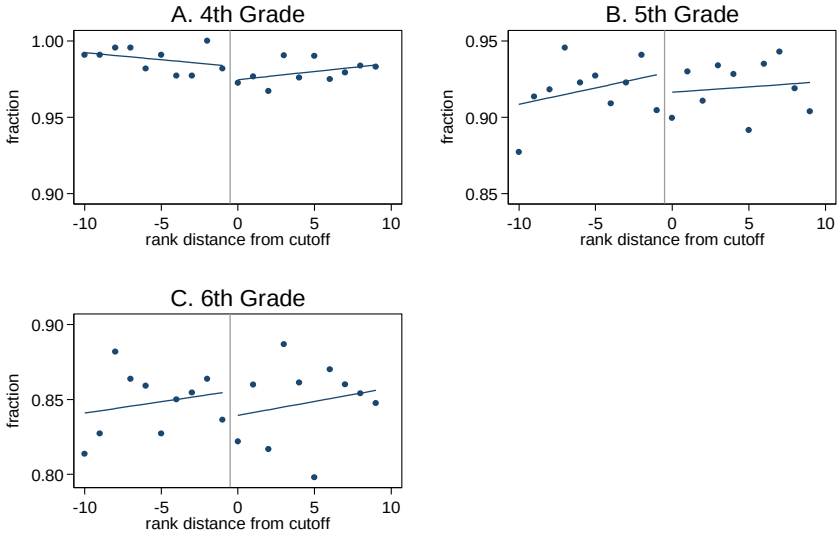
Can Tracking Raise the Test Scores of High-Ability Minority Students?

David Card and Laura Giuliano

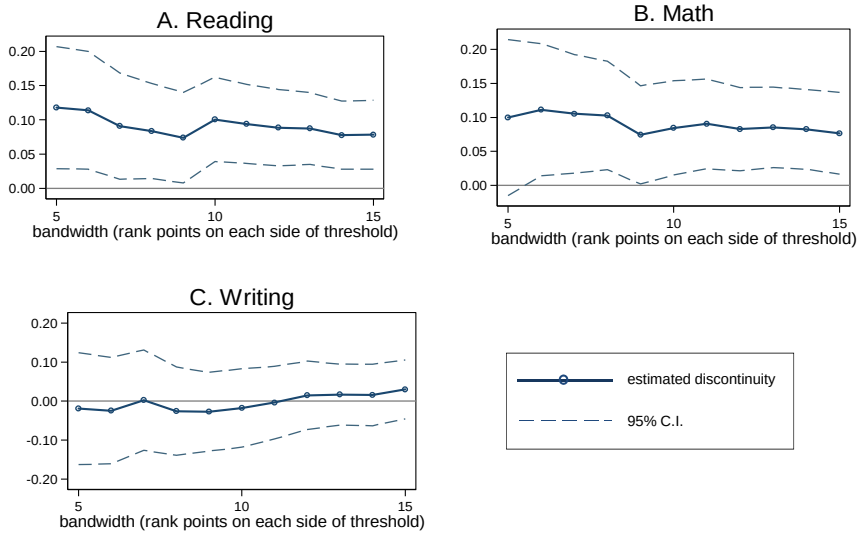
ONLINE APPENDIX

Appendix Figure 1. Fractions of potential high achievers who are observed through end of grades 4, 5, and 6



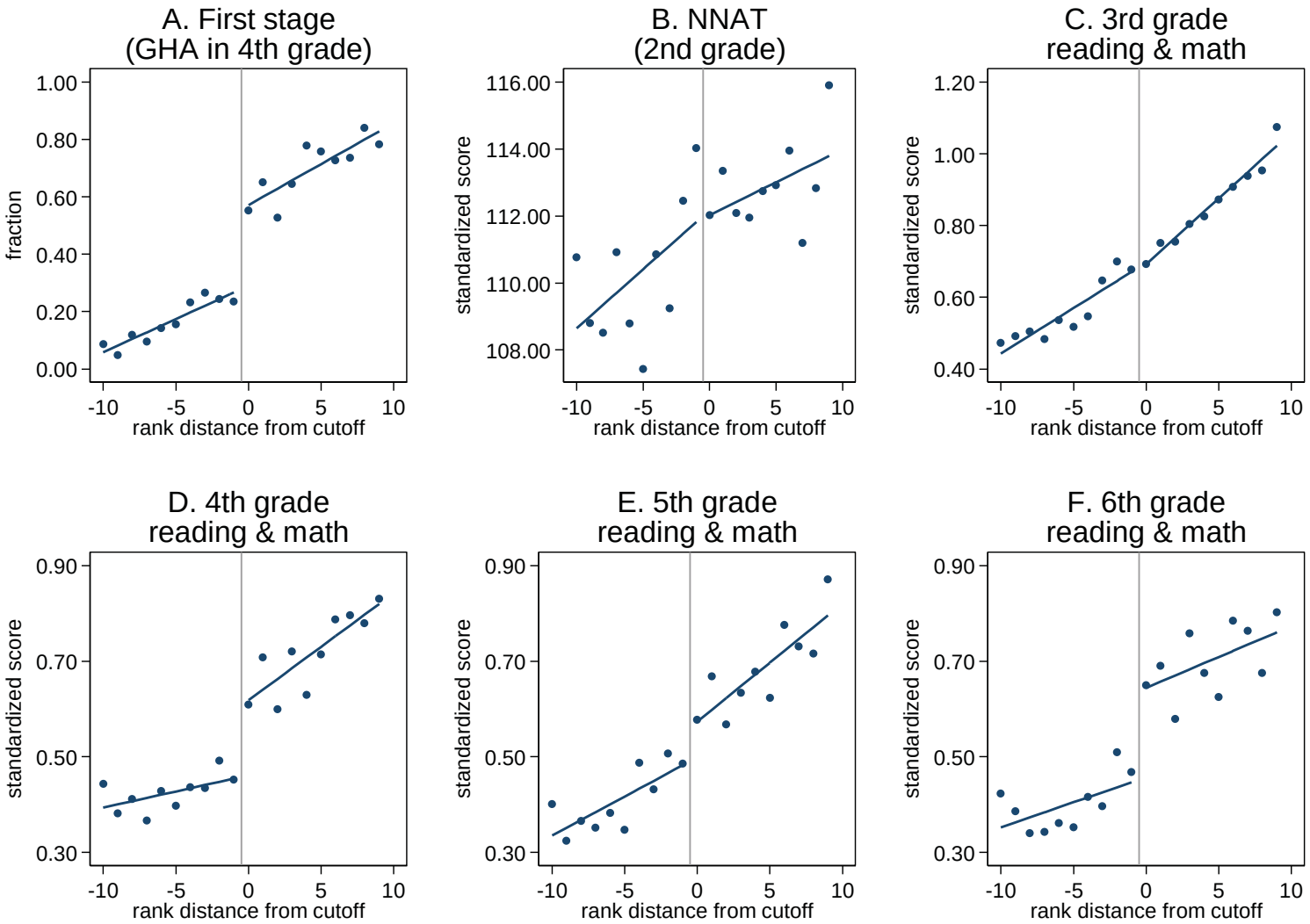
Note: Figures plot means and fitted values from local linear regressions fit separately to students ranked above and below the cutoff for placement in a GHA classroom in fourth grade. Sample is 4,244 students whose rank was +/- 10 from threshold, and who were enrolled in the District in third grade in 2008-2011.

Appendix Figure 2. Estimated discontinuities in fourth-grade scores from local linear regressions with varying bandwidths



Note: Figures plot RD coefficients and 95% confidence intervals from local linear models estimated using bandwidths ranging from 5 to 15 rank points above and below the cutoff for placement in a fourth-grade GHA classroom. All models control for baseline scores, student characteristics, and school dummies, as in row 2 of Table 2.

Appendix Figure 3. Combined reading and math scores of minorities, by grade level



Note: Rank means and fitted values from linear regressions fit separately to students above and below the cutoff for placement in a fourth-grade GHA classroom. Sample is 2,047 black or Hispanic students whose rank on third-grade scores was +/- 10 from cutoff and who were enrolled in the District in third through sixth grade. Panel B is further restricted to 1,473 students who took the NNAT in second grade in 2007-2009. NNAT-based score is scaled to a national norm with a mean of 100 and standard deviation of 15. Reading and math test scores are standardized within district and year before averaging.

Appendix Table 1. OLS and Tobit RD Estimates for Fourth Grade Outcomes

	OLS			Tobit, scores censored at maximum			Tobit, scores censored at 95th percentile		
	4th Grade Reading (1)	4th Grade Math (2)	4th Grade Writing (3)	4th Grade Writing (3)	4th Grade Math (4)	4th Grade Writing (5)	4th Grade Writing (6)	4th Grade Math (7)	4th Grade Writing (8)
1. All students	0.098**	0.081*	-0.005	0.095**	0.081*	-0.005	0.091**	0.088*	0.015
	(0.033)	(0.040)	(0.054)	(0.032)	(0.040)	(0.054)	(0.031)	(0.035)	(0.055)
Sample size	*4,144*	*4,144*	*4,144*	*4,144*	*4,144*	*4,144*	*4,144*	*4,144*	*4,144*
2. White only	-0.026	-0.068	-0.046	-0.029	-0.074	-0.051	-0.035	-0.056	-0.005
	(0.062)	(0.065)	(0.098)	(0.059)	(0.066)	(0.095)	(0.055)	(0.057)	(0.100)
Sample size	*1397*	*1397*	*1397*	*1397*	*1397*	*1397*	*1397*	*1397*	*1397*
3. Minorities	0.176**	0.142**	0.001	0.175**	0.144**	0.004	0.170**	0.147**	0.009
	(0.045)	(0.048)	(0.063)	(0.044)	(0.048)	(0.062)	(0.042)	(0.044)	(0.062)
Sample size	*2323*	*2323*	*2323*	*2323*	*2323*	*2323*	*2323*	*2323*	*2323*

Note: 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). Columns (1)-(3) reproduce the estimates from Table 2, row 2. Columns (4)-(6) report estimates from Tobit models in which the data is assumed to be censored at the maximum value of each test score; in columns (7)-(9) the data is assumed to be censored at the minimum across the four sample cohorts of the cohort 95th percentile. Parentheses contain standard errors, clustered by school. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix Table 2. RD Heterogeneity Analysis for Attrition and Classroom Placement in Fifth and Sixth Grades

	Fifth Grade			Sixth Grade Outcomes		
	Prob. Stayed in District (1)	Prob. in 4th Grade GHA Classroom (2)	Prob. in 5th Grade GHA Classroom (3)	Prob. Stayed in District (4)	Prob. in 4th Grade GHA Classroom (5)	Prob. in 6th Grade Advanced Math (6)
1. Full sample	0.002	0.309**	0.068*	0.002	0.316**	0.052*
	(0.017)	(0.027)	(0.027)	(0.021)	(0.027)	(0.024)
Sample size	*4144*	*3901*	*2768*	*4144*	*3598*	*3598*
2. White only	0.001	0.320**	-0.014	-0.029	0.353**	0.005
	(0.027)	(0.045)	(0.049)	(0.039)	(0.050)	(0.039)
Sample size	*1397*	*1321*	*945*	*1397*	*1187*	*1187*
3. Black and Hispanic only	0.003	0.290**	0.122**	0.022	0.282**	0.077*
	(0.020)	(0.039)	(0.038)	(0.026)	(0.040)	(0.037)
Sample size	*2323*	*2193*	*1552*	*2323*	*2047*	*2047*

Note: 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). The analysis samples in columns 2, 5 and 6 consist of the subset of students from the main analysis sample who were observed in the District through the end of the relevant grade. The samples in column 3 are reduced by roughly 30% due to our inability to match students to fifth-grade classrooms at schools where students rotate between teachers in fifth grade (see Appendix A for details). Parentheses contain standard errors, clustered by school. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix Table 3. Heterogeneity Analysis for Discontinuities in Potential Mechanisms

	Prob(TVA is non-missing)	Teacher value added	Peer avg. lagged test scores	Peer std. dev. lagged test scores	Peer fraction suspended in 3rd grade	Peer fraction female	Peer fraction minorty male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. Full sample	0.02	-0.01	0.86**	-0.08**	-0.02**	0.04*	-0.08**
(n=3685)	(0.05)	(0.03)	(0.06)	(0.03)	(0.01)	(0.02)	(0.02)
2. By Race/Ethnicity							
2a. White	0.06	-0.01	0.88**	-0.08*	-0.01	0.03	-0.05*
(n=1266)	(0.08)	(0.05)	(0.10)	(0.04)	(0.01)	(0.03)	(0.02)
2b. Black and Hispanic	0.01	-0.01	0.83**	-0.05	-0.03*	0.03	-0.10**
(n=2040)	(0.09)	(0.04)	(0.09)	(0.04)	(0.01)	(0.02)	(0.02)
2c. Black Only	0.15	-0.05	0.70**	-0.07	-0.05*	-0.00	-0.06
(n=1017)	(0.14)	(0.06)	(0.18)	(0.06)	(0.02)	(0.04)	(0.04)
2d. Hispanic Only	-0.09	0.06	0.89**	-0.04	-0.01	0.07*	-0.16**
(n=1023)	(0.13)	(0.05)	(0.10)	(0.07)	(0.01)	(0.03)	(0.03)
3. Black and Hispanic Only, by FRL Status							
3a. FRL-eligible	-0.09	-0.06	0.81**	-0.01	-0.06**	0.03	-0.12**
(n=1340)	(0.14)	(0.07)	(0.13)	(0.06)	(0.02)	(0.04)	(0.04)
3b. Non-FRL-eligible	0.04	0.05	0.86**	-0.12*	0.00	0.05	-0.10**
(n=700)	(0.10)	(0.04)	(0.10)	(0.05)	(0.01)	(0.03)	(0.03)
4. Black and Hispanic Only, by Number of Gifted Students in School/Cohort							
4a. 1-4 Gifted	0.33**	-0.02	0.79**	-0.03	-0.05*	-0.01	-0.05
(n=931)	(0.12)	(0.06)	(0.12)	(0.04)	(0.02)	(0.03)	(0.04)
4b. 5 or more Gifted	-0.13	-0.00	0.83**	-0.03	-0.01	0.07*	-0.14**
(n=1085)	(0.13)	(0.05)	(0.13)	(0.07)	(0.01)	(0.03)	(0.03)
5. Black and Hispanic Only, by Gender							
5a. Girls	-0.01	-0.06	0.80**	-0.06	-0.03	0.07*	-0.10**
(n=1073)	(0.11)	(0.06)	(0.08)	(0.07)	(0.02)	(0.03)	(0.03)
5b. Boys	0.01	0.03	0.84**	-0.02	-0.03	0.02	-0.12**
(n=967)	(0.13)	(0.06)	(0.15)	(0.07)	(0.02)	(0.03)	(0.04)

Note: Estimates from two-stage least squares RD models with school and year fixed effects and student controls, as in Table 2, row 2 (see Table 2 note for details). Estimation samples (and indicated sample sizes) exclude students for whom teacher value added cannot be estimated because the teacher is only observed in one year. (See Appendix B for description of the model used to estimate TVA.) In all models the first-stage model is for the probability of being in the fourth-grade GHA classroom (first-stage estimates are reported in column 3 of Table 3). Parentheses contain standard errors, clustered by school. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix Table 4. Estimated Impact of Classroom Characteristics on Gain Scores in Reading and Math

	Reading						Math					
	Full sample		White only		Minority only		Full sample		White only		Minority only	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Classroom characteristic:	Univariate model	Joint model	Univariate model	Joint model	Univariate model	Joint model	Univariate model	Joint model	Univariate model	Joint model	Univariate model	Joint model
1. Teacher value added	0.40**	0.40**	0.41**	0.42**	0.39**	0.39**	0.69**	0.63**	0.73**	0.67**	0.67**	0.61**
	(0.03)	(0.04)	(0.05)	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)	(0.07)	(0.06)	(0.05)	(0.05)
2. Average of peers' lagged test scores	0.04**	-0.00	0.04+	-0.00	0.05**	0.01	0.13**	0.09**	0.13**	0.08**	0.12**	0.09**
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
3. Std. dev. of peers' lagged test scores	-0.04+	-0.04+	-0.02	-0.03	-0.06*	-0.06*	0.09**	0.15**	0.09*	0.14**	0.10**	0.16**
	(0.02)	(0.02)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.04)	(0.04)	(0.04)
4. Peer fraction suspended in 3rd grade	-0.06	-0.02	-0.24	-0.20	-0.00	0.05	-0.28+	-0.14	-0.41	-0.21	-0.21	-0.09
	(0.11)	(0.10)	(0.27)	(0.28)	(0.11)	(0.10)	(0.16)	(0.14)	(0.27)	(0.26)	(0.16)	(0.15)
5. Peer fraction female	-0.01	0.08	0.00	0.07	-0.02	0.09	-0.03	-0.01	-0.01	0.01	-0.05	-0.01
	(0.05)	(0.06)	(0.08)	(0.09)	(0.06)	(0.08)	(0.06)	(0.06)	(0.09)	(0.10)	(0.06)	(0.07)
6. Peer fraction minority male	0.06	0.13*	0.07	0.13+	0.06	0.15+	-0.03	0.07	-0.07	0.04	-0.00	0.09
	(0.05)	(0.06)	(0.07)	(0.08)	(0.05)	(0.08)	(0.06)	(0.06)	(0.09)	(0.09)	(0.07)	(0.08)

Notes. Coefficients from models of test scores gains between 3rd and 4th grade, estimated for all students enrolled in a regular district elementary school in fourth grade in 2009-2012. All models include school fixed effects, a dummy for whether the student is in a GHA classroom, and controls for student's age, gender, race/ethnicity, FRL and ELL status, and median household income in the student's neighborhood. Models in odd-numbered columns include only one classroom characteristic. Models in even-numbered columns simultaneously control for all six classroom characteristics. The full sample has 47,890 observations; the white student sample has 14,771 observations and the minority student sample has 29,529 observations. All estimation samples exclude students for whom teacher value added cannot be estimated because the teacher is only observed in one year. (See Appendix B for description of the model used to estimate TVA.) Parentheses contain standard errors, clustered by school. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix Table 5. Estimated Achievement Gaps in Third Grade, by Race and Ethnicity

	(1)	(2)	(3)	(4)	(5)	(6)
Black	-0.725**	-0.462**	-0.461**	-0.287**	-0.280**	-0.213**
	(0.029)	(0.020)	(0.020)	(0.012)	(0.009)	(0.008)
Hispanic	-0.320**	-0.263**	-0.262**	-0.160**	-0.208**	-0.153**
	(0.017)	(0.013)	(0.013)	(0.009)	(0.008)	(0.008)
Asian	0.117**	-0.003	-0.006	0.037*	0.018	0.045**
	(0.023)	(0.015)	(0.015)	(0.015)	(0.016)	(0.016)
FRL eligible				-0.333**		-0.244**
				(0.013)		(0.008)
Control for Ability Index	none	linear	quadratic	quadratic	quadratic	quadratic
school/cohort FEs	no	no	no	no	yes	yes

Note: Estimated coefficients from regressions of average reading and math scores in third grade on race/ethnicity dummies, controlling for nonverbal ability index. Ability index is constructed from second-grade NNAT score and is scaled to a national norm with a mean of 100 and standard deviation of 15. Sample is 76,727 students who took the NNAT because they were enrolled in the District in second grade between 2005-2009, and who were enrolled in District for third grade the following year. Omitted race category is white. Parentheses contain standard errors, clustered by school. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix Table 6. RD Heterogeneity Analysis for Unexcused Absences and Suspensions in Grades 4-6

	Prob. >1 Unexcused Absence, Grades 4-6		Prob. Suspended >1 time, Grades 4-6	
	(1) mean below cutoff	(2) RD estimate	(3) mean below cutoff	(4) RD estimate
1. Full sample	0.66	-0.06+	0.08	-0.05**
(n=3596)		(0.03)		(0.01)
2. By Race/Ethnicity				
2a. White	0.53	-0.00	0.03	-0.02
(n=1187)		(0.06)		(0.02)
2b. Black and Hispanic	0.74	-0.08*	0.11	-0.06**
(n=2045)		(0.04)		(0.02)
2c. Black Only	0.73	-0.06	0.16	-0.08*
(n=1060)		(0.06)		(0.04)
2d. Hispanic Only	0.74	-0.12*	0.05	-0.03
(n=985)		(0.06)		(0.03)
3. Black and Hispanic Only, by FRL Status				
3a. FRL-eligible	0.78	-0.02	0.14	-0.08*
(n=1376)		(0.04)		(0.03)
3b. Non-FRL-eligible	0.64	-0.23**	0.04	-0.04
(n=669)		(0.08)		(0.03)
4. Black and Hispanic Only, by Number of Gifted Students in School/Cohort				
4a. 1-4 Gifted	0.76	-0.03	0.16	-0.07+
(n=950)		(0.05)		(0.04)
4b. 5 or more Gifted	0.71	-0.11	0.06	-0.04
(n=1068)		(0.07)		(0.03)
5. Black and Hispanic Only, by Gender				
5a. Girls	0.73	-0.01	0.06	-0.02
(n=1074)		(0.05)		(0.03)
5b. Boys	0.75	-0.13+	0.16	-0.10*
(n=971)		(0.07)		(0.04)

Note: Odd columns contain group means among students whose rank is up to ten places below the school-specific cutoff for placement in a GHA classroom in fourth grade; even columns contain estimated discontinutities at the cutoff, from models with controls 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. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix A: Matching Students to Classrooms and Identification of GHA Classrooms

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

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

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

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

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

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

We used a similar procedure to match students to classrooms in fifth grade and to construct an identifier for being in a GHA classroom in fifth grade. Because the practice of rotating classrooms is more common in fifth grade, only 71% of students observed through fifth grade could be matched to a classroom. This is the reason for the reduced sample size in row 1, column 3 of Table 3.

Construction of High Achiever Sample and Estimation of Cutoff Scores

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

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

In principle, the cutoff for placement in the GHA classroom of a given cohort is the test score of the lowest-scoring non-gifted child in the GHA classroom, or the score just above that of the highest-scoring child in a regular classroom. But non-compliance can cause these two scores to differ—and use of either one of these measures leads to misleading mappings between relative rank and placement in the GHA class. To circumvent this problem, we employ a two-step procedure that starts with an initial estimate based on the number of seats in the classroom, and then makes adjustments that reduce misclassification due to measurement error.

Specifically, for each of school/cohort, we estimated the cutoff rank for placement in the GHA classroom as follows:

1. First, using the District's prescribed rule, we assigned a within-cohort rank to each non-gifted fourth-grade student with non-missing third-grade test scores. The rule is a lexicographic formula that first groups students based on their "achievement levels" on the reading and math portions of the third grade statewide achievement test. These achievement levels range from 1-5 and are based on the scale scores (which range from 100 to 500), with cutoffs set each year by the state. Students who achieve level 5 (the highest) in *both* reading and math are given highest priority, followed by students with a level 5 in reading and a 4 in math; those with a 4 in reading and 5 in math; those with a 4 in both reading and math, and so on. Within each of these groups, students are ranked using the sum of their scale scores in reading and math.
2. Next, we calculated an initial estimate of the cutoff rank, c, as the rank of the N^{th} ranked non-gifted student, where N is the number of non-gifted students in the GHA classroom.
3. Classroom reassignments and errors in matching students to classrooms lead to measurement error in the classroom size N and thus in the initial cutoff estimate c. To reduce this measurement error, we replaced c with $c' \in (c\text{-}10, c\text{+}9)$, where c' is chosen using an iterative procedure to minimize the misclassification rate of students whose scores are outside an interval around the potential cutoff. Specifically, letting $c'=c$ be the initial

estimate of the cutoff rank, we replaced c' with c'+1 if $\sum_{r=c'-3}^{c'-2} T_r < \sum_{r=c'+1}^{c'+2} T_r$ or with c'-1 if $\sum_{r=c'-3}^{c'-2} T_r > \sum_{r=c'+1}^{c'+2} T_r$, where T_r is a dummy variable for the student with rank r being in the GHA classroom. We repeated this step until no further reduction in mismatch was possible.

4. After estimating a cutoff for each cohort, we eliminated cohorts where there was still substantial mismatch or "non-compliance" with the assignment rule based on the estimated cutoff. For each cohort we examined placement rates of students with $r \in (c'-10, c'+9)$, and we kept cohorts for which a one-tailed test of H_0: $\left(E(T_r|r \geq c') - E(T_r|r < c')\right) = 0$ has a z-statistic of >1. This resulted in our estimation sample of 4,144 fourth grade students in 220 school/cohorts.

Finally, we investigated the causes of mismatch and the determinants of being excluded from our sample. Our analysis showed four patterns. First, the rate at which cohorts are dropped from the sample due to mismatch is highest in the first year that the rule was prescribed by the District—suggesting some non-compliance due to weak initial enforcement. We dropped 60% of the 2009 fourth-grade cohorts compared to 47% of the cohorts in 2010 and 32% in 2011 and 2012. Second, the mismatch rate is significantly higher in cohorts where the measured class size is larger than the target class size of 20-24 students. Third, mismatch is also higher in cohorts where we are missing test scores for students in the GHA classroom (which may occur, for example, when a student transfers into the District in fourth grade from elsewhere in the state). The effects of measured class size and missing test scores on mismatch both point to measurement error and misclassification as an explanation for much of the "non-compliance" with our estimated cutoffs. Finally, and importantly, the likelihood of being excluded from our sample is *not* significantly correlated with school characteristics such as the fraction of students who are FRL eligible or the fraction who are black or Hispanic.

Comparison with Alternative Methods for Identifying Cutoff Scores

As a check on our procedure, and to compare the robustness of our main findings to other possible way of identifying the cutoff score for entry to the GHA class, we re-calculated the cutoff scores using three alternative procedures. The first (alternative procedure 1), sets the cutoff score as the score that, when used as a cutoff threshold, yields the highest fraction of correct assignments (i.e., the highest compliance rate) among non-gifted students in the school/cohort ranked 1-50 using the District's ranking formula. (In cases where 2 or more scores yield the same fraction of correct assignments we choose the highest). The second (alternative procedure 2) sets the cutoff as the lowest rank among all non-gifted students assigned to the GHA class, with the proviso that the cutoff must be no higher than 50 (otherwise we exclude the entire school/cohort). The third (alternative procedure 3) sets the

cutoff as 1 plus the highest rank among all non-gifted students who are not assigned to the GHA class, with the proviso that the top-ranked student must be assigned to the GHA class (otherwise we exclude the entire school/cohort).

The results from using each of these procedures are summarized in the following series of tables and figures. We present figures that show the probability of placement in a GHA class, the relationships with baseline reading and math scores, and the relationships with fourth-grade reading and math scores, as well as tables showing estimation results for the corresponding RD models.

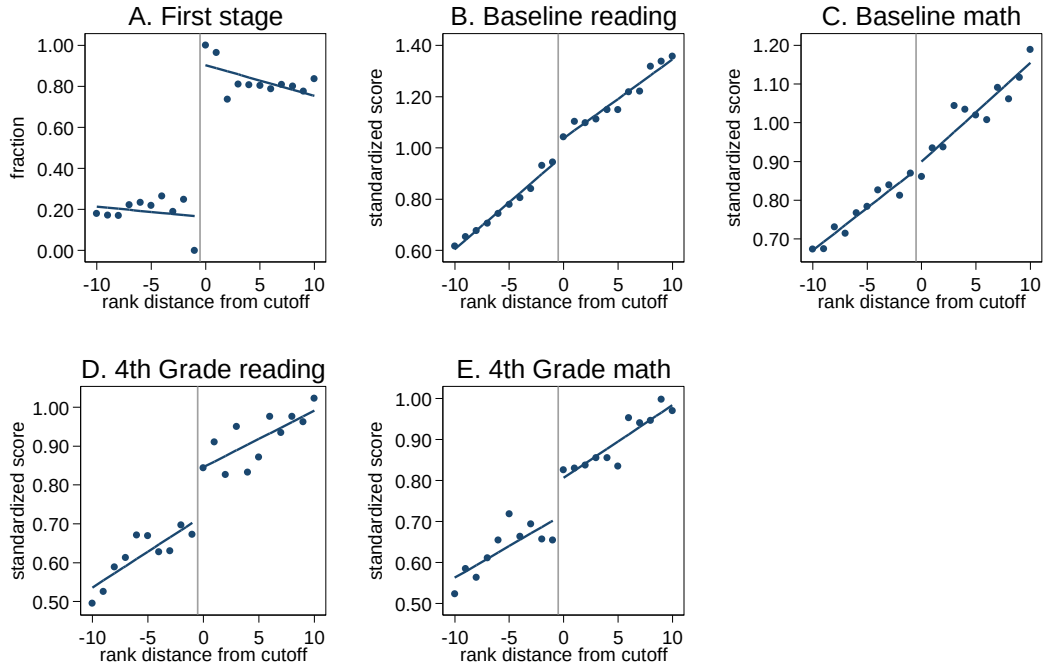
Alternative procedure 1 (Appendix Figure A1 and Table A1) yields a first-stage relationship between relative rank and the probability of placement in a GHA class that shows a large jump at the cutoff, but is downward sloping to the right and left of the cutoff. This arises because a procedure that maximizes the correct classification rate will always choose a cutoff such that the student just to the right of the cutoff is assigned to the GHA class, and the student just to the left is not. By contrast, our preferred procedure avoids this problem by maximizing the compliance rate for students *outside* an interval around the threshold.

Procedure 1 also generates a discontinuous relationship between relative ranks and baseline reading scores. The estimated reduced form impacts on fourth grade scores using this procedure are positive, but show some sensitivity to the controls used in the RD model (unlike the reduced-form impacts from our preferred procedure).

By construction alternative procedure 2 (Appendix Figure A2 and Table A2) yields a first-stage relationship that shows zero probability of placement in a GHA class for all students ranked below the cutoff and a 100% probability for the student in each school/cohort ranked just above the cutoff. However, the average placement rate for students ranked 2-5 above the cutoff is relatively flat at about 40%. This procedure generates a positive discontinuity in baseline reading scores and a negative discontinuity in baseline math. The reduced-form impacts on fourth-grade reading and math are positive, but smaller in magnitude than the estimates from our preferred procedure, and also sensitive to specification.

By construction alternative procedure 3 (Appendix Figure A3 and Table A3) yields a first-stage relationship that shows a 100% probability of placement in a GHA class for all students ranked above the cutoff and a zero probability for the student ranked just below the cutoff in each school/cohort. However, the average placement rate for students ranked 2-5 below the cutoff is 55-60%. This procedure generates relatively small and insignificant discontinuities in baseline reading and math scores. The reduced-form impacts on fourth-grade reading and math are positive and significant, about the same magnitude as the estimates from our preferred procedure, and not very sensitive to choice of specification for the RD model.

Appendix Figure A1. GHA placement, baseline scores, and fourth grade outcomes by rank, cutoff estimated using alternative procedure 1

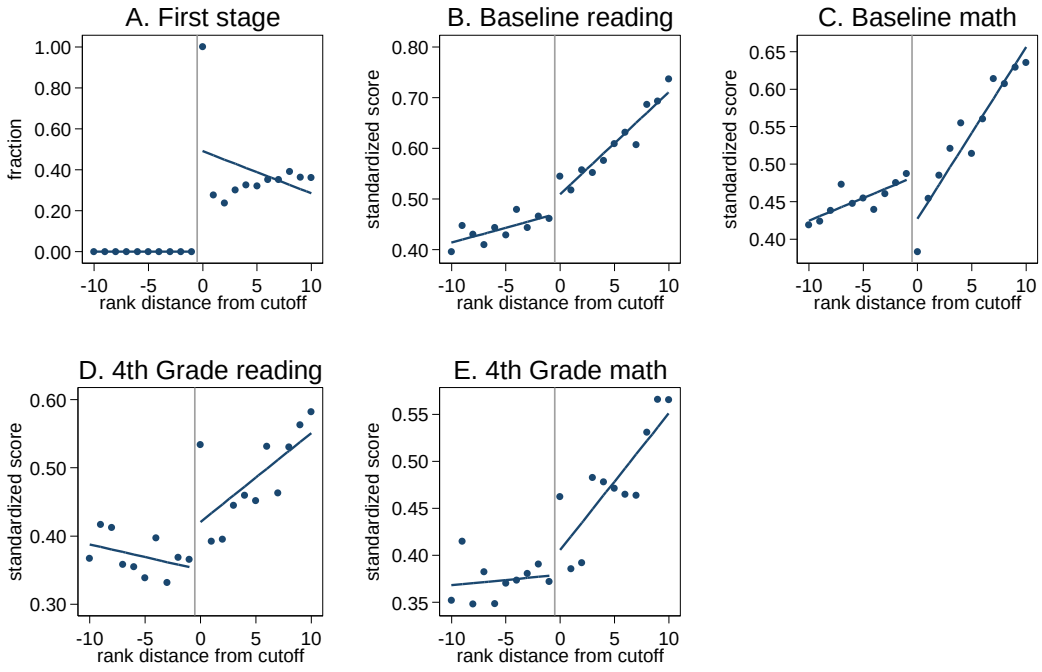


Appendix Table A1. Regression discontinuity estimates for GHA placement, baseline scores, and fourth grade outcomes; cutoff estimated using alternative procedure 1

	Baseline achievement		First stage	Reduced-form estimates		
	3rd grade reading	3rd grade math	Prob. in GHA classroom	4th grade reading	4th grade math	4th grade writing
	(1)	(2)	(3)	(4)	(5)	(6)
1. No controls	0.066*	0.009	0.741**	0.125**	0.089*	0.091†
	(0.028)	(0.035)	(0.014)	(0.032)	(0.034)	(0.050)
2. School & year fixed effects; student controls	0.056*	0.000	0.735**	0.087**	0.066*	0.064
	(0.025)	(0.032)	(0.014)	(0.029)	(0.029)	(0.046)
3. Differenced specification	--	--	--	0.054†	0.058†	--
				(0.031)	(0.032)	
Sample size	*6,029*	*6,029*	*6,029*	*6,029*	*6,029*	*6,009*

Note: Estimates from models of dependent variables as a function of a student's rank (within school-year cohort) on third-grade test scores. See Table 2 note for details on model specifications. 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. The cutoff score is estimated as the score that yields the highest fraction of correct assignments among non-gifted students in the school cohort ranked 1-50 using the District's ranking formula.

Appendix Figure A2. GHA placement, baseline scores, and fourth grade outcomes by rank, cutoff estimated using alternative procedure 2

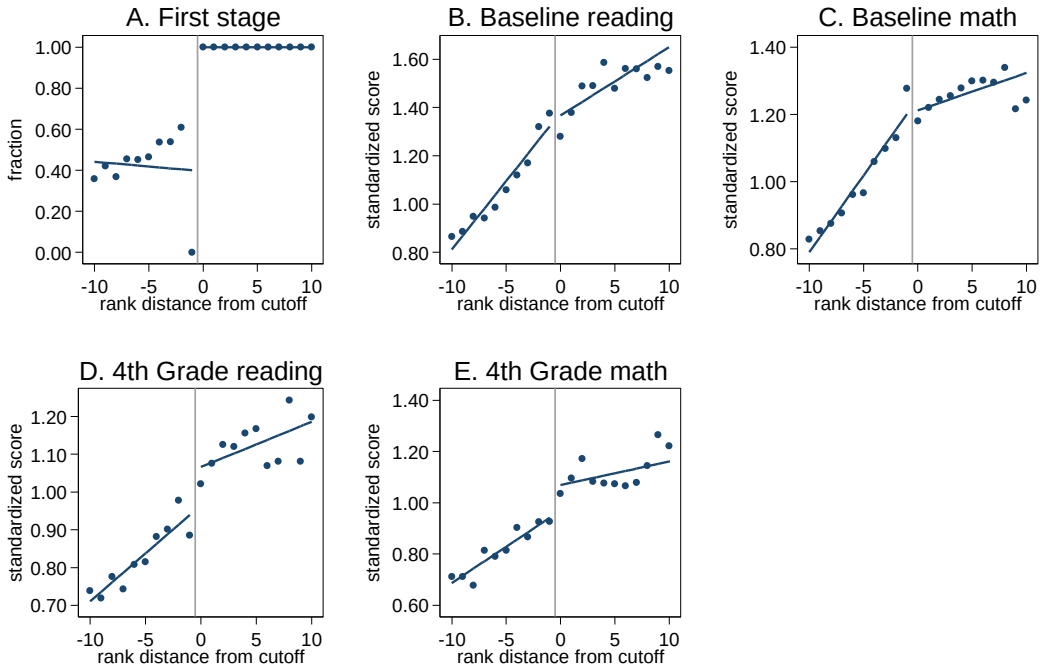


Appendix Table A2. Regression discontinuity estimates for GHA placement, baseline scores, and fourth grade outcomes; cutoff estimated using alternative procedure 2

	Baseline achievement		First stage	Reduced-form estimates		
	3rd grade reading	3rd grade math	Prob. in GHA classroom	4th grade reading	4th grade math	4th grade writing
	(1)	(2)	(3)	(4)	(5)	(6)
1. No controls	0.037	-0.058+	0.492**	0.070**	0.026	0.057
	-0.024	-0.03	-0.015	-0.026	-0.031	-0.041
2. School & year fixed effects; student controls	0.046**	-0.053*	0.486**	0.061*	0.046†	0.051
	-0.017	-0.026	-0.015	-0.024	-0.026	-0.039
3. Differenced specification	--	--	--	0.043	0.064*	--
				-0.027	-0.029	
Sample size	*6,578*	*6,578*	*6,578*	*6,578*	*6,578*	*6,578*

Note: Estimates from models of dependent variables as a function of a student's rank (within school-year cohort) on third-grade test scores. See Table 2 note for details on model specifications. 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. The cutoff rank is estimated as the lowest rank among all non-gifted students assigned to the GHA class (see text of Appendix A for details).

Appendix Figure A3. GHA placement, baseline scores, and fourth grade outcomes by rank, cutoff estimated using alternative procedure 3



Appendix Table A3. Regression discontinuity estimates for GHA placement, baseline scores, and fourth grade outcomes; cutoff estimated using alternative procedure 3

	Baseline achievement		First stage	Reduced-form estimates		
	3rd grade reading	3rd grade math	Prob. in GHA classroom	4th grade reading	4th grade math	4th grade writing
	(1)	(2)	(3)	(4)	(5)	(6)
1. No controls	-0.012	-0.035	0.605**	0.102**	0.099*	0.029
	(0.047)	(0.041)	(0.018)	(0.039)	(0.039)	(0.048)
2. School & year fixed effects; student controls	0.037	-0.015	0.572**	0.110**	0.108**	0.051
	(0.038)	(0.037)	(0.018)	(0.033)	(0.034)	(0.045)
3. Differenced specification	--	--	--	0.088*	0.111**	--
				(0.037)	(0.040)	
Sample size	*4,844*	*4,844*	*4,844*	*4,844*	*4,844*	*4,844*

Note: Estimates from models of dependent variables as a function of a student's rank (within school-year cohort) on third-grade test scores. See Table 2 note for details on model specifications. 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. The cutoff rank is estimated as the 1 plus the highest rank among all non-gifted students who are not assigned to the GHA class (see text of Appendix A for details).

Appendix B: Construction of Teacher Value-Added

To construct a value-added model of teacher quality, we use data on all teachers who are observed teaching fourth grade in two or more years between 2005 and 2012, and we estimate teacher fixed effects from a model of average 4th-grade test scores in reading and math. Specifically, we estimate:

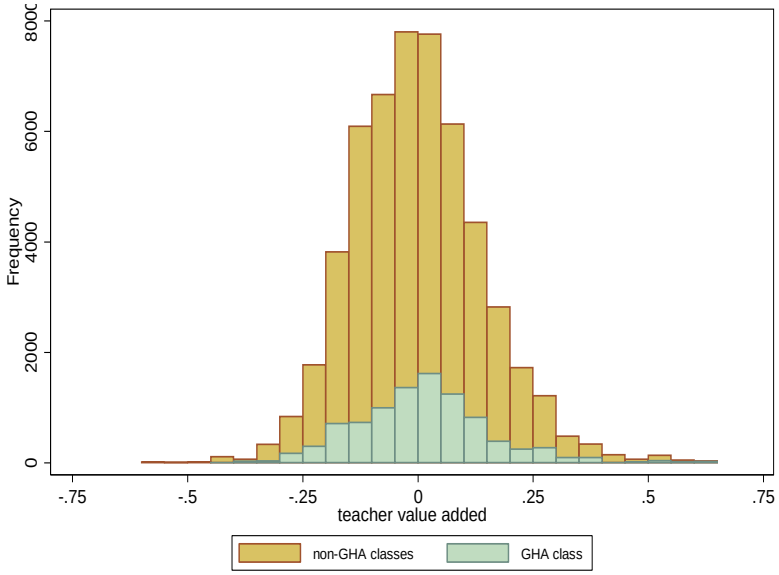
$$\text{(B1)} \qquad Y_{\text{isjt}} = \beta_0 + Y_{\text{isjt}-1}\beta_1 + Y^2_{isjt-1}\beta_2 + X_{\text{isjt}}\beta_3 + S_{\text{isjt}}\beta_4 + T_{\text{isjt}}\theta_j + \varepsilon_{\text{isjt}},$$

where Y_{isjt} is the average of standardized test scores in reading and math for student *i* at school *s* with teacher *j* in year *t*. T_{isjt} is a vector of teacher dummy variables, and the parameters θ_j are the estimates of teacher value added (TVA). We control for a vector of student characteristics, X_{isjt}, that includes dummy variables for student gender, race, ethnicity, and for FRL, ELL, and gifted status. We also control for a vector of school/cohort and classroom characteristics, S_{isjt}, that includes: dummies for being in a GHA classroom and for being in a non-GHA special-education classroom; interactions of the GHA classroom dummy with the race indicators; and school/cohort-level controls for: the total number of students enrolled in fourth grade; the number who are gifted; the fraction of students who are in a GHA classroom; average lagged reading & math scores; and the fractions of students who are FRL, white, black, and Hispanic.

We estimate equation (B1) separately for four samples—in each case excluding one of the four years of our RD estimation sample (2009, 2010, 2011 or 2012). When assigning teacher value added to fourth-grade teachers in a given year, we use the estimates from the sample that excludes that year.

Appendix Figure B1 shows the distribution of TVA among teachers of fourth-grade GHA and regular classrooms in 2009-2012. The full distribution has a standard deviation of 0.14 σ. On average, teachers of GHA classes have slightly higher (0.015 σ) TVA than those in non-GHA classes.

Appendix Figure B1. Teacher value added in reading and math, fourth-grade GHA and regular classrooms in 2009-2012



Appendix Table B1 presents estimates from of models in which teacher value-added is the dependent variable and the controls include school fixed effects. The estimates confirm that within schools, GHA classrooms are assigned slightly better teachers on average; however the difference is not statistically significant. Further, estimates from models that include student's lagged test scores indicates that the sorting of better students to better teachers extends beyond the GHA classroom. In particular, column (2) shows that lagged math scores are significantly correlated with measured teacher value added—suggesting that students who are better in math receive slightly better teachers even if they are not GHA participants. Finally, column (3) shows that conditional on lagged test scores, minorities are assigned to teachers with slightly lower value-added—which is suggestive of race-based bias outside the GHA classroom.

Appendix Table B1. Within-school sorting of students and teachers

	(1)	(2)	(3)
Student is in GHA classroom	0.020	0.013	0.013
	(0.012)	(0.012)	(0.012)
3rd-grade math score (standardized)		0.006**	0.006**
		(0.002)	(0.002)
3rd-grade reading score (standardized)		0.001	0.001
		(0.001)	(0.001)
Student is a minority (Black or Hispanic)			-0.003*
			(0.001)

Note: Dependent variable is estimated teacher value added for fourth-grade reading and math. All test scores are standardized across the district within year and grade. All regression models include school-year fixed effects. Estimation sample is 52,034 students enrolled in the district in fourth grade in 2009-2012. Parentheses contain standard errors, clustered by school. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix C: Effect of Misclassification Error on RD Estimates

The following model formalizes the effect of misclassification errors on the first stage and reduced form estimates from an RD analysis when observed GHA participation status is potentially mis-measured.

Let x denote the observed relative rank of a given student, and assume that $x = 0$ corresponds to the cutoff rank. Let GHA^* denote the student's true GHA status, and let GHA denote her observed status. Assume that:

$$\begin{aligned} P(GHA = 1|GHA^* = 1, x) &= q_1(x) \\ P(GHA = 1|GHA^* = 0, x) &= q_0(x). \end{aligned}$$

Here, $1 - q_1(x)$ is the false negative rate for a student with rank x, and $q_0(x)$ is the corresponding false positive rate. We assume that $q_1(x) > q_0(x)$ and that

$$lim_{x \to 0^-} q_j(x) = lim_{x \to 0^+} q_j(x) = q_j(0),\ j = 0, 1$$

i.e., that the error rates for students ranked just below and just above the cutoff rank are the same. Finally, assume that the *true* first stage relationship is:

$$P(GHA^* = 1|x) = \pi(x)$$

with a discontinuity of size π_1 at $x = 0$:

$$\begin{aligned} lim_{x \to 0^-} \pi(x) &= \pi_0 \\ lim_{x \to 0^+} \pi(x) &= \pi_0 + \pi_1. \end{aligned}$$

Under these assumptions the relationship between *observed* GHA status and rank is:

$$P(GHA = 1|x) = q_0(x) + \pi(x)(q_1(x) - q_0(x)),$$

which implies that the first-stage discontinuity in observed GHA status at $x = 0$ is:

$$D_{FS} = \pi_1(q(0) - q_0(0)).$$

If for example $q_1(0) = 0.9$ and $q_0(0) = 0.1$ (i.e., 10% false negative rate and a 10% false positive rate for students around the cutoff rank) then the observed first stage discontinuity is attenuated by 20% relative to the true discontinuity.

Next, assume that the conditional expectation of a student's achievement scores (y) given her actual GHA status and relative rank can be written as:

$$E[y|GHA^*, x] = \beta GHA^* + f(x)$$

where $f(x)$ is some smooth function of relative rank, and β is the causal effect of GHA participation. Using the expressions above,

$$lim_{x \to 0^-} E[y|x] = \beta\pi_0 + f(0),$$

and

$$lim_{x \to 0^+} E[y|x] \quad = \quad \beta(\pi_0 + \pi_1) + f(0)$$

so the reduced form-discontinuity in test scores is:

$$D_{RF} \quad = \quad \beta\pi_1$$

The probability limit of the two stage least estimate of the effect of participating in a GHA class is the ratio of the discontinuities in the reduced form model and the first stage model, which is:

$$\frac{D_{RF}}{D_{FS}} = \frac{\beta}{q_1(0) - q_0(0)}.$$

Thus, the presence of misclassification error leads to an over-estimate of the treatment-on-the-treated effect.