimates. We use this design to estimate separate impacts for white and minority students, and to assess the mediating role of several factors that could differ between the regular and GHA classes—including teacher quality and peer characteristics.

Despite the appeal of an RD design, the interpretation of the resulting estimates depends critically on whether there are spillover effects on students who narrowly miss the GHA cutoff.[12] In our second "between-school/cohort" design, we ask how specific rank groups (e.g., students ranked 15-20) perform when there is or is not a GHA classroom available for their cohort. Specifically, we focus on school/cohorts with between 0 and 4 gifted students. As shown in Figure 1, most of the top-ranked students in these school/cohorts are black or Hispanic, so this design is informative about GHA impacts on minority students. Using this design we estimate effects for different rank groups that are likely to participate in a GHA class if one is offered, which allows us to assess "rank effects" and also provides a second estimate of the impact of participating in the class. Moreover, we also estimate effects for rank groups that are likely to just miss the cutoff for a GHA class if one is offered—providing direct evidence on spillover effects.

[12] In general, the RD design identifies the difference in the effect of the presence of a GHA class on marginally eligible and marginally ineligible students. To see this, note that (up to a scale factor reflecting the first-stage effect) the RD estimator identifies $\tau = \lim_{r\downarrow 0} y(r) - \lim_{r\uparrow 0} y(r)$ where y(r) is the expected achievement of students at rank r when a GHA class is present, and r=0 is the cutoff threshold. Decompose $\tau = \tau_1 - \tau_0$ where $\tau_1 = \lim_{r\downarrow 0} y(r) - y^n(0)$, $\tau_0 = \lim_{r\uparrow 0} y(r) - y^n(0)$, and $y^n(0)$ is the expected achievement of students of rank 0 in the absence of a GHA class (which is the same for marginally eligible and ineligible students). τ_1 is the treatment effect on barely eligible participants relative to the counterfactual of no GHA, and τ_0 is the corresponding treatment effect on barely ineligible non-participants.

C. *Analysis Samples*

Our analysis is based on administrative data for students who completed third grade in the years from 2008 to 2011; entered fourth grade the next year at one of the District's 140 larger elementary schools; and remained in the District through at least the end of fourth grade.[13] Our data set includes age, gender, race, ethnicity, home zip code, eligibility for a free or reduced-price lunch (FRL), English language learner (ELL) status, and student scores on state-wide reading and math tests (administered at the end of third through tenth grades), writing (fourth grade), and science (fifth grade). We also have scores from the Naglieri Nonverbal Ability Test (NNAT), an IQ-like test that was administered to second graders in the years from 2005 to 2009 as a screening test to identify potentially gifted students (see Card and Giuliano, 2015).

Table 1 shows characteristics of all third graders in the District from the included school/cohorts (column 1), as well as characteristics of students in our RD analysis samples (columns 2-4) and our between-school analysis samples (columns 5-6). As shown in column 1, the District is highly diverse, with 28 percent white non-Hispanics, 39 percent black non-Hispanics, 27 percent Hispanics, and 3 percent Asians. Roughly half of third graders are eligible for free or reduced price lunches and 10 percent are English language learners. The mean score on the NNAT test (for students that took the test) was 105 – slightly above the nationally normed mean of 100. Overall, about 6 percent of students were classified as gifted by fourth grade and 13 percent were assigned to a GHA classroom.

[TABLE 1 ABOUT HERE]

[13] We exclude students who were in fourth grade at small schools (including charter schools) since these schools do not have enough students to create both a GHA and regular class. However, we continue to follow students in fifth and sixth grades as long as they remain at any District school. As discussed below, we find no evidence that GHA classroom assignment affects attrition from the District either by the end of fourth grade or beyond.

The middle rows of Table 1 report mean scores on third- and fourth-grade statewide achievement tests; mean third-grade scores of students' school-wide peers; and mean third-grade scores of participants in their schools' fourth-grade GHA classrooms (if one was created).[14] Following standard practice we standardize the achievement scores to have mean 0 and standard deviation 1 within each grade/year cohort. Finally, in the bottom rows of the table we report some characteristics of the schools attended by students in our analysis samples. A typical school has 4-6 fourth-grade classes, one of which is a GHA class, with 23-24 students per class.

To construct our RD analysis sample, we first identified fourth graders who had test scores from third grade and who were in school/cohorts with a fourth-grade GHA class. Under the District's rules, open seats in the GHA classroom are supposed to be filled by non-gifted students in the same grade with the highest scores on the previous year's statewide tests, using a specific formula to combine math and reading scores. The formula was only introduced in 2009 so we limit our entire analysis to cohorts in third grade in 2008 or later. A few schools also appear to ignore the District formula, so we exclude students at these schools. Using the District formula and an algorithm described in Appendix A, we then calculated school/cohort-specific cutoff scores for admission to the fourth-grade GHA class and selected the first 10 students with scores above this cutoff and the first 10 with scores below it.[15] We explore the use of a wider range of ranks below.

We emphasize that the ranks used to select the sample are based on the *District formula* and not on students' actual placement in the GHA class (which

[14] For all students (column 1 of Table 1) the mean characteristics of school wide peers and class peers are equal to the mean characteristics of all students in the District, so we do not display these.

[15] The number of students above the cutoff is less than 10 if the gifted class has <10 extra seats. Appendix A also compares the algorithm we use to find the cutoff to three alternative methods.

can deviate from the formula). In total we have 4,144 students who are ranked within 10 places from the estimated school/cohort cutoff. About two-thirds are in school/cohorts with 5 or more gifted students (column 3); the others are in school/cohorts with 1-4 gifted students (column 4).

Relative to the all third-grade students in the District, those ranked within 10 positions of the GHA cutoff (column 2) are more likely to be white, less likely to be FRL-eligible, and have higher test scores. They also have slightly better-than-average school wide peers (reflecting the fact that students at schools with no GHA class are excluded). Highly ranked students from schools with a larger number of gifted students (column 3) are even more selected, while those from schools with fewer gifted students (column 4) are closer to the District-wide average.

For our between-school analysis sample (columns 5-6) we focus on students from school/cohorts with 0 to 4 fourth-grade gifted students, and we use the District formula to construct their within-school/cohort ranks. Students ranked 1-20 (column 5) are likely to move to a GHA class if one is offered, while those ranked 25-44 (column 6) are likely to stay in a regular class regardless of the availability of a GHA class. Notice that even the highly ranked students in these school/cohorts are over 70 percent black or Hispanic, and have an average FRL rate of over 70 percent, reflecting the concentration of schools with few gifted children per grade in lower-income neighborhoods.

II. RD-Based Analysis of GHA Participation

A. *Validity of RD Design*

In this section we present an RD-based analysis of the effect of GHA participation on non-gifted high achievers. Figure 2 shows the frequency distribution of within-school/cohort ranks for fourth-grade cohorts with a GHA

class. The distribution is smooth around the cutoff rank for GHA eligibility, as would be expected in a valid RD design (Lee, 2008; McCrary, 2008). The fall-off in frequencies to the right of the cutoff arises because some school/cohorts have fewer than 10 open seats for high achievers in the GHA class.

[FIGURE 2 ABOUT HERE]

Further evidence on the validity of our design is presented in Figure 3, which shows the relationship between a student's relative rank and four key variables: their third-grade reading and math scores (panels A and B); their NNAT test scores (panel C); and their predicted average fourth-grade reading and math scores, estimated from a regression model that includes third-grade scores, age, race, gender, FRL and ELL status, and school dummy variables (Panel D). All four panels show a smooth evolution through the cutoff rank, confirming that students just above and just below the cutoff are very similar. More formal tests are presented below in Table 2.

[FIGURE 3 ABOUT HERE]

B. *Differential Attrition*

To be included in our analysis sample, a student must be enrolled in a regular District elementary school at the beginning of fourth grade and remain at *any* District school through the end of the school-year when the outcome is measured. A threat to our design could arise if students assigned to GHA classes are more (or less) likely to remain in the District than those in regular classes.[16] We address this concern in Appendix Figure 1, which shows the relationship between a student's relative rank at the start of fourth grade and the probability of remaining in the District through the end of fourth, fifth or sixth grade (Panels A, B, and C respectively). We find no evidence of discontinuities in the retention

[16] Davis et al. (2013) find that students eligible for the gifted program in a mid-western school district are more likely to stay in the district.

rate. Based on these comparisons, and on results from a series of RD models, we conclude that differential attrition is not a concern.

C. *First-Stage Relationship and Impacts on Fourth-grade Achievement*

Figure 4 plots the "first-stage" relationship between student relative ranks and the probability of placement in a GHA class. The graph shows a fuzzy discontinuity in the relationship between rank and placement in a GHA class, with a 25 percent placement rate just to the left of the threshold and a 60 percent rate just to the right. Inspection of the data leads us to conclude that the fuzziness is attributable partly to non-compliance with the District formula. In particular, a pattern of increasing compliance over time suggests that enforcement was initially weak (see Appendix A). The remaining fuzziness is due to data issues (e.g., missing test scores and slippage in our creation of class rosters), which causes some students to be misclassified and may also introduce measurement error in the school-specific cutoffs. As explained in Appendix A, our procedure for calculating the cutoffs is designed to avoid creating spurious first-stage discontinuities or biases in the reduced-form RD estimates. The very smooth patterns in Figures 2 and 3 support the validity of our procedure. We suspect, however, that slippage in the measurement of GHA participation leads to some attenuation in the magnitude of the first-stage discontinuity – on the order of 10-15 percent – that should be taken into account in interpreting our two-stage least squares estimates (see Appendix C for a formal development).[17]

[17] We believe that misclassification errors are most likely when there are more students with missing third-grade test scores who are assigned to the GHA class, and when the *estimated* size of the GHA class is over 24 students (the legal maximum). We therefore fit first-stage and reduced-form models allowing interactions with the number of student with missing scores, and the number of students in excess of 24 assigned to the class. The implied first-stage discontinuity in GHA participation is 0.36 for school cohorts with no missing scores and a GHA class size $\leq$24 (versus 0.32 in a simpler specification with no interactions). Both interaction terms are also negative and significant, and imply a reduction of about 0.01 in the size of the first-stage effect per student with missing scores, and .03 per student in excess of 24 in the GHA class. Importantly, when we add the same interactions to the reduced-form models neither is large in size or close to statistically significant, implying that there is no induced bias in the reduced-form effects.

[FIGURE 4 ABOUT HERE]

Figure 5 shows how fourth-grade test score outcomes vary with a student's relative rank within their cohort and school. Reading scores (panel A) and math scores (panel B) show clear jumps at the cutoff rank, indicating a positive impact from assignment to a GHA class. In contrast, there is no evidence of an effect on writing. Panels D and E show the average test score *gains* in reading and math, formed by subtracting each student's third-grade score from his or her fourth-grade score. (There is no third-grade writing test so we cannot construct a change in writing scores). The downward-sloping pattern in these graphs is driven by mean regression in test scores: the highest scores in any year incorporate measurement errors and "good luck" that do not persist to the next year, so test score gains are negatively correlated with rank. Relative to this overall trend, however, the differenced models show clear discontinuities at the GHA threshold, similar in size to the discontinuities in the corresponding test score levels.

[FIGURE 5 ABOUT HERE]

Table 2 presents the estimated discontinuities in baseline test scores (columns 1-2), as well as the first-stage discontinuity in the probability of placement in a GHA class (column 3) and the associated reduced-form discontinuities in fourth-grade reading, math, and writing (columns 4-6). All the models are "local linear" RD specifications with a bandwidth of 10 ranks to the right and left of the threshold for admission to the GHA class. (We discuss alternative bandwidths below.) Row 1 presents basic specifications with no other controls; the models in row 2 include a broad set of additional controls (including school effects); and row 3 presents first differenced specifications.

[TABLE 2 ABOUT HERE]

The baseline test score models (in columns 1 and 2) show only small discontinuities at the threshold rank, consistent with the smooth patterns noted in

Figure 3. The first-stage models in column 3 show a precisely estimated jump of about 33 percentage points in the probability of placement in the GHA class at the threshold. Reassuringly, the magnitude of the jump is virtually identical in specifications with and without student controls and school dummies—again minimizing concerns about possible bias resulting from our method for calculating the thresholds.

The reduced-form models for reading and math in columns 4 and 5 show relatively precisely estimated discontinuities with magnitudes of 0.07 to 0.11 σ's. Taking account of the fuzzy first stage, the implied treatment-on-the-treated effects are in the range of 0.3 σ's. (Corresponding two-stage least squares estimates are reported in the top row of Table 3.) In contrast, the impacts on writing achievement are small and insignificant.

Appendix Figure 2 shows the robustness of the reduced-form impacts to alternative bandwidth choices, using the specification in row 2 of Table 2. Across a range of bandwidths from 5 to 15, the estimated impacts on reading and math are quite stable and remain significant at conventional levels, while the estimated impacts on writing are uniformly small.

D. *Heterogeneity in Impacts*

The results in Figure 5 and Table 2 suggest that participants in GHA classes experience significant average achievement gains in reading and math. A key question is whether these gains are similar for different groups of students. We study this issue in Table 3. Each row presents results for a different subgroup. We show the estimated discontinuities in their baseline test scores in columns 1-2; the first-stage discontinuity in their probability of placement in a GHA class in column 3; and the implied effects on their reading and math scores in columns 4-7. For ease of comparison across groups with different first-stage discontinuities, we report two-stage least squares estimates of the treatment-on-

the-treated effects for the outcomes.[18]

For reference, row 1 of the Table presents results for our overall sample. On average, entry to a GHA classroom is associated with gains in fourth-grade reading and math scores of about 0.3 σ's per treated student. Row 2a and 2b present results for white students and minority (black or Hispanic) students. The first-stage models in column 3 show that the two groups have similar jumps in the probability of placement in a GHA class at the threshold rank. The impacts on achievement, however, are quite different: there are small and insignificant effects for whites, but large positive effects for minorities. This difference is confirmed visually in Figure 6, where we show reduced-form plots of average reading and math scores for the two groups. The plots for white students show no evidence that placement in a GHA class matters. The plots for minorities, in contrast, show clear jumps at the GHA threshold.[19]

[TABLE 3 ABOUT HERE]

[FIGURE 6 ABOUT HERE]

Rows 2c and 2d of Table 3 present separate estimates for black and Hispanic students. Though the estimates are a little imprecise, they are very similar—leading us to conclude that the benefits from participating in a GHA class are about the same for blacks and Hispanics, and together account for virtually all of the average effect observed in the pooled sample.[20]

---

[18] As noted above attenuation of the first stage due to misclassification may cause the 2SLS estimates to overstate the true TOT effect by about 10-15 percent. Our analysis suggests that the first-stage attenuation is only slightly larger for minorities than for whites (15 percent vs. 11 percent), consistent with greater mobility across schools and classrooms (the main sources of misclassification) among minorities.

[19] One concern is that gains for high achieving whites may be attenuated by "topping out" on the standardized tests. In fact only 2 percent of students achieve the top score in reading, though in math 10 percent are topped out. Estimates from Tobit models (reported in Appendix Table 1) are very similar to the models in Table 3 and show no indication that censoring accounts for the small impacts for whites.

[20] Blacks and Hispanics account for 56 percent of all students in our main RD sample. The 2SLS impacts in row 1 are very close to 0.56 times the corresponding impacts in row 3.

One concern with the comparison in rows 2a and 2b is that the minority students in our RD sample are drawn from schools with fewer gifted students and higher fractions of FRL-eligible students. This raises the question of whether part of the difference across groups is attributable to some factor other than race/ethnicity. In rows 3a and 3b of Table 3 we compare the impacts of participating in GHA classes for black and Hispanic students who are FRL-eligible and ineligible. The estimated treatment-on-the-treated impacts are very similar for the two groups. In rows 4a and 4b we compare the estimates for minorities in cohorts with relatively few (1-4) or relatively many (5+) gifted children. Again, the treatment-on-the-treated impacts are similar.

Finally, rows 5a and 5b present separate results for minority girls and minority boys. Overall, the impacts appear to be larger for boys than girls, particularly in reading. The large readings gains for boys are interesting because minority boys tend to have particularly low average reading scores.[21] Nevertheless, the gender-specific samples are relatively small, and one might be concerned about multiple testing. At a minimum it seems safe to conclude that GHA classes are, if anything, more effective for minority boys than minority girls.

We have also fit a parallel set of models for subgroups of whites. The resulting estimates are all small in magnitude and uniformly insignificant. Even for lower-income white students who are FRL eligible we find no evidence of an effect of participating in a GHA class. Hence we conclude that race/ethnicity – not family income – is the fundamental dimension of heterogeneity in the GHA treatment effect.

[21] Among minority students ranked 5 points below the cutoff, the average reading score for boys is .38σ while for girls it is .54 σ. The average math score is .43σ for both these groups.

E. *Impacts on Later Grades*

Next we look at the impacts of participating in a GHA classroom in fourth grade on achievement scores in fifth and sixth grades.[22] Table 4 presents 2SLS RD estimates for the impacts on fifth and sixth grade reading and math, and fifth-grade science. We follow the same format as Table 3, presenting results for the overall sample in row 1, and for various subgroups in later rows of the table. For the overall sample we find mixed evidence that fourth-grade GHA participation affects fifth-grade test scores. The impacts on reading are positive but modest in size and statistically insignificant, while the effects for math are larger and marginally significant. For sixth grade, however, the results are more clearly positive, with marginally significant impacts of about 0.2σ's for reading and significant impacts of about 0.4σ's for math.

[TABLE 4 ABOUT HERE]

Rows 2a and 2b present separate results for whites and minorities. Consistent with the findings in Table 3, the effects of a fourth-grade GHA class are relatively small (and mainly negative) for whites, but the effects for black and Hispanic students are positive and, at least in sixth grade, statistically significant. The impacts for minorities are illustrated in Appendix Figure 3, which shows the first-stage relationship between GHA placement and relative rank for minority students who remained in the District through sixth grade, as well as reduced-form graphs of their NNAT scores and combined math and reading scores in third through sixth grade. The pattern of sixth-grade scores is particularly suggestive of a persistent effect of admission to the fourth-grade GHA class. Reassuringly, we

[22] A possible concern about effects on later grades is differential attrition out of the District. Appendix Table 1 presents RD models for the probability that students remain at a District school through the end of fifth grade (column 1) and through the end of sixth grade (column 4). Consistent with the visual evidence in Appendix Figure 1, we see no indication of a discontinuity in longer-run attrition rates at the threshold for admission to a GHA class in fourth grade.

continue to find no evidence of discontinuities in either the second-grade NNAT or the third-grade reading and math scores in this sample.

The estimates in the later rows of Table 4 suggest that, as in Table 3, the longer run impacts of fourth-grade GHA participation are not too different for FRL-eligible and ineligible minority students, or for those in schools with more or less gifted students. Comparisons by gender also reinforce the conclusion from Table 3 that GHA participation has a larger effect for minority boys than for minority girls.

One explanation for the persistent effects of fourth-grade GHA classes is that students who are placed in a GHA class in fourth grade are more likely to be placed in advanced classes in later grades. The District offers GHA classes in fifth grade with the same admission rules as the fourth-grade classes, and advanced-track sixth-grade classes for math where students are admitted on the basis of fifth-grade standardized scores. High achievers who participate in a fourth-grade class are not guaranteed a seat in the fifth-grade class, but if the fourth-grade GHA class causes large achievement gains there will be an induced discontinuity in the probability of fifth-grade placement (and likewise for sixth-grade advanced math).

In Appendix Table 2 we show that there are in fact significant discontinuities in the probability of placement in a fifth-grade GHA class and in a sixth-grade advanced math class at the threshold for admission to the *fourth-grade* class.[23] As expected, these jumps are driven entirely by outcomes for black and Hispanic students: there are no discontinuities in placement for white students. Even for minorities, however, the jumps are relatively small, suggesting that participation in later advanced classes is not the main explanation for the

[23] The analysis sample for placement in a fifth-grade GHA class is reduced by roughly 30 percent due to our inability to match students to classrooms at schools where students rotate between teachers in fifth grade (see Appendix A for details). For this reason, we are also unable to directly estimate the effect of GHA assignment in fifth grade.

persistent effect of fourth-grade GHA classes.[24] Moreover, the advanced class in sixth grade is only for math, and we think it is unlikely to have a major effect on sixth-grade reading scores. Instead, we interpret the impacts in later grades as arising mainly from the fourth-grade GHA class itself, either through a dynamic "learning begets learning" process, or through behavioral or other changes attributable to the GHA class that raise students' productivity at school.

F. *Causal Mechanisms*

In this section we ask to what extent the impacts of fourth-grade GHA participation can be explained by the quality of GHA teachers and/or the characteristics of peers in GHA classes. We follow a two-step approach. First, using data on all fourth-grade classes in the District we estimate a series of value added (or "gain score") models in which the dependent variables are test score gains between third and fourth grade, and the explanatory variables include one of six class features—namely, a measure of value added for the teacher or one of five peer composition variables.[25] We estimate separate models for each class feature, as well as a joint model that combines all six features. We also compare pooled models for all students with separate models for whites and minorities to check for heterogeneity in the effects of the various features. These gain score

[24] We also investigated whether admission to a fourth-grade GHA class affects the probability of remaining at a regular District elementary school (versus a charter or Montessori school) in fifth grade. We find a small but statistically insignificant discontinuity in this probability among minority students, with those who just miss the cutoff being 2 percentage points more likely to transfer to a non-traditional school. If the alternative schools are of higher quality than the non-GHA classrooms in the schools these students would otherwise attend, this would tend to reduce our RD estimates for fifth-grade outcomes.

[25] These models also include controls for individual student characteristics and school fixed effects. Teacher value-added is estimated for each teacher in a given year using data in *other* years when the teacher is observed teaching fourth-grade in the District, including the four years prior to our sample (i.e., 2005-2008). Appendix B describes our model in more detail. In brief, we regress fourth-grade scores (for students in other years) on lagged test scores, individual student characteristics, school-wide average demographics, a dummy for GHA class, and teacher dummies. We use the estimated teacher effects (with no "shrinkage adjustment") as measures of value added.

models will provide unbiased estimates of the effects of teacher value added and peer characteristics under the same assumptions that are widely adopted in the "teacher effects" literature (e.g., Chetty, Friedman, and Rockoff, 2014). [26] As noted by Rothstein (2015), these assumptions are unlikely to hold exactly, particularly in models that examine only one class feature at a time, but we suspect that violations are likely to lead us to overstate the causal effects of the various class features.

In the second step, we re-estimate the RD models in Table 3 including teacher value-added (TVA) or a peer composition variable, but restricting the coefficients to the values estimated in the first stage. The estimated RD coefficients from these second step models show the impact of GHA classes net of the effect of any differences in teacher or peer characteristics between regular and GHA classes.

We use this approach – instead of simply adding teacher or peer characteristics to the RD models – for two reasons. First, data for the overall population of fourth-grade students yield much more precise estimates of teacher and peer effects. Second, in a fuzzy RD design with incomplete compliance, a direct approach can yield biased estimates for the control variables because teacher and peer characteristics are endogenously determined by the actual assignment of students to classes.[27] Average peer test scores, for example, will be much higher for students who are placed in a GHA class, regardless of whether their rank is above or below the GHA threshold, so this variable is correlated with compliance status.

For a classroom feature to explain part of the measured effect of GHA

[26] Formally, the required assumption is that the unobserved determinants of a student's test score gain between third and fourth grades are uncorrelated with the measured class features, conditional on the observed control variables, which include school dummies and age, race, gender, FRL and ELL controls.

[27] See Rosenbaum (1984) for an analysis of this problem in the context of experimental and observational designs.

classes, two things must be true. First, there must be a discontinuous change in the average value of the characteristic between students who are barely eligible for a GHA class and those who are barely ineligible. Second, the characteristic must be causally associated with higher test scores. Evidence on the first issue is presented in Figure 7, which shows the relationship between a student's relative rank and each of the six features of the student's classroom: mean third-grade test scores of classroom peers (panel A); the within-class standard deviation in third-grade scores (panel B); the fraction of classroom peers who were suspended in third grade (panel C); the fraction of female peers (panel D); the fraction of minority male peers (panel E); and teacher value added (panel F).

[FIGURE 7 ABOUT HERE]

As expected given the admission process for GHA classes, mean lagged test scores of classroom peers jump up at the cutoff rank, while the dispersion in peer scores falls. There is also a small reduction in exposure to peers who were suspended in the previous grade, a small increase in the fraction of girls in the class, and a modest reduction in the fraction of minority boys in the class. In contrast, estimated TVA is smooth through the cutoff.

Appendix Table 3 presents 2SLS-RD models for the changes in these characteristics at the threshold for admission to a GHA class for our overall sample and for the subgroups identified in Tables 3 and 4.[28] As suggested by the graphs in Figure 7, we find that TVA evolves smoothly at the GHA threshold for

[28] Because we measure TVA using data on students taught by a given teacher in *other* years, we cannot assign TVA to students whose teachers appear in the District in only one year. Rothstein (2015) shows that treatment of missing data can affect the validity of TVA estimates. To assess whether selective availability of TVA is an issue for our analysis, we show RD estimates for the probability of having an estimate of TVA of in column 1 of this table; apart from one subgroup, none of these estimates is large or significant. We conclude that missing TVA data is unlikely to be a problem for our analysis, and we focus on the subset of students in our RD analysis sample with a TVA estimate for their fourth-grade teacher. This choice allows us to hold constant the sample as we compare models that adjust for the effects of different classroom features, though our conclusions are quite robust to allowing for different samples in different models.

all students in our sample. The treatment-on-the-treated (TOT) estimate is -0.01 with a standard error of 0.03, ruling out discontinuities in TVA larger than +0.05. Compared to the TOT estimates for fourth-grade math and reading of about 0.30 (row 1 of Table 3) this is a small effect. Importantly, we find small and insignificant changes in TVA at the GHA threshold for both whites and minorities, and for various subgroups of minorities. These patterns mean that teacher quality (as measured by TVA) cannot be the primary explanation for the effect of GHA classes.

In contrast, we find that the five peer composition variables all change at the GHA threshold. The largest change is for mean lagged test scores: the TOT estimate is 0.86 for all students, 0.88 for white students and 0.83 for black and Hispanic students. In fact, we cannot reject an effect of 0.85 for any subgroup, suggesting that the impact on peer lagged achievement is relatively homogeneous. The effects on the other characteristics are smaller in magnitude, with some evidence of heterogeneity across subgroups in the fraction of classmates who were suspended, and in the fraction of minority male classmates.

How much do the various classroom features matter for student achievement gains? Coefficients from student achievement models fit to data on fourth graders in the larger elementary schools in the District are presented in Appendix Table 4. We show separate estimates for reading and math for all students, white students, and minority students. We also compare coefficient estimates from value-added specifications that include each classroom feature separately, and from a specification that controls for all six features at once.

Consistent with the large literature on teacher effects, we find that the most important classroom feature is teacher quality. A one-unit increase in estimated TVA is associated with a 0.4 increase in fourth-grade reading scores,

and a 0.6 to 0.7 increase in math scores.[29] The effects on both domains are highly significant and very similar in magnitude for white and minority students. In contrast, the effects of measured peer characteristics are much smaller, particularly for reading achievement. For math achievement, both the mean and standard deviation of lagged peer test scores appear to matter, with coefficients of 0.08 to 0.12 for mean lagged peer scores, and similar coefficients for the standard deviation of scores.[30] The fractions of female peers and minority male peers have very small and insignificant effects for both reading and math scores of all groups.

With these estimates in hand, we turn to the adjusted RD models in Table 5. The top row of the table shows estimated reduced-form impacts of participating in a GHA class on fourth-grade reading and math scores from our basic RD models, using the specification in row 3 of Table 2. We show effects for all students and for white and minority students separately. Each of the other rows reports the same RD coefficient from specifications that control for the classroom feature indicated in the row heading, using estimates of the effect of this feature reported in Appendix Table 4. Finally, the bottom row of the table reports the RD coefficient from a model that controls for all six classroom features, using coefficients from the joint model shown in Appendix Table 4.

[TABLE 5 ABOUT HERE]

For reading achievement, teacher value added and the five peer characteristics have essentially no power to explain the effects of GHA participation. For math achievement, the addition of controls for the mean lagged scores of classroom peers (row 3) explains about 25 percent of the effect of GHA

[29] In principle the coefficients should be close to 1, after adjusting for sampling error in the estimates of TVA (i.e., "shrinking" the estimates). We have not attempted to perform these adjustments, since we are only interested in how controlling for TVA affects the main RD coefficients. For simplicity we also only produce one estimate of TVA for each teacher based on their impact on combined student outcomes in other years.

[30] The *positive* effect of within-class dispersion on math scores is the opposite of what we would expect if teachers can better match their lessons to a typical student in more homogeneous classes.

participation on all students, and about 17 percent of the effect on minority students. However, in our preferred specification that controls for all six classroom features together (row 8), the explained shares fall to around 10 percent.[31] Overall, we conclude that measures of teacher quality and classroom composition explain none of the effect of GHA classes on reading, and at most a small share of the effect on math achievement, particularly for minorities.

We emphasize that the findings in Table 5 do *not* mean that teacher quality or peer characteristics are unimportant. In the case of teacher quality, which has a strong effect on student scores, the lack of impact on our RD estimates arises because there is no discontinuity in TVA at the threshold for admission to GHA classes. There is a relatively large discontinuity in peer quality at the GHA threshold, but mean third-grade scores of peers have no effect on reading achievement and only a modest effect on math achievement. So the share of the GHA effect on reading scores explained by peer quality is negligible, while the share of the effect on math is relatively small.

III. Between-School/Cohort Design

A concern with our RD estimates is that they may be driven in part by impacts of GHA classes on *non-participants*. The District's policy of offering a GHA classroom if and only if there is at least one gifted child in the fourth-grade cohort provides a design for measuring potential spillover effects.[32] Consider fourth graders ranked just below the top 20 in their cohort at schools with either zero or a small number of gifted students in their cohort. If there are no gifted

[31] The reduction in explanatory power from the combined model is explained by the fact that in the models that control *only* for mean lagged peer scores, the estimated effects of mean lagged peer scores are in the range from 0.12 to 0.13, while in the joint model the estimated effects of mean lagged peer scores are smaller – see Appendix Table 3.

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

children these students will participate in regular classes with the top 20 high achievers. If there is at least one gifted child, however, the top 20 students will be moved to a GHA class. We can thus compare how the test score gains of students ranked just below 20 vary as the number of gifted children varies, looking for an effect when the number exceeds 0.

We can also use a between-school/cohort design to estimate impacts on the test score gains for different subgroups of students ranked in the top 20 students of their cohort. These students will either participate in regular or GHA classes, depending on whether is at least one gifted child. Comparisons between cohorts with or without a gifted child therefore allow us to estimate the *average* treatment effect for the top 20 students in each school, or for narrower groups, like those ranked 1-5.

To implement this design, we identified a set of fourth-grade school/cohorts with 0 to 4 gifted students.[33] We then identified the students in each cohort ranked from 1 to 20 and from 25 to 44 on the previous year's achievement scores, and computed the average changes in math and reading scores from third to fourth grade for students in the two groups from different school/cohorts.

The first-stage relationship for this alternative design and the corresponding reduced-form effects are illustrated in Figure 8. Panel A shows the fraction of the top 20 students in a school/grade cohort who were placed in a GHA class for cohorts with 0, 1, 2, 3, or 4 gifted children. The relationship is nearly linear for schools with 1 or more gifted students, with an intercept of just under 40 percent. Since high achievers at schools with 0 gifted children have no chance of placement in a GHA classroom, the effect of having at least 1 gifted

[33] This sample comprises 28,177 students in 255 fourth-grade cohorts during the years 2009-12 (when we know the District's ranking formula) and includes roughly half of all the fourth-grade cohorts during these years.

child in the cohort is about 40 percentage points. Panel B shows 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 GHA classroom even if one is present. The implied effect of having at least 1 gifted child is only about 5 percent.[34]

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 gain power, and show regression-adjusted impacts that control for cohort-wide average test scores in third grade, the cohort-wide fraction of FRL students, cohort size, and year dummies. For the 1-20 rank group, the pattern in panel C suggests a reduced-form impact of around 0.1 σ's. The impact on the 25-44 ranked group, by comparison, is close to 0.

[FIGURE 8 ABOUT HERE]

Table 6 presents a series of models for the outcomes of different rank groups at school/cohorts with no more than 4 gifted children. Panel A presents results for students ranked 1-20 in their school/cohort; panel B presents results for the 25-44 rank group; panel C presents results for the combined 1-44 group. For the dependent variable identified in each column and the alternative specification in each row, we show the coefficient of a dummy for having at least one gifted child in the school/cohort. Column 1 presents models for third-grade (i.e., pre-intervention) average reading and math scores. Column 2 presents models for the fraction of the group that is placed in a GHA classroom. Finally, columns 3 and 4 show models for the fourth-grade average reading and math scores, and the

[34] The <100 percent impact on the 1-20 group and the positive impact on the 25-44 group are due to a combination of non-compliance and measurement error in the cutoff scores. As explained in Appendix A, there are also a few cohorts where a student is classified as gifted but there is no gifted classroom—e.g., because the school received a waiver or because the student was identified after the school year began.

change in average reading and math scores between third and fourth grades.

[TABLE 6 ABOUT HERE]

The models in row 1 of each panel include the same controls used in Figure 8 (a linear control for the number of gifted students in the cohort; a set of year dummies; and cohort-level controls for average third-grade test scores, mean FRL eligibility, and total enrollment). The models in row 2 of each panel add school fixed effects, and therefore identify the within-school effects of a GHA class for fourth grade. The models in row 3 have the same controls as in row 1, but are estimated on the subsample of school/cohorts with either 0 or 1 gifted student.

In column 1, the estimates for the baseline scores show only small differences between school/cohorts with and without a GHA classroom. In column 2, the models for the probability of placement in a GHA class show that for the 1-20 rank group, the presence of a gifted child raises the probability of being assigned by 30-40 percentage points. For the 24-44 rank group the corresponding effect is 6-8 points, while for the combined 1-44 group the effect is 11-12 points.

The models in columns 3 and 4 suggest that the presence of a GHA class has a positive impact on average scores of the top 20 students, no effect on students ranked 24-44, and an effect on the overall 1-44 group about one-half as large as the effect on the 1-20 group. To interpret the impacts in Panel A, note that the estimate in row 1, column 4 implies a TOT effect of 0.27 σ's on the combined reading and math scores of the 1-20 group. This is remarkably similar to the estimated TOT effects of 0.29 σ's and 0.34 σ's for reading and math, respectively, from our main RD specification (see row 1 of Table 3). There are two offsetting factors that complicate this comparison. On one hand, the fraction of minority students in our between-school sample is higher than in our RD sample (73 percent versus 56 percent). Taking this into account the comparable RD estimates

are arguably a little larger. On the other hand, the RD estimates are based on comparisons for marginally eligible GHA participants. To the extent that the impacts of a GHA class are bigger (or smaller) for marginally eligible students, the RD estimates will be bigger (or smaller) than the average effects for the 1-20 group.

Table 7 uses the between-school design to examine the impacts for narrower rank groups. Specifically, in Panel A we divide highly ranked students in each school/cohort into 4 groups, while in Panel B we divide the students ranked 21 and lower into 5 quintiles within their school/cohort.[35] Given the small samples for these subgroups we present specifications similar to the models in row 1 of panels A and B in Table 6 (with a set of year dummies and cohort-level controls).

[TABLE 7 ABOUT HERE]

For reference, column 1 shows the fraction of black or Hispanic students in each subgroup. This ranges from 68 to 75 percent within the subgroups of the top 20 group, and is even higher for the lower ranked quintile groups. Column 2 shows the estimated impacts of having a GHA class on third-grade (pre-intervention) average math and reading scores. These are all small and insignificant. The first-stage models for placement in a GHA class show relatively large effects for the top 15 students (35-44 percentage points); a smaller – but still highly significant – effect for the marginally eligible 16-20 rank group (around 23 points); and very small effects for the lower ranked students.

The reduced-form impacts for reading and math scores (columns 4 and 5) suggest that the effects of participating in a GHA class are smaller for highly ranked students and larger for those who barely qualify. Normalizing the impacts

[35] We use quintile groups because the total number of students varies across school/cohorts (the mean is 110 and standard deviation is 39). Since the typical class has 23 students, the highest ranked quintile group among the students ranked 21 or lower are, on average, the top 4-5 students in each regular class.

by the first-stage effects, the TOT effects in column 6 range from 0.12 σ's for the top 10 students to 0.5 σ's for the next 10. Given the relatively large standard errors, the subgroup-specific estimates should be interpreted carefully. However, the evidence suggests that GHA classes have, if anything larger impacts for lower-ranked students – the opposite of the pattern predicted by mismatch or invidious comparison effects.

The results in panel B are also interesting and show no evidence of negative (or positive) spillover effects on lower-ranked groups.[36] The reduced-form estimates for the quintile groups are relatively precise, allowing us to rule out spillover effects any larger than +/- 0.05 σ's for students who would be close to the top of the regular class in the presence of a GHA class, for example.

Evidence from "Global" RD Estimates.—In light of these results it is interesting to return to the RD framework – this time comparing actual fourth-grade scores to *predicted* scores for students ranked above and below the GHA eligibility cutoff. Figure 9 plots the means of observed and predicted fourth-grade math and reading scores for students ranked from 50 below the cutoff to 15 above it.[37] We also show quadratic models for both actual and predicted scores, fit separately to the students ranked above and below the cutoff. Panel A shows results for all students while Panel B shows results for black and Hispanic students.

The figures suggest that actual and predicted fourth-grade scores are very similar for all students ranked below the GHA cutoff. In particular, there is no evidence of spillover effects on students just below the cutoff. For students above the cutoff, however, the actual scores are above the predicted scores – with the

[36] The first-stage estimates imply that 5 percent of the quintile 1 group actually participate in a GHA class. Assuming a treatment on the treated effect of 0.5 σ's we would expect to see reduced-form impacts of 0.025 σ's, which we interpret as negligible.

[37] As in Figure 3, predicted scores are based on a regression model that includes third-grade scores, other student characteristics, and school dummies.

largest gap among students just above the cutoff, and little or no gap for students ranked 15 above the cutoff. These patterns are very similar to the pattern of reduced-form effects in Table 8, and provide additional confirmation that: (1) GHA classes have no spillover effects on lower-ranked students, and (2) the impacts of GHA participation are larger for marginally eligible students, contrary to the predicted pattern from mismatch or invidious comparison effects.

[FIGURE 9 ABOUT HERE]

IV. Interpretation of the GHA Effects

The results from our between-school analysis confirm that the presence of a GHA class leads to significant achievement gains for black and Hispanic students who are admitted to these classes, with no spillover effects on non-participants. In light of these findings, we return to the question of how to interpret the impacts of a GHA class.

The analysis in section II.F leads us to conclude that these impacts are not mediated by teacher quality or the peer composition channels that have been highlighted in the tracking literature. Moreover, neither these channels, nor explanations based on the match between student ability and the level/pace of instruction, can readily explain the combination of large positive effects for minority students and small, insignificant effects for whites, since these mechanisms should also affect white students.

Instead, we hypothesize that higher-ability minority students face specific obstacles in a regular classroom environment that lead them to under-perform relative to their potential, and that some of these impediments are reduced in a GHA class. A key feature of this hypothesis is that it is consistent with both the absence of effects on white students, *and* the absence of spillover effects on students who remain in regular classes when a GHA class is established.

Evidence of systematic under-performance by minority students is presented in Appendix Table 5. Here we show a series of models that relate average third-grade reading and math scores to measures of cognitive ability based on NNAT scores (used by the District between 2005 and 2009 to screen second graders for gifted status) and dummies for race and ethnicity. These models show that black students' third-grade achievement scores are 0.2 to 0.45 σ's below those of white students with similar cognitive ability (depending on whether we also control for school dummies and FRL status), while Hispanics have scores that are 0.15 to 0.25 σ's below those of whites.

Turning to our RD sample, if we measure relative performance in fourth grade using residuals from a regression of fourth-grade reading and math scores on NNAT scores, the white-minority achievement gap is 0.39 σ's among students who just miss the GHA cutoff.[38] Estimates from linear RD models that use residual (NNAT-adjusted) achievement as the dependent variable show a significant reduced-form discontinuity of 0.14 σ's for minorities but none for whites. Scaling up the reduced-form estimate for minorities implies a treatment-on-treated effect of roughly 0.40— a big enough effect to close the 0.39 σ achievement gap for minority participants.[39]

What specific obstacles might cause minorities to under-perform in regular classrooms? First, a pattern of under-performance by minorities could be perpetuated in the regular classroom because of lowered teacher expectations about student potential. Such a pattern might be mitigated in a GHA class for at

[38] Residual scores are estimates from a linear regression of the average of fourth-grade reading and math scores on the NNAT score, using all students with NNAT scores who are in fourth grade during our sample period and not in a GHA classroom. Specification tests show that the white-minority gap in relative achievement is similar among high achievers and the broader student population. The RD sample in this analysis is limited to the subset of students for whom NNAT scores available; this sample includes 978 white students and 1,634 minorities.

[39] To calculate the TOT effect, we normalize the RD estimate of 0.14 by a first-stage estimate for minorities of .35. The latter is the first-stage estimate reported in row 2b of Table 3, scaled up by 15 percent to adjust for attenuation due to misclassification (see footnotes 16 and 17).

least two reasons. Since teachers know that all students in the GHA class have relatively high achievement scores, they may revise their views about the ability levels of their minority students. Also, GHA teachers have received some training on the education of special populations of gifted students emphasizing stereotyping and intercultural competence.

Evidence consistent with a cycle of minority underperformance and low teacher expectations comes from studying the impact of the District's universal screening program for identifying potentially gifted students. Introduced in 2005, this program changed the process by which students are referred for gifted evaluation, replacing a system based on teacher nominations with an automated process based on NNAT scores. As discussed in Card and Giuliano (2015), it uncovered large numbers of minority students with relatively high cognitive ability—ultimately leading to increases of 80 percent and 130 percent, respectively, in the numbers of black and Hispanic gifted students in the District. The newly identified gifted students had IQ scores similar to those who were identified in earlier cohorts, but lower achievement—suggesting that teachers are often unaware of the cognitive ability of minority students who under-perform in class.

A large body of ethnographic research suggests another cause of minority under-performance in regular classrooms – namely, that minority students face peer pressure against high academic performance (or "acting white").[40] Arguably, such pressures are reduced in a GHA class where all students are labelled as either gifted or high achieving. Bursztyn and Jensen (2015), for example, find that Hispanic high school students in non-honors classes are less likely to enroll in SAT preparation classes when their enrollment choices are revealed to classmates,

[40] See Fryer and Torelli (2010) for a brief summary of this literature. Similar hypotheses could explain the larger benefits for minority participants in studies of class size reductions (Krueger 1999; Krueger and Whitmore 2001), and admission to selective colleges and universities (Dale and Krueger 2002).

but that the negative effect of information revelation is eliminated in honors classes. Fryer and Torelli (2010) show that the social pressure against academic achievement is particularly strong for boys, so this channel could also explain the larger GHA impact for minority boys.

If minority students have more positive interactions with teachers and peers in the GHA classrooms, this might be reflected in observable student behavior such as attendance and disciplinary actions. Appendix Table 6 presents estimates from reduced-form RD models in which the dependent variables measure the likelihood of having more than one unexcused absence or being suspended more than once in grades 4-6. We find significant negative discontinuities in these outcomes for minorities but none for whites. Moreover, the discontinuities are especially large for minority boys, who have the highest rates of suspension and unexcused absence when placed in regular classrooms, and who also show the largest achievement gains from GHA participation.

These results suggest to us that the GHA impacts on achievement are at least partly mediated by changes in high-ability minority students' behavior and motivation to perform well in school. A recent study by Dee and Penner (2016) finds a similar pattern of effects from a “culturally relevant” high school curriculum program for lower-achieving minority students. Interestingly, Angrist et al. (2013, Table 5) find much larger effects of urban charter schools on black and Hispanic students than on whites -- effects that they attribute in part to a culture of high expectations at so-called "No Excuses" charter schools.

V. Conclusions

The District's policy of creating GHA classrooms provides valuable new evidence on the effects of a within-school tracking program for high-achieving students. We use two complementary approaches to study the impact of GHA

classes – a regression discontinuity design and a between-cohort design. Our RD design shows that participation in GHA classes leads to significant achievement gains for minority participants, with treatment-on-the-treated effects of around 0.5 σ's on fourth-grade reading and math, and persistent impacts to at least sixth grade. We confirm these effects using comparisons between school/cohorts with no gifted children (and no GHA class) and those with 1-4 gifted children (and a GHA class with seats for 20-23 high achievers). This design also allows us to check for effects on students who remain in regular classes when a GHA class is established. Importantly, we find no spillover effects on non-participants, including those who narrowly miss the cutoff for the class.

We also use the RD design to investigate several possible explanations for the large impact of GHA classes, including differences in the quality of teachers assigned to these classes, and peer effects associated with the average ability of GHA classmates, their gender and race composition, and the presence of disruptive students. Although we find that teacher quality (measured by average value added) is an important determinant of student achievement, there is no discontinuous change at the threshold for entry to the GHA class, so this channel cannot explain the impacts we observe. We find modest effects of lagged peer test scores on math achievement, accounting for a relatively small share (around 10 percent) of the GHA impact on math, but none of the effect on reading achievement. Moreover, neither these channels nor potential differences in instruction or pacing can explain the combination of large, positive effects for minorities and small, insignificant effects for white GHA participants. Instead, we argue that the likely explanation for our findings is that higher-ability minority students tend to under-perform academically in a regular classroom environment, and that some of the obstacles faced by these students – including negative peer pressure and low expectations of teachers – are reduced or eliminated in a GHA class.

While an important limitation of our analysis is that it pertains to only a single school district, nevertheless the student population in the District is highly diverse and arguably representative of the student population in many other large urban districts. Overall, our results suggest that a comprehensive tracking program that establishes a separate classroom in every school for the top-performing students has the potential to significantly boost the performance of higher-achieving minority students – even in the poorest neighborhoods of a large urban school district. Given the high degree of economic and racial segregation in many urban districts, such a program could effectively serve large numbers of high achieving and minority students, and it could do so at little or no cost to other students or school district budgets.

References

Abdulkadiroglu, Atila, Joshua D. Angrist and Parag A. Pathak. 2014. “The Elite Illusion: Achievement Effects at Boston and New York Exam Schools.” *Econometrica* 82(1): 137-196.

Angrist, Joshua D., Parag Pathak, and Christopher R. Walters. 2013. "Explaining Charter School Effectiveness." *American Economic Journal: Applied Economics* 5(4): 1-27.

Arcidiacono, Peter, Esteban Aucejo, Patrick Coate and V. Joseph Hotz. 2014. "Affirmative Action and University Fit: Evidence from Proposition 209." *IZA Journal of Labor Economics* 3(7). doi:10.1186/2193-8997-3-7.

Betts, Julian R. 2011. “The Economics of Tracking in Education.” In *Handbook of the Economics of Education*, Volume 3, edited by Eric A. Hanushek, Stephen Machin and Ludger Woessmann, 341-381. Amsterdam: North Holland.

Betts, Julian R. 2003. Andrew C. Zhao and Lorien A. Rice. *Determinants of Student Achievement: New Evidence from San Diego.* San Francisco: Public Policy Institute of California.

Booij, Adam S., Edwin Leuven and Hessel Oosterbeek. 2015. "Ability Peer Effects in University: Evidence from a Randomized Experiment." Institute for the Study of Labor Working Paper 8769.

Bowen, William G. and Derek Bok. 1998. *The Shape of the River: Long-term Consequences of Considering Race in College and University Admissions*. Princeton NJ: Princeton University Press.

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

Bursztyn, Leonardo and Robert Jensen. 2015. "How Does Peer Pressure Affect Educational Investments?" *The Quarterly Journal of Economics* 130 (3): 1329-1367.

Card, David and Laura Giuliano. 2014. "Does Gifted Education Work? For Which Students?" National Bureau of Economic Research Working Paper 20453.

Card, David and Laura Giuliano. 2015. "Can Universal Screening Increase the

Representation of Low Income and Minority Students in Gifted Education?" National Bureau of Economic Research Working Paper 21519.

Carrell, Scott E., Bruce I. Sacerdote and James E. West. 2013. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81(3): 855-882.

Carrell, Scott E. and Mark Hoekstra. 2010. Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone's Kids." *American Economic Journal: Applied Economics* 2(1): 211-228.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104(9): 2593–2632.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor. 2009. "The Academic Achievement Gap in Grades 3 to 8." *Review of Economics and Statistics* 91(2): 398-419.

Dale, Stacy Berg and Alan B. Krueger. 2002. "Estimating The Payoff Of Attending A More Selective College: An Application Of Selection On Observables And Unobservables." *Quarterly Journal of Economics* 117(4): 1491-1527.

Davis, Billie, John Engberg, Dennis Epple, Holger Sieg and Ron Zimmer. 2013. "Bounding the Retention Effects of a Gifted Program Using a Modified Regression Discontinuity Design." *Annals of Economics and Statistics* 111-112: 10-34.

Dee, Thomas and Emily Penner. 2016. "The Causal Effects of Cultural Relevance: Evidence from an Ethnic Studies Program." NBER Working Paper 21865.

Dieterle, Steven G., Cassandra Guarino, Mark Reckase and Jeffrey M. Wooldridge. 2015. "How Do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value Added." *Journal of Policy Analysis and Management* 34(1): 32-58.

DiNardo, John and David S. Lee. 2011. "Program Evaluation and Research Design." In *Handbook of Labor Economics*, edited by Orley Ashenfelter and David Card, (4a): 463-536. New York: Elsevier.

Dobbie, Will and Roland G. Fryer Jr. 2014. “The Impact of Attending a School with High-Achieving Peers: Evidence from New York City Exam Schools.” *American Economic Journal—Applied Economics* 6(3): 58-75.

Donovan, M. Suzanne and Christopher T. Cross. 2002. *Minority Students in Special and Gifted Education*. Washington DC: National Academies Press.

Dotter, Dallas. 2013. "Is Gifted and Talented Education a Gift that Keeps on Giving?" Unpublished Manuscript, UC San Diego Department of Economics.

Duflo Esther, Pascaline Dupas, and Michael Kremer. 2011. “Peer Effects, Teacher Incentives and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya.” *American Economic Review* 101(5): 1739-74.

Fryer, Roland G. and Steven D. Levitt. 2006. "The Black-White Test Score Gap Through Third Grade." *American Law and Economics Review* 8(2): 249-281.

Fryer, Roland G. and Paul Torelli. 2010. "An Empirical Analysis of Acting White." *Journal of Public Economics* 94(5-6): 380-396.

Hanushek, Eric and Stephen G. Rivkin. 2009. "Harming the Best: How Schools Affect the Black-White Achievement Gap." *Journal of Policy Analysis and Management* 28(3): 366-393.

Hoxby, Caroline M. 2000. "Peer Effects in the Classroom: Learning from Gender and Race Variation." National Bureau of Economic Research Working Paper 7867.

Hoxby, Caroline M. and Gretchen Weingarth. 2005. “Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects.” Unpublished manuscript.

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

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497-532.

Krueger, Alan B. and Diane Whitmore. 2001. "The Effect of Attending a Small

Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *Economic Journal* 111(468): 1-28.

Lavy, Victor and Analia Schlosser. 2011. "Mechanisms and Impacts of Gender Peer Effects at School." *American Economic Journal*: *Applied Economics* 3(2): 1-33.

Lazear, Edward P. 2001. "Educational Production." *Quarterly Journal of Economics* 116(3): 777-803.

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

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

Murphy, Richard and Felix Weinhardt. 2014. "Top of the Class: The Importance of Ordinal Rank." CESifo Working Papers 4815.

Oakes, Jeannie. 1985. *Keeping Track: How Schools Structure Inequality.* New Haven, CT: Yale University Press.

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

Rivkin, Steven G., Eric A. Hanushek and John F. Kain. 2005. "Teachers, Schools and Academic Achievement." *Econometrica* 73(2): 417-458.

Rothstein, Jesse. 2015. "Revisiting the Impacts of Teachers." eml.berkeley.edu/~jrothst/workingpapers/Rothstein_cfr_oct2015.pdf

Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable that has Been Affected by the Treatment." *Journal of the Royal Statistical Society* 147(5): 656-666.

Sacerdote, Bruce. 2011. "Peer Effects in Education: How Might They Work, How Big Are They, and How Much Do We Know So Far?" In *Handbook of the Economics of Education*, edited by Eric A. Hanushek, Stephen Machin and Ludger Woessmann, (3): 249-278. Amsterdam: Elsevier.

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

U.S. Department of Education, Office for Civil Rights. 2014. *Data Snapshot: College and Career Readiness*. Issue Brief No. 3.

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

Table 1. Sample Characteristics

		RD samples (students ranked +/- 10 from cutoff)			Between-school/cohort sample (0-4 gifted in cohort)	
	All in 3rd grade in 2008-11	All school /cohorts	Cohorts with 5+ gifted	Cohorts with 1-4 gifted	Ranked 1-20 in cohort	Ranked 25-44 in cohort
	(1)	(2)	(3)	(4)	(5)	(6)
Student demographics						
Female (percent)	48	51	51	52	52	50
White (percent)	28	34	41	23	20	16
Black (percent)	39	28	18	45	52	58
Hispanic (percent)	27	28	30	24	21	21
Asian (percent)	3	5	7	3	3	2
Free lunch eligible (percent)	58	49	36	71	72	79
English language learner (percent)	10	1	1	2	3	7
Median income in ZIP ($1,000s)	57.8	59.6	66.9	47.8	46.0	45.9
Mean NNAT (screening test)	104.6	113.3	116.1	109.4	112.0	103.5
Percent taking test	*72*	*70*	*66*	*76*	*72*	*74*
Third grade state test scores						
Mean reading test	-0.01	0.86	1.04	0.57	0.81	0.03
Mean math test	0.00	0.80	0.99	0.49	0.74	0.04
Fourth grade state test scores						
Mean reading test	0.01	0.69	0.85	0.43	0.60	-0.06
Mean math test	0.03	0.69	0.83	0.45	0.61	-0.05
Characteristics of school-wide peers						
Mean 3rd grade reading test	--	0.09	0.22	-0.12	-0.16	-0.16
Mean 3rd grade math test	--	0.07	0.20	-0.15	-0.19	-0.19
Free lunch eligible (percent)	--	58	48	76	81	81
Characteristics of school's fourth grade GHA classroom						
Mean 3rd grade reading test	--	0.95	1.10	0.72	0.60	0.60
Mean 3rd grade math test	--	0.91	1.04	0.70	0.57	0.57
Non-gifted high achievers (percent)	--	73	63	87	89	89
Number of observations	*70,058*	*4,144*	*2,553*	*1,591*	*4,767*	*5,016*

(continued)

Table 1. Sample Characteristics, cont'd.

		RD samples (students ranked +/- 10 from cutoff)			Between-school/cohort sample (0-4 gifted in cohort)	
	All in 3rd grade in 2008-11	All school /cohorts	Cohorts with 5+ gifted	Cohorts with 1-4 gifted	Ranked 1-20 in cohort	Ranked 25-44 in cohort
	(1)	(2)	(3)	(4)	(5)	(6)
Characteristics of school's fourth grade GHA classroom						
Mean 3rd grade reading test	--	0.95	1.10	0.72	0.60	0.60
Mean 3rd grade math test	--	0.91	1.04	0.70	0.57	0.57
Non-gifted high achievers (percent)	--	73	63	87	89	89
Characteristics of school/cohorts						
Total 4th grade enrollment	--	132.62	143.92	114.48	108.67	108.67
Number of 4th grade classrooms	--	5.77	6.26	4.98	4.72	4.72
Number of GHA classrooms	--	1.16	1.26	1.00	0.65	0.65
Number of students per classroom	--	23.58	23.54	23.64	23.27	23.27
Fraction of students in GHA classroom	--	0.19	0.19	0.19	0.12	0.12
Number of observations	*70,058*	*4,144*	*2,553*	*1,591*	*4,767*	*5,016*

Notes: Sample in column 1 includes one observation per student observed in a non-charter district elementary school in third grade between 2008 and 2011. Sub-samples in columns 2-4 include students in fourth grade in 2009-2012 in school/cohorts that complied with District's ranking formula and had at least one fourth-grade GHA classroom, and whose scores fell within +/- 10 rank points of the eligibility cutoff for their school/cohort. Sub-samples in columns 5-6 include students in 4th grade in 2009-2012 in cohorts with between 0-4 gifted students. 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. Non-verbal ability index (NNAT) is measured at the end of second grade for students who took the test between 2007 and 2009. State test scores are standardized across district within year and grade.

Table 2. Regression Discontinuity Estimates for Fourth Grade Outcomes

	Baseline Achievement		First Stage	Reduced-Form Outcomes		
	3rd Grade Reading (1)	3rd Grade Math (2)	Probability in GHA Classroom (3)	4th Grade Reading (4)	4th Grade Math (5)	4th Grade Writing (6)
1. No controls	0.008	-0.046	0.323	0.092	0.073	-0.011
	(0.029)	(0.043)	(0.025)	(0.034)	(0.039)	(0.054)
2. School and year fixed effects and student controls	0.015	-0.044	0.319	0.093	0.087	-0.012
	(0.027)	(0.040)	(0.026)	(0.031)	(0.035)	(0.051)
3. Differenced specification (based on change in test scores)	--	--	--	0.092	0.105	--
				(0.033)	(0.041)	
Number of observations	*4,144*	*4,144*	*4,144*	*4,144*	*4,144*	*4,144*

Notes: Estimates from models of dependent variable in column heading as a function of a student's rank (within school/cohort) on third-grade test scores. Entries are estimated coefficients on a dummy for the student's rank exceeding the cohort-specific cutoff for placement in the fourth-grade GHA classroom. All models include a linear term in rank interacted with the dummy. Models in rows 2-3 control for student demographics (age, gender, race/ethnicity, and median household income in ZIP code), dummies for year in fourth grade, and a complete set of school dummies. Models in row 2, columns 3-6 also control for third-grade scores in math and reading. Analysis sample is described in column 2 of Table 1. Parentheses contain standard errors, clustered by school.

Table 3. RD Heterogeneity Analysis for Fourth Grade Outcomes

	Baseline Scores		First Stage	Fourth Grade Test Scores (2SLS)			
	3rd Grade Reading	3rd Grade Math	Probability in GHA Classroom	Reading	Diff. Reading	Math	Diff. Math
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. Full sample	0.02	-0.04	0.32	0.29	0.29	0.28	0.34
(n=4144)	(0.03)	(0.04)	(0.03)	(0.10)	(0.10)	(0.11)	(0.13)
2. By Race/Ethnicity							
2a. White	0.02	-0.07	0.33	-0.07	-0.07	-0.13	-0.04
(n=1397)	(0.05)	(0.07)	(0.04)	(0.17)	(0.19)	(0.17)	(0.21)
2b. Black and Hispanic	-0.01	-0.02	0.30	0.61	0.64	0.52	0.56
(n=2323)	(0.04)	(0.04)	(0.04)	(0.15)	(0.16)	(0.16)	(0.17)
2c. Black Only	-0.00	0.00	0.27	0.65	0.65	0.69	0.69
(n=1180)	(0.05)	(0.05)	(0.05)	(0.25)	(0.27)	(0.23)	(0.26)
2d. Hispanic Only	-0.02	-0.07	0.31	0.69	0.76	0.42	0.53
(n=1143)	(0.05)	(0.05)	(0.06)	(0.20)	(0.22)	(0.22)	(0.24)
3. Black and Hispanic Only, by FRL Status							
3a. FRL-eligible	-0.05	-0.02	0.24	0.56	0.70	0.47	0.55
(n=1538)	(0.04)	(0.05)	(0.05)	(0.20)	(0.23)	(0.23)	(0.25)
3b. Non-FRL-eligible	0.06	-0.02	0.44	0.74	0.67	0.63	0.63
(n=785)	(0.06)	(0.07)	(0.07)	(0.24)	(0.25)	(0.21)	(0.21)
4. Black and Hispanic Only, by Number Gifted in School/Cohort							
4a. 1-4 Gifted	-0.08	0.04	0.27	0.62	0.73	0.79	0.80
(n=1072)	(0.04)	(0.06)	(0.05)	(0.19)	(0.19)	(0.23)	(0.25)
4b. 5 or more Gifted	0.05	-0.10	0.32	0.63	0.59	0.36	0.49
(n=1223)	(0.06)	(0.06)	(0.05)	(0.23)	(0.25)	(0.21)	(0.23)
5. Black and Hispanic Only, by Gender							
5a. Girls	0.02	-0.00	0.31	0.35	0.31	0.43	0.42
(n=1216)	(0.05)	(0.05)	(0.05)	(0.19)	(0.21)	(0.20)	(0.22)
5b. Boys	-0.03	-0.03	0.27	0.92	1.04	0.69	0.77
(n=1107)	(0.05)	(0.06)	(0.05)	(0.26)	(0.29)	(0.29)	(0.30)

Notes: Estimates from RD models with school and year fixed effects and student controls, as in Table 2, row 2 (see Table 2 note for details). In all two-stage least squares models (columns 4-7) the first-stage model is for the probability of being in the fourth-grade GHA classroom. Parentheses contain standard errors, clustered by school.

Table 4. Two-Stage Least Squares RD Heterogeneity Analysis for Fifth and Sixth Grade Outcomes

	Fifth Grade Outcomes					Sixth Grade Outcomes			
	Reading (1)	Diff. Reading (2)	Math (3)	Diff. Math (4)	Science (5)	Reading (6)	Diff. Reading (7)	Math (8)	Diff. Math (9)
1. Full sample	0.12	0.10	0.16	0.20	0.13	0.23	0.21	0.41	0.45
(n=3598)	(0.12)	(0.12)	(0.10)	(0.11)	(0.11)	(0.12)	(0.12)	(0.12)	(0.14)
2. By Race/Ethnicity									
2a. White	-0.11	-0.11	-0.10	-0.05	-0.23	-0.20	-0.20	0.12	0.17
(n=1187)	(0.18)	(0.18)	(0.19)	(0.20)	(0.17)	(0.17)	(0.19)	(0.19)	(0.21)
2b. Black and Hispanic	0.20	0.21	0.34	0.38	0.32	0.62	0.64	0.68	0.73
(n=2047)	(0.19)	(0.20)	(0.15)	(0.18)	(0.19)	(0.21)	(0.21)	(0.21)	(0.23)
3. Black and Hispanic Only, by FRL Status									
3a. FRL eligible	0.24	0.32	0.20	0.26	0.27	0.78	0.89	0.83	0.90
(n=1378)	(0.29)	(0.32)	(0.25)	(0.29)	(0.31)	(0.33)	(0.36)	(0.31)	(0.36)
3b. Non-FRL eligible	0.35	0.27	0.63	0.63	0.43	0.65	0.56	0.56	0.56
(n=669)	(0.26)	(0.27)	(0.19)	(0.19)	(0.25)	(0.27)	(0.28)	(0.27)	(0.27)
4. Black and Hispanic Only, by Number Gifted in School/Cohort									
4a. 1-4 Gifted	0.20	0.18	0.24	0.42	0.28	0.67	0.65	0.58	0.75
(n=1070)	(0.24)	(0.26)	(0.18)	(0.21)	(0.22)	(0.27)	(0.29)	(0.25)	(0.28)
4b. 5 or more Gifted	0.23	0.28	0.45	0.37	0.36	0.56	0.63	0.80	0.70
(n=977)	(0.28)	(0.30)	(0.24)	(0.28)	(0.30)	(0.28)	(0.29)	(0.31)	(0.34)
5. Black and Hispanic Only, by Gender									
5a. Girls	0.09	0.05	0.34	0.34	-0.06	0.34	0.29	0.36	0.36
(n=1074)	(0.21)	(0.24)	(0.19)	(0.21)	(0.19)	(0.25)	(0.28)	(0.23)	(0.25)
5b. Boys	0.50	0.60	0.49	0.57	0.90	0.93	1.07	1.21	1.28
(n=973)	(0.32)	(0.34)	(0.26)	(0.29)	(0.37)	(0.36)	(0.40)	(0.39)	(0.41)

Notes: Estimates from two-stage least squares RD models where the first-stage effect is on the probability of being in the fourth-grade GHA classroom. All models control for school and year fixed effects and student characteristics (as in Table 2, row 2). Analysis samples include all students in the main analysis sample who are in the relevant sub-population and who are observed in the District through the end of sixth grade. Parentheses contain standard errors, clustered by school.

Table 5. RD Estimates of GHA Impact on Gain Scores in Reading and Math, Adjusted for Effects of Changes in Classroom Characteristics

	Reading			Math		
Classroom Characteristic:	All Students (1)	Whites only (2)	Minorities only (3)	All Students (4)	Whites only (5)	Minorities only (6)
1. None	0.10	-0.00	0.19	0.11	0.01	0.17
	(0.04)	(0.07)	(0.05)	(0.04)	(0.08)	(0.05)
2. Teacher value added	0.10	-0.00	0.19	0.12	0.02	0.17
	(0.04)	(0.07)	(0.05)	(0.04)	(0.08)	(0.05)
3. Average of peers'	0.09	-0.01	0.18	0.08	-0.02	0.14
lagged test scores	(0.03)	(0.07)	(0.05)	(0.04)	(0.08)	(0.05)
4. Std. dev. of peers'	0.10	-0.00	0.19	0.12	0.02	0.17
lagged test scores	(0.04)	(0.07)	(0.05)	(0.04)	(0.08)	(0.05)
5. Peer fraction suspended	0.10	-0.00	0.19	0.11	0.01	0.17
in 3rd grade	(0.04)	(0.07)	(0.05)	(0.04)	(0.08)	(0.05)
6. Peer fraction female	0.10	-0.00	0.19	0.12	0.01	0.17
	(0.04)	(0.07)	(0.05)	(0.04)	(0.08)	(0.05)
7. Peer fraction minority	0.10	-0.00	0.19	0.11	0.01	0.17
male	(0.04)	(0.07)	(0.05)	(0.04)	(0.08)	(0.05)
8. All mechanisms	0.10	0.00	0.19	0.10	-0.00	0.15
	(0.03)	(0.07)	(0.05)	(0.04)	(0.08)	(0.05)
Number of observations	*3685*	*1266*	*2040*	*3685*	*1266*	*2040*

Notes: Table reports reduced-form RD estimates from models for test score gains in reading and math between 3rd and 4th grade. Estimation samples (and indicated sample sizes) exclude students for whom teacher value added (TVA) cannot be estimated because the teacher is only observed in one year. (See Appendix B for description of the model used to estimate TVA.) Row 1 shows the estimates from models with controls as in row 3 of Table 2. Rows 2-6 report estimates from RD models that also control for the specified classroom characteristic and constrain the coefficient to equal the estimated effect of that characteristic, as reported in the odd columns of Appendix Table 4 (see text and Appendix Table 4 note for details). Row 7 controls for all classroom characteristics simultaneously and constrains the coefficients to those reported in the even columns of Appendix Table 4. Standard errors, clustered by school, in parentheses.

Table 6. Effect of Having One or More Gifted Students in Fourth Grade School/Cohort

	Average 3rd Grade Reading & Math (1)	Probability in 4th Grade GHA Classroom (2)	Average 4th Grade Reading & Math (3)	Difference, Grade 3-4 Reading & Math (4)
A. Students Ranked 1-20 in School/Cohort				
1. Control for school/cohort characteristics	-0.021	0.372	0.092	0.099
	(0.023)	(0.060)	(0.035)	(0.033)
2. Add school fixed effects	-0.005	0.278	0.072	0.074
	(0.027)	(0.072)	(0.041)	(0.040)
Number of observations	*4767*	*4767*	*4767*	*4767*
3. School/cohort controls; cohorts with 0 or 1 gifted	-0.036	0.438	0.070	0.080
	(0.019)	(0.052)	(0.034)	(0.032)
Number of observations	*2251*	*2251*	*2251*	*2251*
B. Students Ranked 25-44 in School/Cohort				
1. Control for school/cohort characteristics	-0.000	0.071	0.009	0.009
	(0.018)	(0.025)	(0.034)	(0.035)
2. Add school fixed effects	0.001	0.058	-0.017	-0.017
	(0.020)	(0.031)	(0.039)	(0.040)
Number of observations	*5016*	*5016*	*5016*	*5016*
3. School/cohort controls; cohorts with 0 or 1 gifted	-0.001	0.089	-0.009	-0.009
	(0.016)	(0.014)	(0.029)	(0.031)
number of observations	*2243*	*2243*	*2243*	*2243*
C. Students Ranked 1-44 in School/Cohort				
1. Control for school/cohort characteristics	-0.016	0.214	0.048	0.052
	(0.015)	(0.035)	(0.029)	(0.030)
2. Add school fixed effects	-0.004	0.164	0.021	0.022
	(0.015)	(0.043)	(0.033)	(0.033)
Number of observations	*10774*	*10774*	*10774*	*10774*
3. School/cohort controls; cohorts with 0 or 1 gifted	-0.017	0.255	0.025	0.030
	(0.011)	(0.027)	(0.026)	(0.027)
Number of observations	*4953*	*4953*	*4953*	*4953*

Notes: Estimates are coefficients on a dummy for having at least one gifted student in one's fourth-grade school/cohort from models that control linearly for the number of gifted students in the cohort. Additional controls in all models include the cohort average of third-grade reading and math scores, cohort % FRL, cohort size, and a set of year dummies. The models in row 2 (but not row 3) add school fixed effects. Estimation sample in rows 1 & 2 of each panel includes students in the specified rank group in one of 256 fourth-grade cohorts with between 0 and 4 gifted students. Sample in row 3 is restricted to cohorts with only 0 or 1 gifted students. Parentheses contain standard errors, clustered by school.

Table 7. Estimated Effect of Having at Least one Gifted Student in Fourth Grade School/Cohort, Allowing for Heterogeneity by Relative Achievement within Classroom

		Baseline	First Stage	Reduced Form Outcomes		2SLS (TOT)
	Percent Black or Hispanic in Group	Average 3rd Grade Reading & Math	Probability in 4th Grade GHA Classroom	Average 4th Grade Reading & Math	Difference, Grade 4-3 Reading & Math	Difference, Grade 4-3 Reading & Math
	(1)	(2)	(3)	(4)	(5)	(6)
A. Students Ranked 1-20 within School/Cohort						
All	72.4	-0.021	0.372	0.092	0.099	0.266
		(0.023)	(0.060)	(0.035)	(0.033)	(0.095)
#1-5	67.5	-0.034	0.442	0.036	0.051	0.115
		(0.049)	(0.074)	(0.058)	(0.058)	(0.127)
#6-10	71.8	-0.015	0.416	0.049	0.053	0.127
		(0.034)	(0.070)	(0.046)	(0.048)	(0.112)
#11-15	74.7	0.018	0.348	0.164	0.162	0.464
		(0.025)	(0.068)	(0.045)	(0.045)	(0.150)
#16-20	75.4	-0.020	0.233	0.122	0.128	0.547
		(0.029)	(0.057)	(0.045)	(0.045)	(0.232)
B. Students Ranked 21 and Lower within School/Cohort, by Quintile						
Quintile 1	77.0	-0.009	0.051	0.006	0.007	--
(highest)		(0.014)	(0.012)	(0.022)	(0.023)	
Quintile 2	81.1	-0.011	0.044	0.022	0.023	--
		(0.012)	(0.014)	(0.031)	(0.031)	
Quintile 3	82.1	-0.005	0.022	0.026	0.027	--
		(0.012)	(0.012)	(0.026)	(0.027)	
Quintile 4	85.4	0.004	0.023	0.021	0.021	--
		(0.014)	(0.013)	(0.030)	(0.030)	
Quintile 5	87.1	0.048	0.021	0.004	-0.013	--
(lowest)		(0.029)	(0.011)	(0.037)	(0.037)	

Notes: Estimates in columns 2-6 are coefficients on a dummy for having at least one gifted student in one's fourth grade school/cohort, from models with year dummies and cohort-level controls as in row 1 of Panels A & B in Table 6. See Table 6 note for details. Two-stage least squares estimates in column 6 are normalized by the first-stage effect shown in column 3. Estimation samples are subgroups, based on third-grade test-score rank, from a sample of 28,177 students in one of 256 4th grade cohorts with between 0 and 4 gifted students. Parentheses contain standard errors, clustered by school.

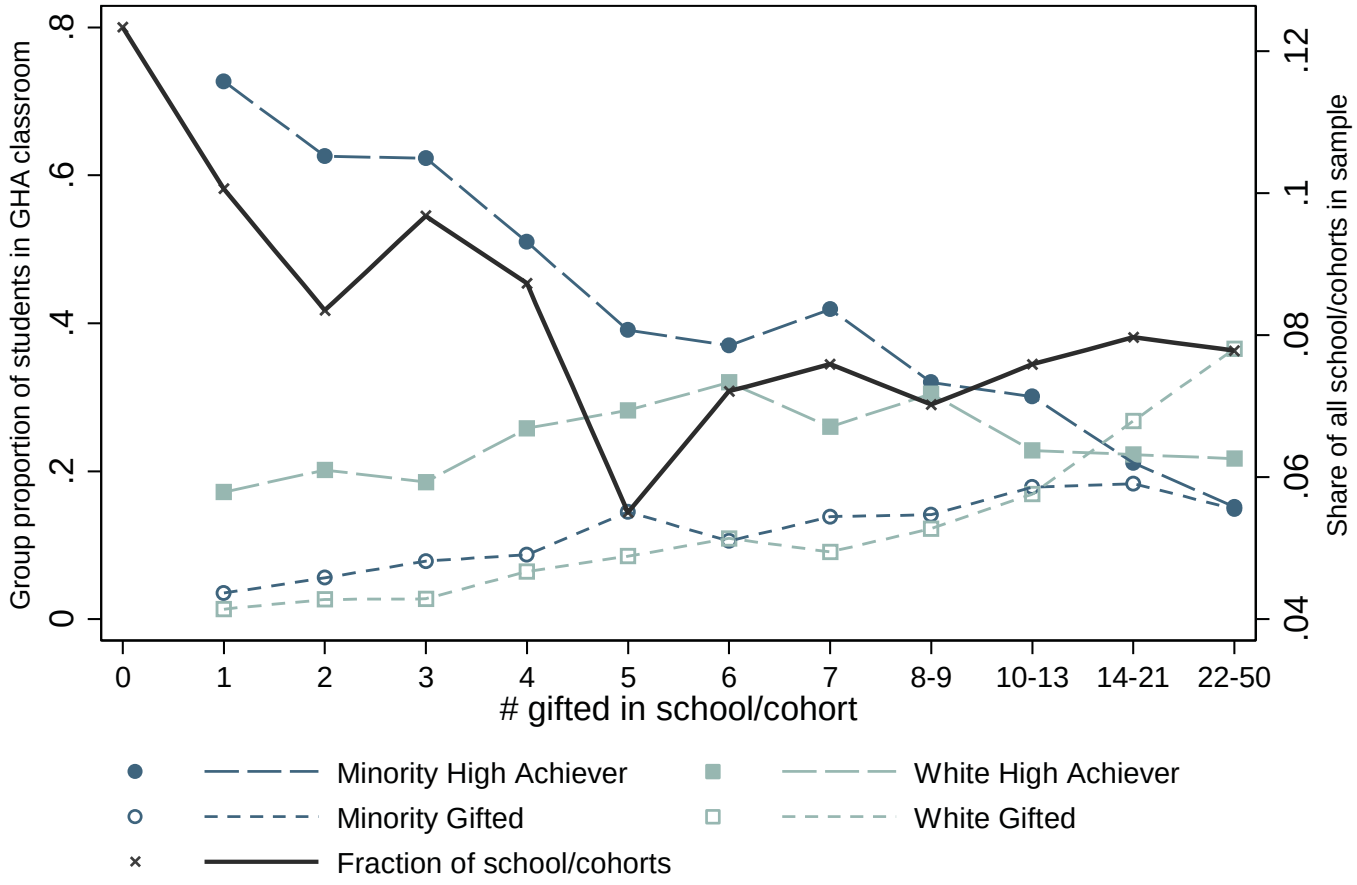

FIGURE 1. FRACTION OF SCHOOL/COHORTS IN SAMPLE AND COMPOSITION OF FOURTH-GRADE GHA CLASSROOMS, BY NUMBER GIFTED IN SCHOOL/COHORT

Notes: Four series showing composition of GHA classroom (left axis) are based on sample of 68,263 students in a fourth-grade GHA classroom in the 2008/09 through 2011/12 school years. Share of school/cohorts (right axis) is based on 527 cohorts during the same period. Minority is defined as black or Hispanic; Asians and other non-white minorities make up less than ten percent of all students in GHA classrooms at all schools.

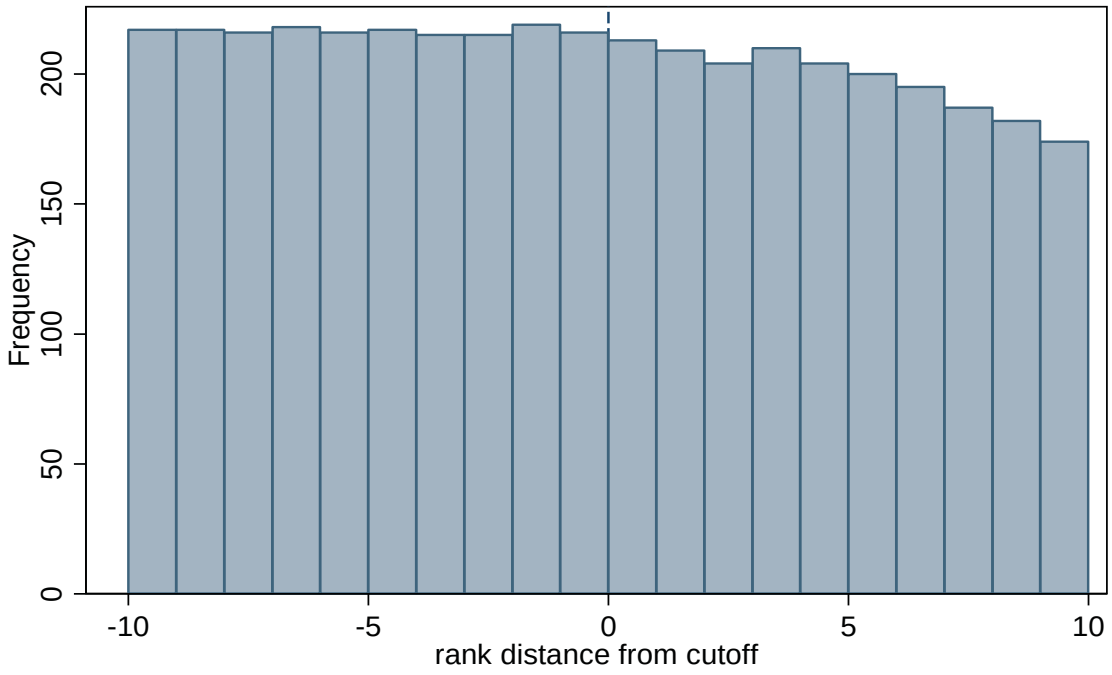

FIGURE 2. HISTOGRAM OF RUNNING VARIABLE FOR RD ANALYSIS (CLASS RANK IN THIRD-GRADE ACHIEVEMENT SCORES)

Notes: Sample is all non-gifted students enrolled in fourth grade between 2009-2012 whose class rank on third-grade achievement scores was within +/- 10 points of the placement cutoff for the fourth-grade GHA classroom in their school/cohort (as in Table 1, column 2). Sample size is 4,144.

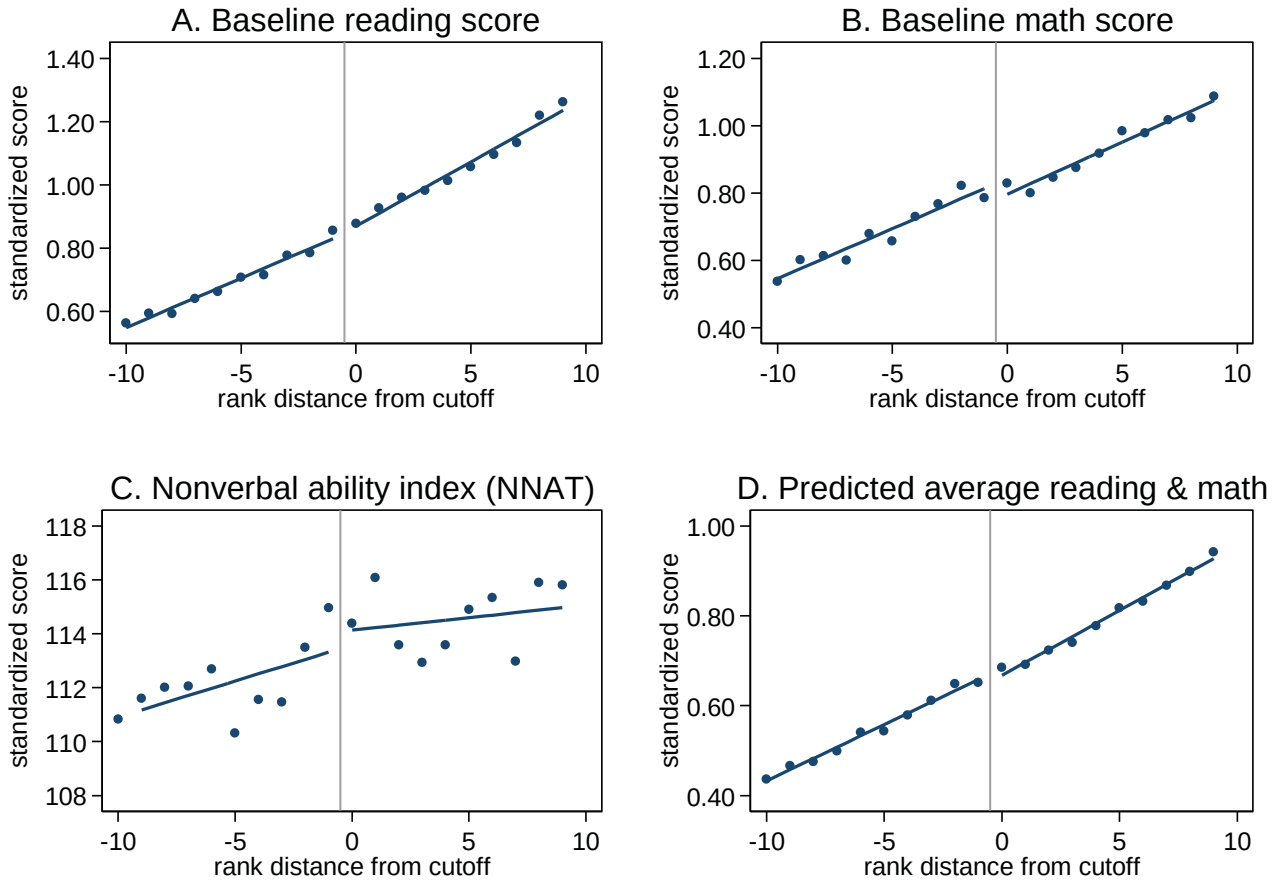

FIGURE 3. STUDENT BASELINE SCORES AND PREDICTED SCORES

Notes: Rank means and fitted values from linear regressions fit separately to students ranked above and below the cutoff for placement in a fourth-grade GHA classroom. Baseline test scores (Panels A and B) are third-grade scores standardized within district and year. Nonverbal ability index (Panel C) is scaled to a national norm with a mean of 100 and standard deviation of 15. Predicted scores (Panel D) are from a regression of fourth-grade scores on third-grade scores, observed student characteristics, and school dummies. Sample in panels A, B and D is 4,144 students whose rank on third-grade scores was +/- 10 from cutoff and who were enrolled in the District in third and fourth grades. Panel C sample is restricted to 2,984 students who took the NNAT because they attended a District school in second grade in 2007-2009.

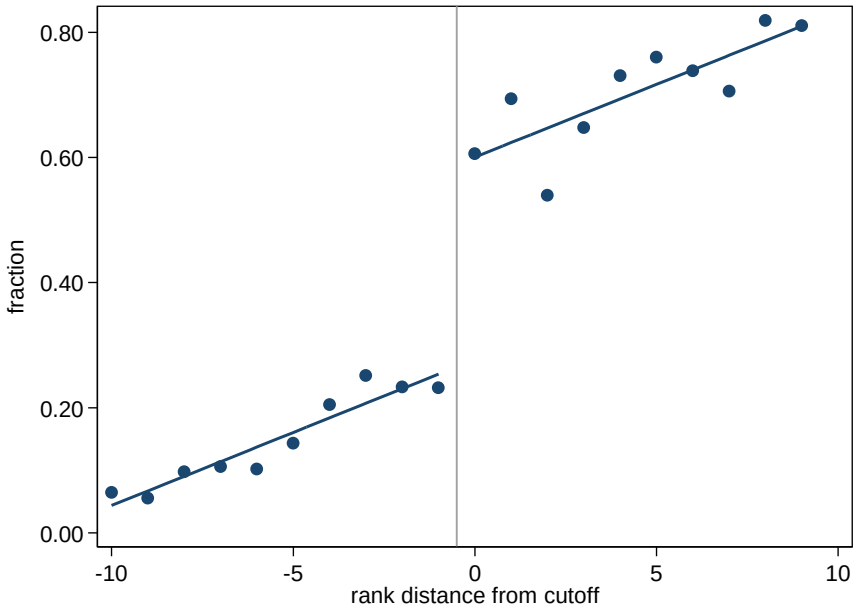

FIGURE 4. FIRST-STAGE RELATIONSHIP FOR PLACEMENT IN GHA CLASSROOM

Notes: Rank means and fitted values from linear regressions fit separately to students ranked above and below the cutoff for placement in a fourth-grade GHA classroom. Sample is 4,144 students whose rank on third-grade scores was +/- 10 from cutoff and who were enrolled in the District in third and fourth grades. All test scores are standardized within district and year. See text for more details.

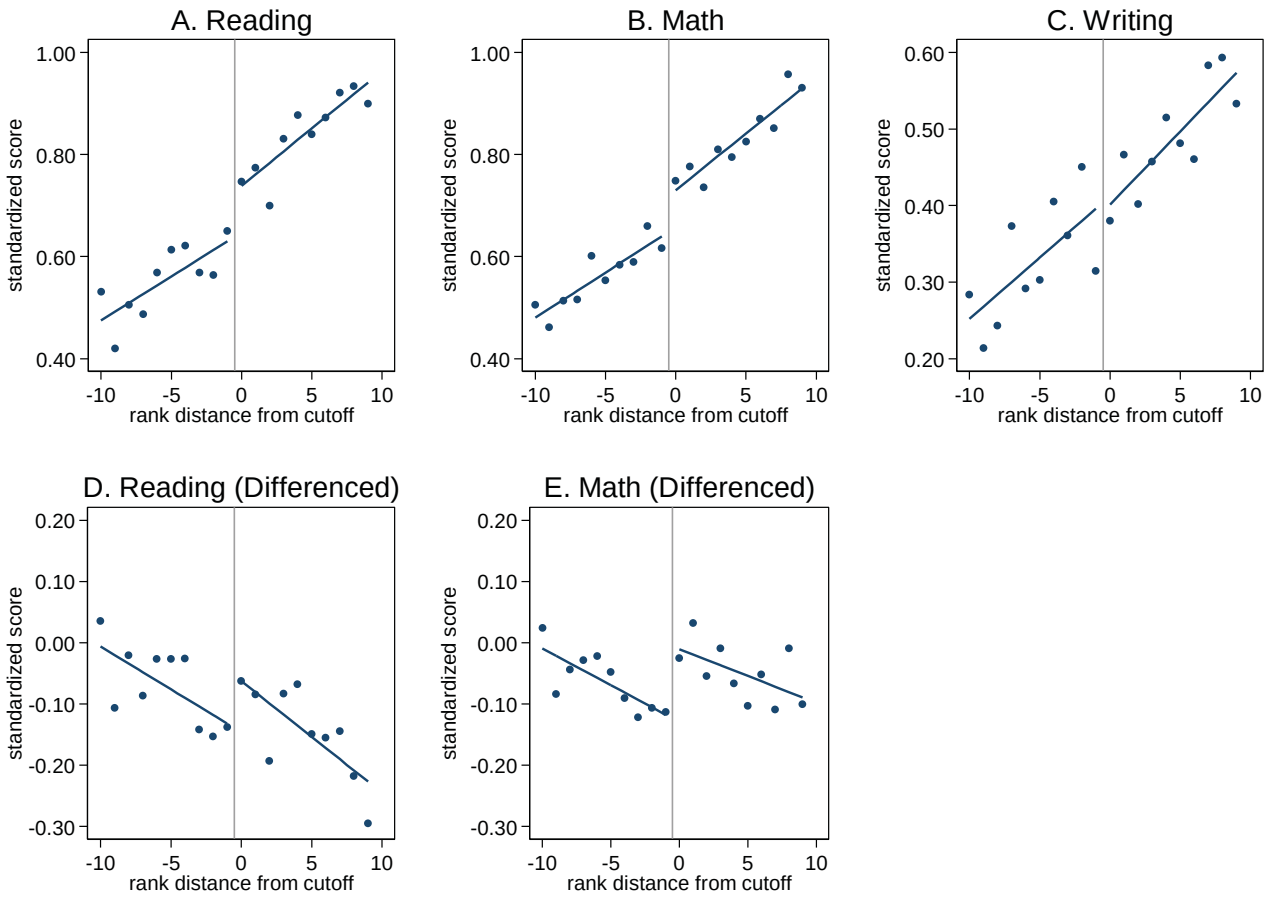

FIGURE 5. FOURTH-GRADE STANDARDIZED TEST SCORES

Notes: Rank means and fitted values from linear regressions fit separately to students ranked above and below the cutoff for placement in a fourth-grade GHA classroom. Sample is 4,144 students whose rank on third-grade scores was +/- 10 from cutoff and who were enrolled in the District in third and fourth grades. All test scores are standardized within district and year.

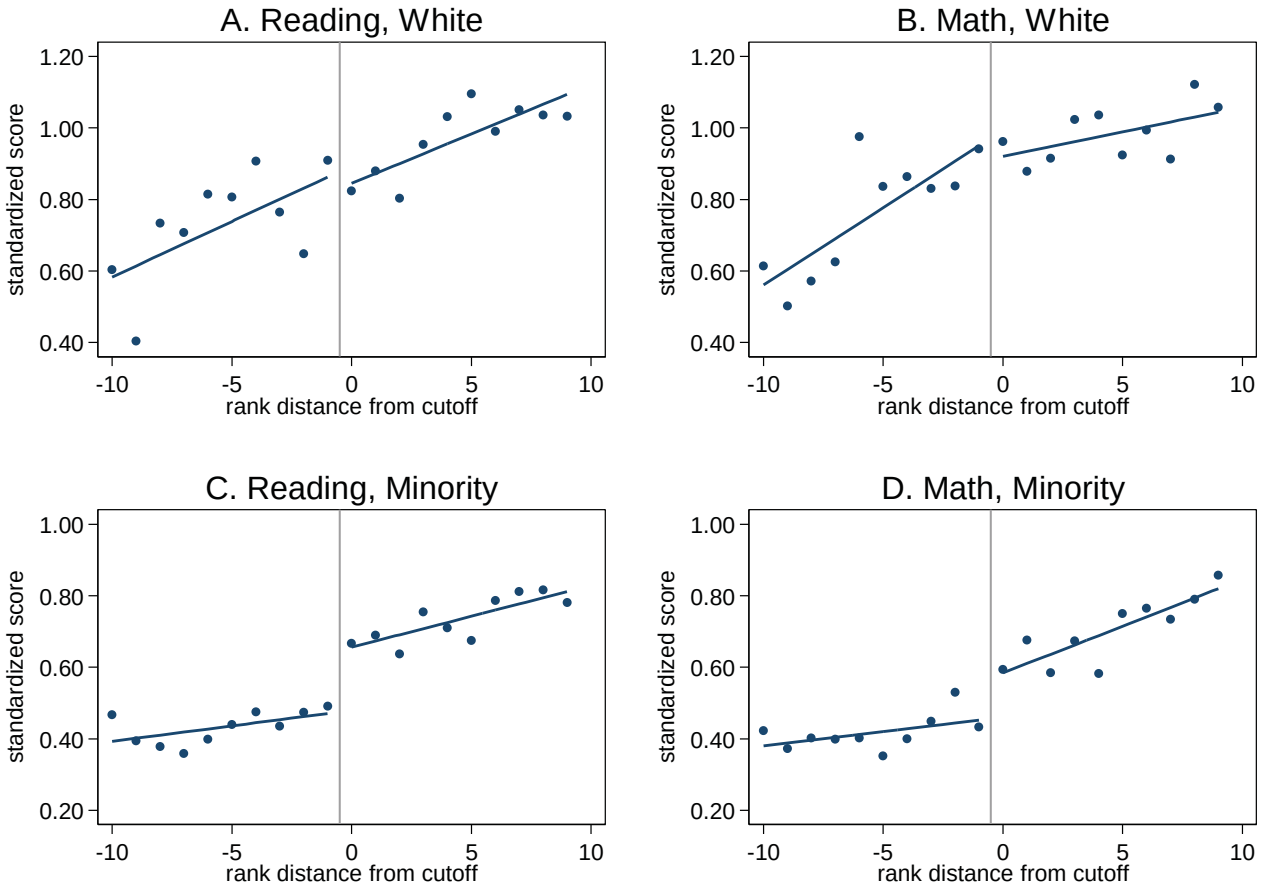

FIGURE 6. FOURTH-GRADE STANDARDIZED TEST SCORES, BY MINORITY STATUS

Notes: Rank means and fitted values from linear regressions fit separately to students ranked above and below cutoff for placement in a fourth-grade GHA classroom. Sample is 1,397 white students and 2,323 black or Hispanic students whose rank on third-grade scores was +/- 10 from cutoff and who were enrolled in the District in third and fourth grades. Test scores are standardized within district and year.

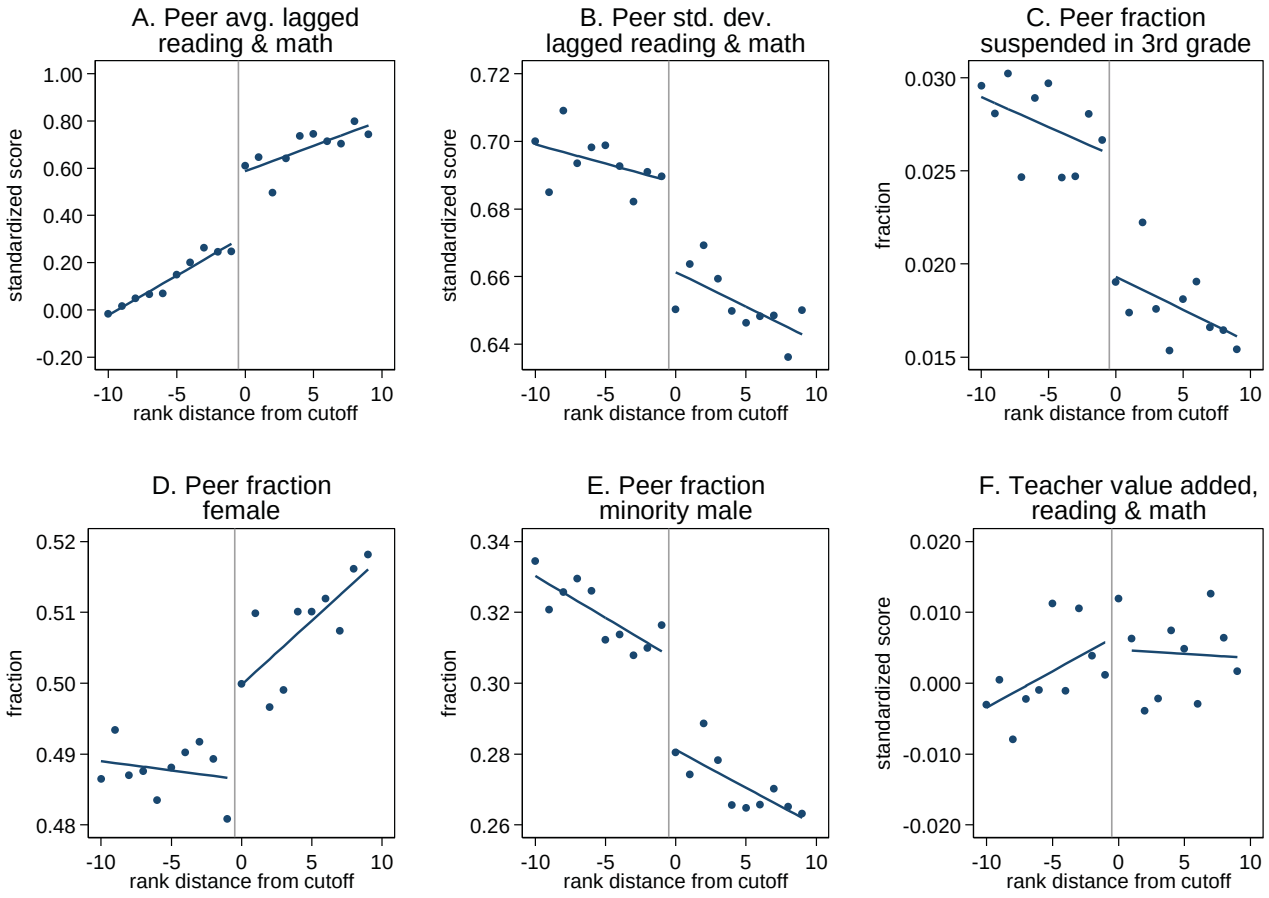

FIGURE 7. CHARACTERISTICS OF FOURTH-GRADE CLASSMATES AND TEACHERS

Notes: Rank means and fitted values from linear regressions fit separately to students ranked above and below the cutoff for placement in a fourth-grade GHA classroom. Sample is 4,144 students whose rank on third-grade scores was +/- 10 from cutoff and who were enrolled in the District in third and fourth grades. All test scores are standardized within district and year. See text for more details.

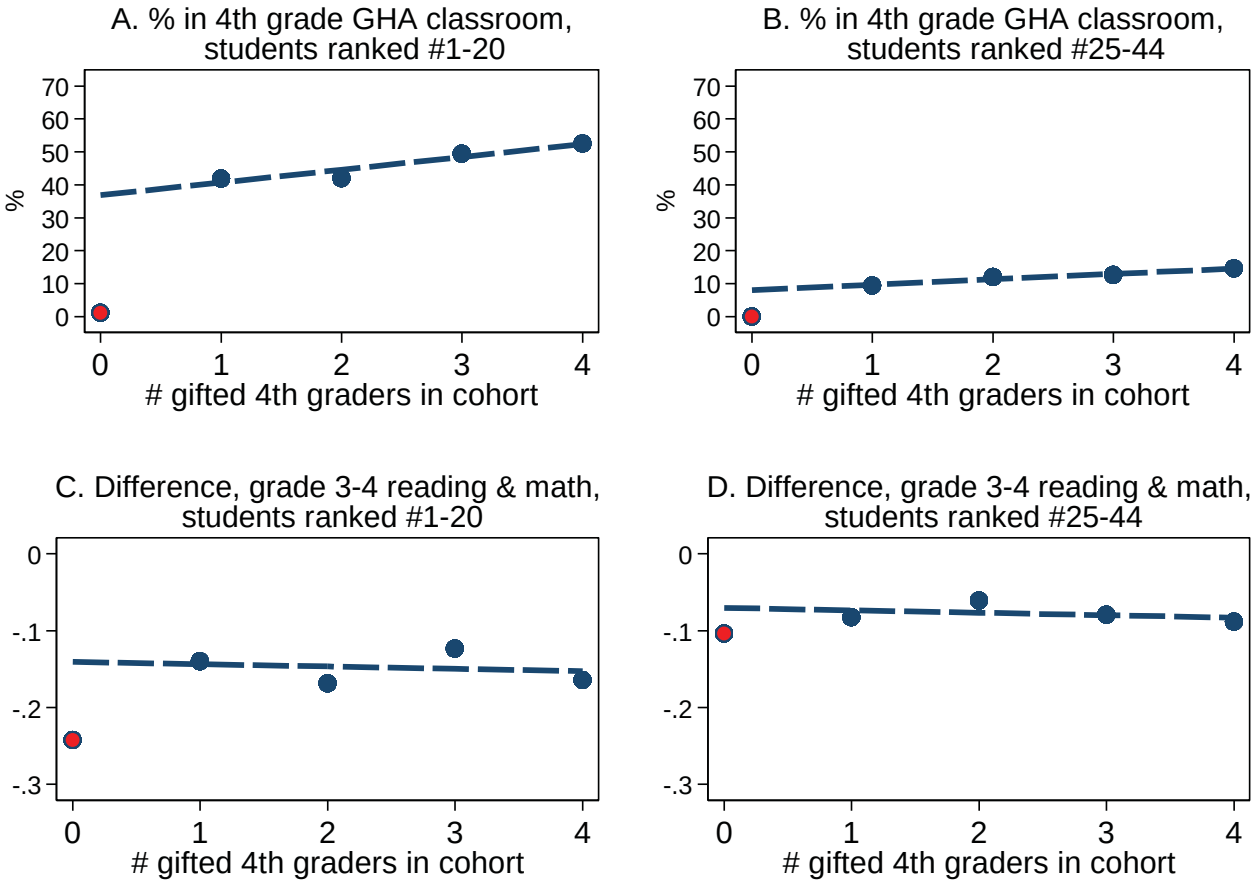

FIGURE 8. BETWEEN-SCHOOL/COHORT ANALYSIS, PLACEMENT IN GHA CLASSROOM AND STUDENT ACHIEVEMENT IN FOURTH GRADE

Notes: Lines are fitted values from linear regressions of GHA placement (Panels A & B) or the average change in reading and math scores between third and fourth grades (Panels C & D), on the number of gifted students in the school/cohort, estimated for students in cohorts with 1-4 gifted students. Regressions include year dummies and cohort-level controls (see text for details). Plotted points are observed GHA placement rates or regression-adjusted means.

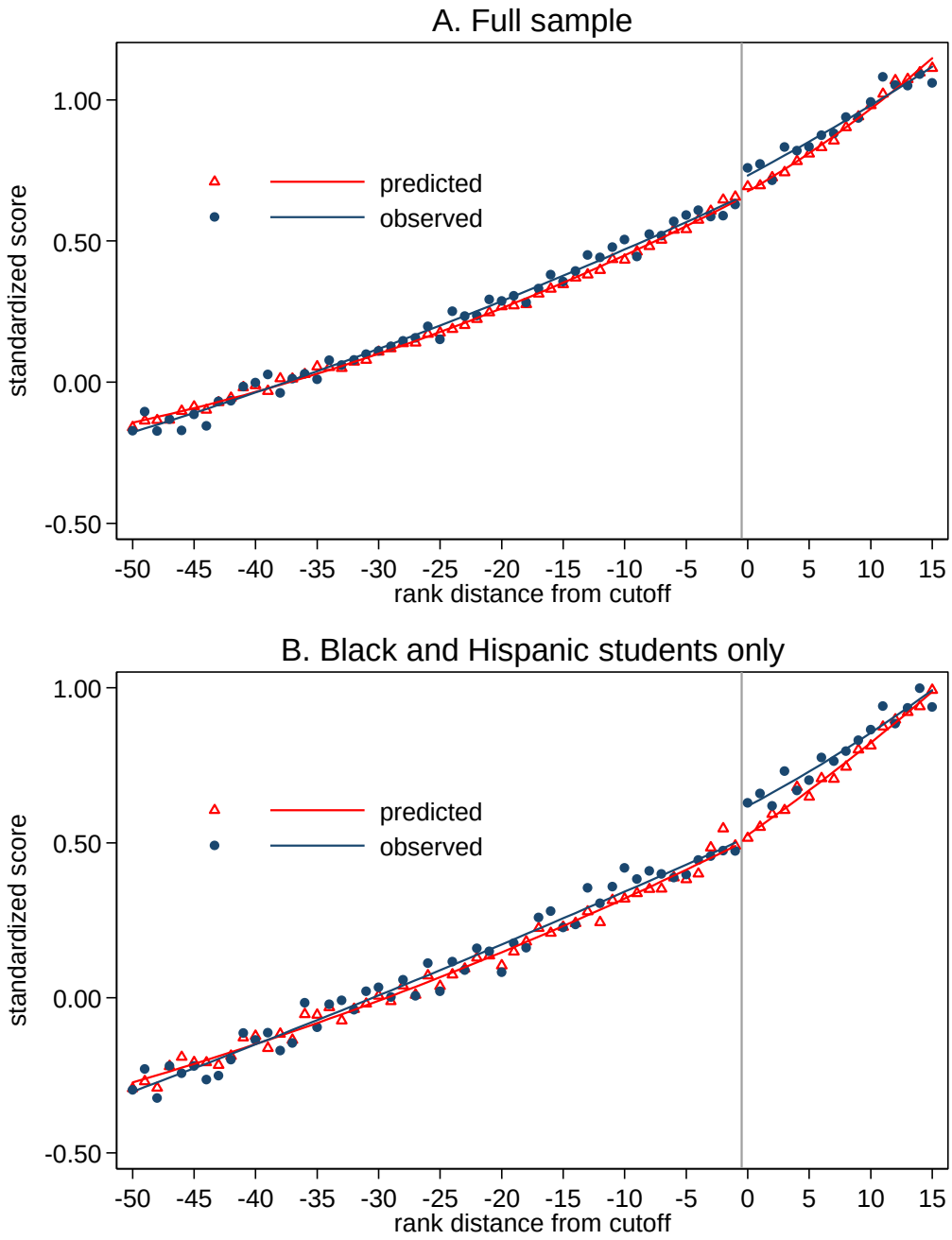

FIGURE 9. PREDICTED AND OBSERVED FOURTH-GRADE TEST SCORES

Notes: Plotted points are means of predicted or observed test scores in fourth grade. Observed scores are averages of standardized fourth-grade reading and math scores. Predicted scores are constructed from a regression of fourth-grade average of reading and math score on: third-grade reading score, third-grade math score, age, dummies for FRL, ELL, race/ethnicity, and gender, and a full set of school dummies, using all fourth-grade students in District in 2009-2012 for whom third-grade scores are available from the previous year. Lines are fitted values from quadratic models for predicted or observed fourth-grade scores, fit separately to students ranked above and below the cutoff. Sample size is 13,379 in Panel A, and 8,101 in Panel B.